Structure
Article

Comprehensive Structural Classification of Ligand-Binding Motifs in Proteins

Akira R. Kinjo[1,*] and Haruki Nakamura[1]
[1]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
*Correspondence: akinjo@protein.osaka-u.ac.jp
DOI 10.1016/j.str.2008.11.009

SUMMARY

Comprehensive knowledge of protein-ligand interactions should provide a useful basis for annotating protein functions, studying protein evolution, engineering enzymatic activity, and designing drugs. To investigate the diversity and universality of ligand-binding sites in protein structures, we conducted the all-against-all atomic-level structural comparison of over 180,000 ligand-binding sites found in all the known structures in the Protein Data Bank by using a recently developed database search and alignment algorithm. By applying a hybrid top-down-bottom-up clustering analysis to the comparison results, we determined ~3000 well-defined structural motifs of ligand-binding sites. Apart from a handful of exceptions, most structural motifs were found to be confined within single families or superfamilies, and to be associated with particular ligands. Furthermore, we analyzed the components of the similarity network and enumerated more than 4000 pairs of structural motifs that were shared across different protein folds.

INTRODUCTION

Most proteins function by interacting with other molecules. Therefore, the knowledge of interactions between proteins and their ligands is central to our understanding of protein functions. However, simply enumerating the interactions of individual proteins with individual ligands, which is now indeed possible owing to the massive production of experimentally determined protein structures, would only serve to increase the amount of data, not necessarily our knowledge or understanding of protein functions. What is needed is a classification of general patterns of interactions. Otherwise, it would be difficult to apply the wealth of information to elucidate the evolutionary history of protein functions (Andreeva and Murzin, 2006; Goldstein, 2008), to engineer enzymatic activity (Gutteridge and Thornton, 2005), or to develop new drugs (Rognan, 2007).

In order to classify protein-ligand interactions and to extract general patterns from the classification, it is a prerequisite to compare the ligand-binding sites of different proteins. There are already a number of methods by which to compare the atomic structures or other structural features of functional sites of proteins (see reviews, Jones and Thornton, 2004; Lee et al., 2007).

Applications of these methods lead to the discoveries of ligand-binding site structures shared by many proteins of different folds (Kobayashi and Go, 1997; Kinoshita et al., 1999; Stark et al., 2003; Brakoulias and Jackson, 2004; Shulman-Peleg et al., 2004; Gold and Jackson, 2006). Gold and Jackson (2006) conducted an all-against-all comparison of 33,168 binding sites, the results of which have been compiled into the SitesBase database. They have described several unexpected similarities across different protein folds and applied their method to the annotation of unclassified proteins. More recently, Minai et al. (2008) compared all pairs of 48,347 potential ligand-binding sites in 9,708 representative protein chains, and they demonstrated the applicability of ligand-binding site comparison to drug discovery.

To date, however, no method has been applied to the exhaustive all-against-all comparison of all ligand-binding sites found in the Protein Data Bank (PDB) (Berman et al., 2007), presumably because these methods were not efficient enough to handle the huge amount of data in the current PDB, or because it was assumed that the redundancy (in terms of sequence homology) or some "trivial" ligands (such as sulfate ions) in the PDB did not present any interesting findings. As of June 2008, the PDB contains over 51,000 entries, with more than 180,000 ligand-binding sites, excluding water molecules; hence, naively comparing all the pairs of this many binding sites ($>3 \times 10^{10}$ pairs) is indeed a formidable task. Nevertheless, multiple structures of many proteins that have been solved with a variety of ligands (e.g., inhibitors for enzymes) could provide a great opportunity for analyzing the diversity of binding modes, and some apparently trivial ligands are often used by crystallographers to infer the functional sites from the "apo" structure. In other words, the diversity of these apparently redundant data is too precious a source of information to be ignored.

To handle this huge amount of data, we have recently developed the Geometric Indexing with Refined Alignment Finder (GIRAF) method (Kinjo and Nakamura, 2007). By combining ideas from geometric hashing (Wolfson and Rigoutsos, 1997) and relational database searching (Garcia-Molina et al., 2002), this method can efficiently find structurally and chemically similar local protein structures in a database and produce alignments at atomic resolution independent of sequence homology, sequence order, or protein fold. Using the GIRAF method, we first compile a database of ligand-binding sites into an ordinary relational database management system, and we create an index based on the geometric features with

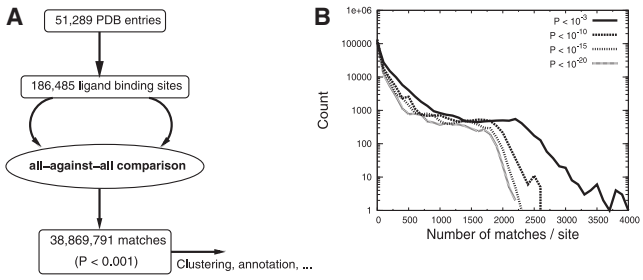


Figure 1. Summary of the Experiment

(A) Flow of the analysis.

(B) Histogram of the number of matches per ligand-binding site.

surrounding atomic environments. Owing to the index, potentially similar ligand-binding sites can be efficiently retrieved, and unlikely hits are safely ignored. For each of the potential hits found, the refined atom-atom alignment is obtained by iterative applications of bipartite graph matching and optimal superposition. In this study, we have further improved the original GIRAF method so that one-against-all comparison takes effectively 1 s, and we applied it to the first all-against-all comparison of all ligand-binding sites in the PDB.

In order to extract recurring patterns in ligand-binding sites, we then classified the ligand-binding sites based on the results of the all-against-all comparison, and defined structural motifs. So far, such structural motifs have been determined either manually (Porter et al., 2004) or automatically (Wangikar et al., 2003; Polacco and Babbitt, 2006). Given the huge amount of data, manual curation of all potential motifs is not feasible, and previously developed automatic methods are computationally too intensive (Wangikar et al., 2003) or limited in scope (e.g., being based on sequence alignment [Polacco and Babbitt, 2006]). Therefore, we first applied divisive (top-down) hierarchical clustering to obtain single-linkage clusters from the similarity network of ligand-binding sites that can be readily obtained from the result of the all-against-all comparison. Based on the hierarchy of the single-linkage clusters, agglomerative (bottom-up) complete-linkage clustering is then applied. Thus, obtained complete-linkage clusters are shown to be well-defined structural motifs, and are then subject to statistical characterization regarding their ligand specificity and protein folds.

Furthermore, based on the result of the all-against-all comparison, we study the structure of the similarity network of ligand-binding sites, and we enumerate interesting similarities shared across different folds. The list of clusters and the list of pairs of ligand-binding sites not sharing the same fold are available online (http://pdbjs6.pdbj.org/~akinjo/lbs/).

RESULTS

All-Against-All Comparison of Ligand-Binding Sites

Out of 51,289 entries in the PDB (Berman et al., 2007) as of June 13, 2008, all 186,485 ligand-binding sites were extracted and compiled into a database. A ligand-binding site is defined as the set of protein atoms that are within 5 Å from any of the corresponding ligand atoms. To define a ligand, we used the annotations in PDB's canonical extensible markup language (XML) files (PDBML) (Westbrook et al., 2005) because these annotations are more accurate than the HETATM record of the flat PDB files. Our definition of ligands includes not only small molecules, but also polymers such as polydeoxyribonucleotide (DNA), polyribonucleotide (RNA), polysaccharides, and polypeptides with less than 25 amino acid residues; water molecules and ligands consisting of more than 1000 atoms were excluded. We did not exclude "trivial" ligands such as sulfate (SO_4^{2-}), phosphate (PO_4^{3-}), and metal ions. We did not use a representative set of proteins based on sequence homology to reduce the data size.

In total, the all-against-all comparison yielded 38,869,791 matches with P-value < 0.001, with 208 matches per site on average (Figure 1A). Whereas 5014 sites found no hits other than themselves, 8369 sites found more than 1000 matches. When we limit the matches to more stringent P-value thresholds (10^{-10}, 10^{-15}, 10^{-20}), the long tail of the large number of matches rapidly disappears (Figure 1B), indicating that many matches reflect partial and weak similarities between sites.

Relationship between Similarities of Protein Sequences and Ligand-Binding Sites

As noted above, the present data set is highly redundant in terms of sequence homology. If the similarity of ligand-binding sites is sharply correlated with that of amino acid sequences, it would have been better to use sequence representatives. To justify the use of the redundant data set, we carried out an all-against-all BLAST (Altschul et al., 1997) search of all protein chains of the present data set, and we checked the correlation between sequence identity and the GIRAF P-value (Figure 2A). It should be noted that a ligand-binding site may reside at an interface of more than two protein subunits (chains), which complicates the notion of representative chains. Therefore, we defined sequence similarity between two PDB entries as the maximum sequence identity of all of the possible pairs of chains from the two PDB entries.

While there was a significant but very weak negative correlation between the GIRAF P-value and the percent sequence identity (Pearson's correlation −0.14), there were many strikingly similar (GIRAF P-value < 10^{-50}) pairs of ligand-binding sites with low (<30%) sequence identity, and there were also many weakly similar ligand-binding sites (GIRAF P-value >10^{-20}) at a high (>90%)-sequence identity region. This tendency was also confirmed by using more conventional measures of similarities. Although the root-mean-square deviation (rmsd) of aligned atoms exhibited a stronger negative correlation with the sequence identity (Figure 2B; Pearson's correlation −0.46), the range of scatter of rmsd was so large that it was not possible to distinguish the range of sequence identity from rmsd values and vice versa. In addition, the number of aligned atoms did not correlate with the sequence identity (Figure 2C), indicating that the local structures of ligand-binding sites can be strictly conserved among distantly related proteins. Visual inspection suggested a few possible reasons for the large deviation in the region of high sequence similarity. First, the binding sites do not necessarily overlap completely when different ligands are complexed with (almost) identical proteins. Second, many binding sites are flexible, yet they are able to bind the same ligand. Third, some ligands are flexible and can be bound as

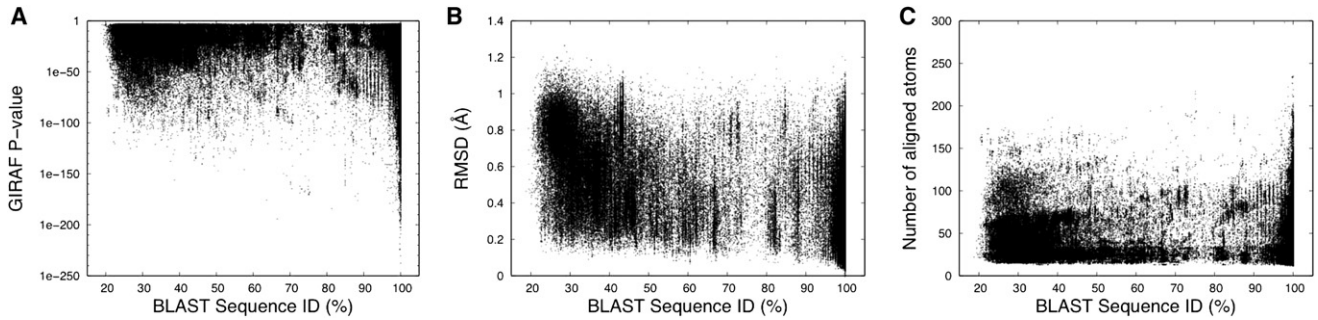


Figure 2. Relationship between Sequence Similarity and Ligand-Binding Site Similarity

(A) Sequence identity of BLAST hits versus GIRAF P-values.

(B) Sequence identity of BLAST hits versus the root-mean-square deviation of aligned ligand-binding sites found by GIRAF.

(C) Sequence identity of BLAST hits versus the number of ligand-binding site atoms aligned by GIRAF.

different conformers, which, in turn, causes structural changes of the binding site.

One of the rationales for an exhaustive all-against-all comparison is that some similarities between nonrepresentative proteins would be ignored when only sequence representatives were used. For example, in the results of a comparison of potential ligand-binding sites of 9708 sequence representative proteins conducted by Minai et al. (2008), the similarity between the ADP-binding sites of human inositol (1,4,5)-triphosphate 3-kinase (PDB: 1W2D [Gonzalez et al., 2004]; SAICAR synthase-like fold) and of *Archaeoglobus fulgidus* Rio2 kinase (PDB: 1ZAR [Laronde-Leblanc et al., 2005]; Protein kinase-like fold) was not detected, although this match was found to have a P-value of 8.1×10^{-17} (40 aligned atoms; rmsd 0.75Å) in the present result. Furthermore, equivalent matches were not found in all homologs of these two proteins. We note, however, that Minai et al. (2008) did find an equivalent similarity between the binding sites of these protein folds, but it was based on apo structures that were not treated here. Thus, the similarity not detected by Minai et al. (2008) is likely to be due to the use of representatives, but not due to the difference in sensitivity of their method and the present one.

We conclude that the similarity of sequences and that of ligand-binding site structures is weakly correlated, but the correlation is not strong enough to infer the one from the other.

Defining Structural Motifs of Ligand-Binding Sites

We have seen that sequence representatives are not suitable for studying the diversity of ligand-binding sites. The use of the raw data of ligand-binding sites for statistical analysis, however, would be problematic due to some overrepresented and underrepresented binding sites. Therefore, it is preferable to remove the redundancy based on the ligand-binding similarity itself. Furthermore, a list of pairwise similarities is not sufficient for characterizing typical patterns of binding modes. Accordingly, we applied the hybrid top-down-bottom-up clustering method to obtain complete-linkage clusters based on P-values. In a complete-linkage cluster (hereafter referred to as "cluster"), any pair of its members are similar within the specified P-value threshold. As such, clusters may be regarded as precisely defined structural motifs of ligand-binding sites; hence, we use the terms "cluster" and "structural motif" (or simply "motif") interchangeably when appropriate. Based on the analysis of similarity networks with varying thresholds (see below), we set the threshold to 10^{-15} in the following analysis.

It is immediately evident that there are a large number of small clusters and a small number of large clusters (Figure 3A). Excluding 58,001 singletons (clusters with only one member), there were 20,224 clusters that accounted for 128,484 (69%) of the 186,485 sites. Out of these clusters, 2959 clusters consisted of at least 10 sites, accounting for 69,748 (37%) sites. The list of these clusters of structural motifs is available online (http://pdbjs6.pdbj.org/~akinjo/lbs/cluster.xml). Since the ligand-binding sites in small clusters are not reliable due to statistical errors, we use only the 2959 clusters consisting of at least 10 sites in the following analysis unless otherwise stated. Furthermore, in the following analysis, redundancy in each cluster was removed by grouping identical binding sites. Two binding sites were defined to be identical if they have the identical ligand and the rmsd of their alignment was less than 0.01 Å. In turn, two ligands were defined to be the same if they had the same InChI (http://old.iupac.org/inchi/) code (available in the PDB chemical component dictionary). This procedure is necessary because some PDB chemical component identifiers are synonyms. For convenience, each InChI code is represented by a representative PDB chemical component identifier in the following. We note, however, that only 2025 binding sites were found to be identical to other sites; hence, the redundancy in clusters is relatively rare.

Diversity of Structural Motifs with Respect to Ligand Types

Although some structural motifs included binding sites for a wide variety of ligand types, this is not always the case (Figure 3B). Here, each PDB chemical component identifier (consisting of 1 to 3 letters) corresponds to a ligand type, except for peptides, nucleic acids, or sugars, which were treated simply as such (i.e., polymer sequence identity is ignored). Large clusters associated with many kinds of ligands were almost always enzymes such as proteases (eukaryotic or retroviral), carbonic anhydrases, protein kinases, and protein phosphatases, whose structures have been solved with a variety of inhibitors. For example, two structural motifs consisting of 245 and 147 ligand-binding sites of eukaryotic (trypsin-like) proteases were

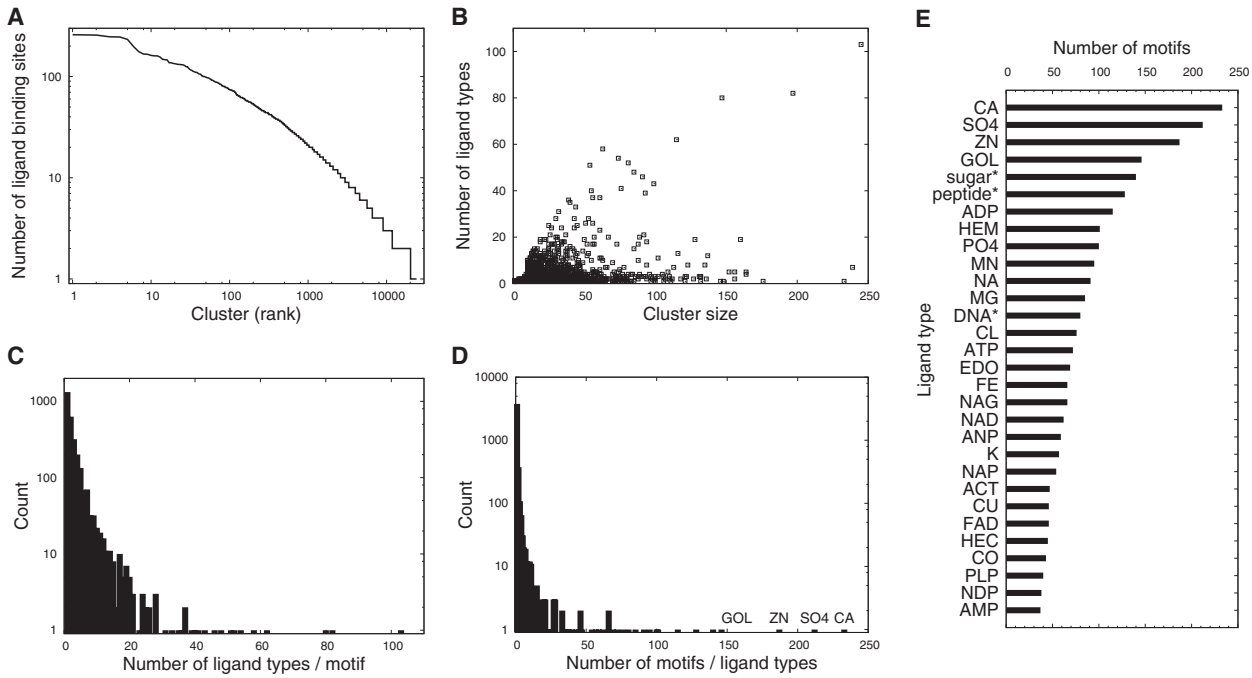


Figure 3. Statistical Properties of Structural Motifs

(A) Size of complete-linkage clusters defined with P-value thresholds of 10^{-15}.

(B) Scatter plot of cluster size versus ligand types found in the cluster.

(C) Histogram of the number of ligand types per structural motif (cluster).

(D) Histogram of the number of structural motifs (clusters) associated with a given ligand type.

(E) The 30 most abundant ligand types (polymer molecules are marked with an asterisk).

associated with 103 and 80 ligand types, respectively; two motifs consisting of 197 and 115 ligand-binding sites of retroviral proteases were associated with 82 and 62 ligand types, respectively; and a motif of 63 ligand-binding sites of protein kinases was associated with 58 ligand types. On the contrary, large clusters with a limited variety of ligands were binding sites for heme (globins and nitric oxide synthase oxygenases) or metal ions. Each structural motif is associated with 3.2 ligand types on average (standard deviation of 5.3): 1323 motifs (45% of 2959 motifs) are associated with only one ligand type, and 2809 motifs (95%) are associated with less than 10 ligand types, whereas only 34 motifs contained more than 20 ligand types (Figure 3C). In general, the diversity of ligand types per structural motif is low.

The converse is also true. That is, the number of structural motifs associated with each ligand type (in terms of InChI code) is generally very limited, with an average of 2.1 motifs (standard deviation 8.4) per ligand type (Figure 3D), and 3770 ligand types correspond to single motifs. Nevertheless, there were some ligands that were associated with many motifs (Figure 3E). As expected, ligands often included in the solvent (e.g., SO_4 [sulfate], MG [magnesium ion], GOL [glycerol], EDO [ethanediol]) were found in many motifs. Reflecting a large number of possible sequences, polymer molecules including peptide, sugar, and DNA were also found to be bound with many motifs. Other than these, mononucleotides and dinucleotides and metal ions exhibited a wide range of binding modes.

Diversity with Respect to Protein Families and Folds

Not many, but some, structural motifs were found to contain ligand-binding sites of distantly related proteins. To quantitatively analyze the diversity of structural motifs in terms of homologous families and global structural similarities, we assigned protein family, superfamily, fold, and classes to each structural motif according to the SCOP (Murzin et al., 1995) database. More concretely, the most specific SCOP code (SCOP concise classification string, SCCS) was assigned to each motif that was shared by all members of the corresponding cluster when it was possible, otherwise (i.e., there is at least one member that is different from other members in the cluster at the class level) motif was categorized as ''others'' (Figure 4A).

Out of 2705 motifs to which SCCS can be assigned, 2637 and 62 motifs shared the same domains at the family and superfamily level, respectively. Thus, more than 99% of the motifs (of at least 10 binding sites) only contained binding sites of evolutionarily related proteins. One motif contained proteins from different superfamilies, but of the same fold. This motif corresponded to the heme-binding site of heme-binding four-helical bundle proteins (SCOP: f.21). Five motifs accommodated similarities across different folds, out of which three were zinc-binding motifs (Krishna et al., 2003). One motif contained a P loop motif that is shared between the P-loop-containing nucleotide triphosphate hydrolases (NTH) (SCOP: c.37) and the PEP carboxykinase-like fold (SCOP: c.91) (Figure 5A) (Tari et al., 1996). One motif was of the nucleotide-binding sites from FAD/NAD(P)-binding domain (SCOP: c.3) and nucleotide-binding domain

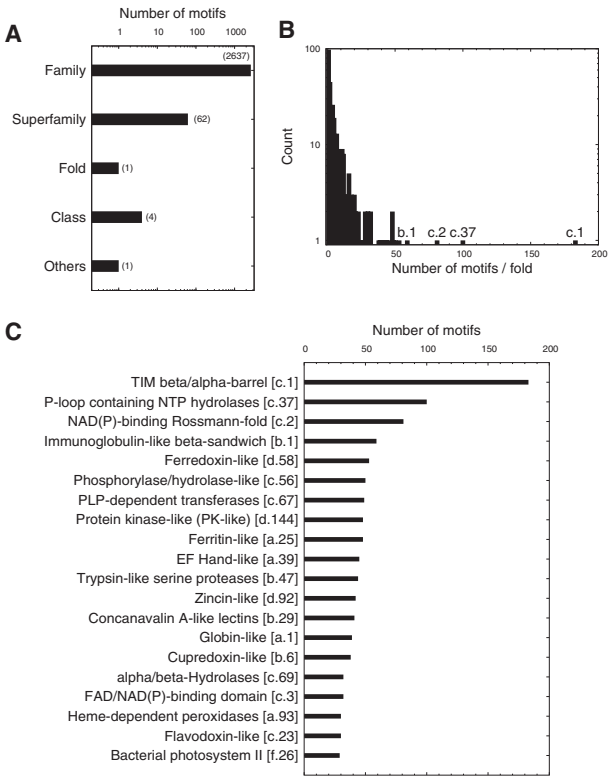


Figure 4. Diversity of Structural Motifs in Terms of Protein Folds

(A) The number of motifs to which the given SCOP (Murzin et al., 1995) hierarchical level (family, superfamily, fold, class) can be assigned.

(B) Histogram of the number of structural motifs associated with each SCOP fold.

(C) The 20 most diverse SCOP folds in terms of the number of associated structural motifs.

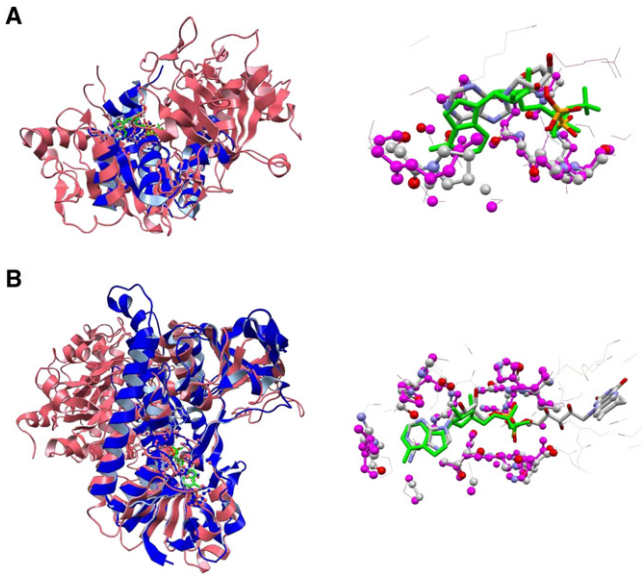


Figure 5. Examples of Structural Motifs Shared by Different Protein Folds

(A and B) The left panel shows the whole protein structures (colored in blue or pink) superimposed based on the alignment of the ligand-binding sites shown in the right panel (colored in the CPK scheme or magenta [protein] and green [ligand], respectively). (A) ADP-binding site of bacterial shikimate kinase (PDB: 2DFT [Dias et al., 2007]; SCOP: c.37; blue/CPK-colored) and ATP-binding site of bacterial phosphoenolpyruvate carboxykinase (PDB: 1AQ2 [Tari et al., 1997]; SCOP: c.91; pink/protein in magenta, ADP in green). (B) FAD-binding site of human glutathione reductase (PDB: 5GRT [Stoll et al., 1997]; SCOP: c.3; blue/CPK-colored) and ADP-binding site of bacterial trimethylamine dehydrogenase (PDB: 2TMD [Barber et al., 1992]; SCOP: c.4; pink/protein in magenta, ADP in green).

(SCOP: c.4) (Figure 5B). These two examples have also been noticed by Brakoulias and Jackson (2004). Note that some PDB entries have not yet been annotated in SCOP. Currently, if such members exist in a cluster, they are simply ignored, and the assigned SCCS is based only on the members whose SCCS is known. Therefore, the number of motifs not sharing the same folds is somewhat underestimated. Nevertheless, it seems to be a general tendency that most motifs are confined within homologous proteins, namely, families or superfamilies.

It was shown above that sequence similarity was only weakly related to the structural similarity of ligand-binding sites (Figure 2). This point can be further clarified by examining motifs of similar binding sites of related proteins. For example, the peptide-binding sites of a pig trypsin (PDB: 1UHB [Pattabhi et al., 2004]) and of a human hepsin (PDB: 1Z8G [Herter et al., 2005]) were both in the same cluster, but they share little sequence similarity (5% sequence identity based on a structural alignment [Kawabata and Nishikawa, 2000; Kawabata, 2003]), whereas the peptide-binding site of bovine trypsin (PDB: 1QB1 [Whitlow et al., 1999]) in another cluster shares 81% sequence identity with the pig thrombin in the previous cluster. This observation can be explained by the fact that different motifs cover different regions of proteins even though they are spatially close or even partially overlapping. The same argument applies to other motifs of related proteins. Thus, the structural motifs distinguish subtle differences in ligand-binding site structures independent of sequence similarity.

It has been known that some protein folds can accommodate a wide range of functions. It is expected that the diversity of function is reflected in that of structures of ligand-binding sites. To analyze such a tendency, we counted the number of motifs that belong to each protein fold (Figure 4B). Only a handful of folds showed a large diversity in terms of structural motifs. On average, 8.9 motifs were assigned to a fold. Out of 332 folds used in the analysis, only 18 contained more than 30 motifs (Figure 4C). Among them, the TIM barrel fold was an extreme case, with 183 motifs assigned, reflecting the great diversity of its functions (Nagano et al., 2002). Some superfolds (Orengo et al., 1994), such as Rossmann-fold, immunoglobulin-like, globin-like, etc., also showed great diversities of ligand-binding sites.

Similarity Network of Ligand-Binding Sites

While each motif defines a precise pattern of ligand-binding mode, the members of different structural motifs share significant structural similarities with each other. To explore the global structure of the "ligand-binding site universe," we constructed a similarity network based on the results of the all-against-all comparison. Each structural motif was represented as a node, and two nodes were connected if a member of one node was

significantly similar to a member of the other node (i.e., the P-value of their alignment was below a predefined threshold). Thus, a constructed network can be decomposed into a number of connected components. When the threshold was greater than 10^{-14}, the size of the largest connected component of the network was one or two orders of magnitude greater than that of the second largest one (Figure 6A). For example, setting the threshold to 10^{-10} yielded the largest connected component consisting of 78,190 sites (i.e., 42% of 186,485 sites). Accordingly, many functionally unrelated binding sites were somehow connected in the largest component, which complicated the interpretation of the component. With the P-value threshold of 10^{-15} or less, the first several connected components were of the same order (Figure 6A), and many members of each component appeared to be more functionally related. Thus, we set $p = 10^{-15}$ for constructing the network described in the following sections (as well as for defining the complete-linkage clusters described above). It is possible that a pair of binding sites from two different complete-linkage clusters (defined with $p = 10^{-15}$) may be similar to each other, with $P < 10^{-15}$. They nevertheless belong to different clusters, otherwise the completeness of the cluster would not hold. This may happen because the similarity is not sufficiently strong and/or the similarity is based on peripheral regions of the binding sites. Therefore, a link between different clusters indicates partial similarities between these motifs.

Excluding 54,092 singleton components (those consisting of only one site), 11,532 connected components were found. The largest component consisted of 7935 sites, and 1881 components contained at least 10 sites (Figure 6A). The network diagrams of the five largest connected components are shown in Figures 6B–6F and are described in the legend.

Main constituents of the largest component were mononucleotide- and phosphate-binding sites (Figure 6B). It is surprising that the heme-binding site of globins (hemoglobins, myoglobins, cytoglobins, etc.) was also included in this component. Nevertheless, it was not directly connected to the main group of P loops, but was indirectly connected via the sparse group consisting of the chloride ion-binding site of T4 lysozymes and sulfate- and phosphate-binding sites of miscellaneous proteins. The binding sites of this latter group were made of regular structures at the termini of α helices. When we used a more stringent P-value threshold (say, 10^{-20}), the groups of globins and lysozymes were detached from the main group, but the main group containing the P-loops was almost unaffected (data not shown). Thus, the matches connecting globins, lysozymes, and P-loop-containing proteins may be considered to be "false" hits. Based solely on structural similarity, however, they are difficult to discriminate from "true" hits (structural matches between functionally related sites) since many functional sites often include regular structures at termini of secondary structures. Nevertheless, the fact that only a subset of regular structures was detected suggests that these matches may correspond to recurring structural patterns often used as building blocks of functional sites. In addition, we point out that weak but meaningful enzymatic functions are sometimes detected experimentally in such "false" hits (Ikura et al., 2008).

Some "false" hits were also found in the fifth largest cluster whose main constituents were the ATP (and inhibitor)-binding sites of protein kinase family proteins (Figure 6F). There was a large, sparse group connected with the main group of protein kinases. In that sparse group, ligand-binding sites of transthyretins (prealbumins) were often found to be directly connected with that of protein kinases, although their folds are different. These binding sites both involve a face of a β sheet, and their similarity was found due to the backbone conformation of the β sheet. Since many proteins bind their ligands on a face of a β sheet, this observation, in turn, explains the origin of the large, sparse group.

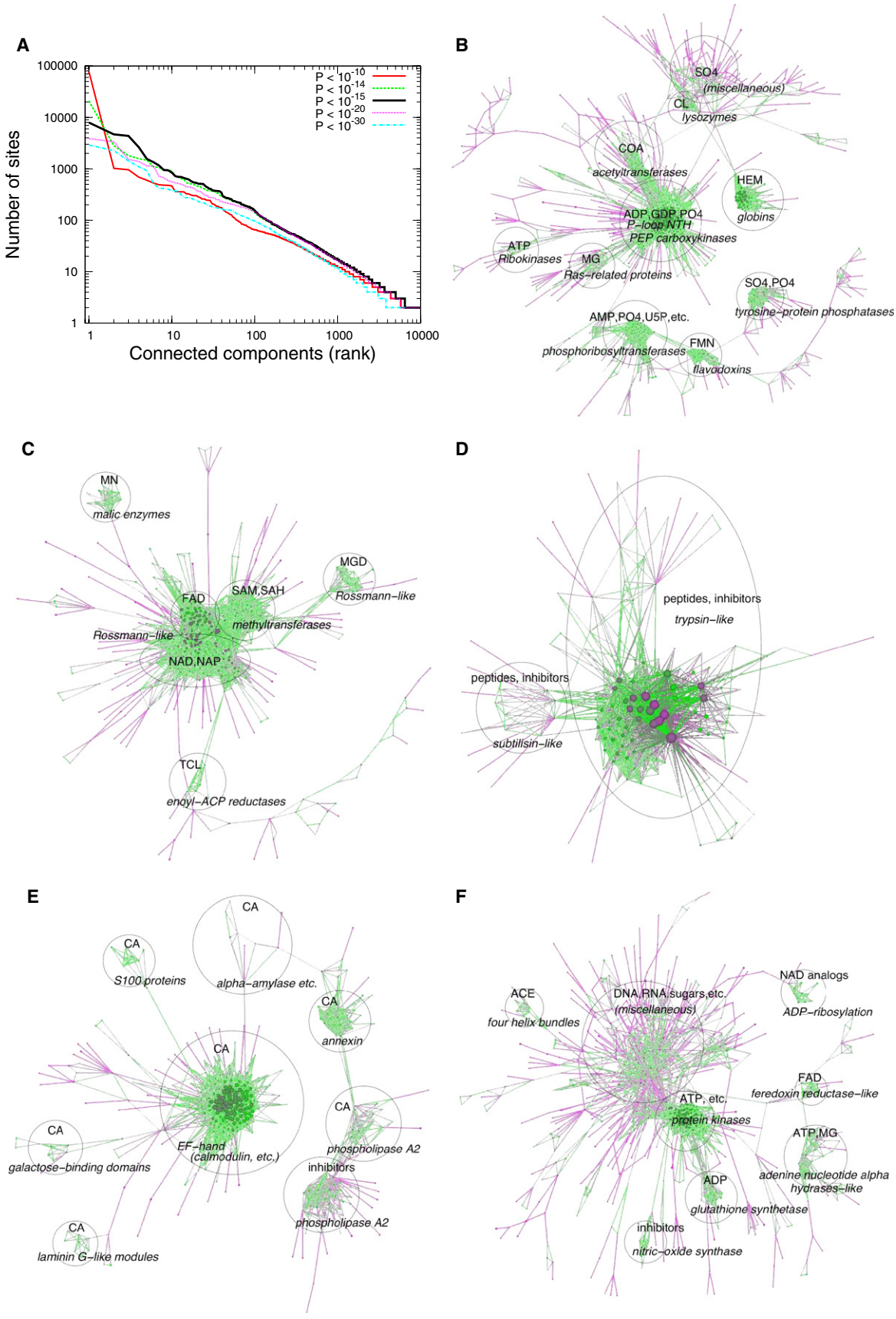
Significant Similarities across Different Folds

The similarity network of ligand-binding sites revealed many structural similarities across different folds. To explore the extent of significant "crossfold" similarities (with $P < 10^{-15}$), we assigned SCOP codes to as many structural motifs as possible, and we enumerated motif pairs whose members were significantly similar but did not share a common fold (Figure 7A). We also examined the ligand pairs in those matches, and we found that most of them were reasonable matches (Figure 7B): metal ions were matched with metal ions, nucleotides with nucleotides or phosphate, and so on. Thus, many of these crossfold similarities are expected to be functionally relevant. The observation that sulfate (SO_4)-binding sites were often found to be matched with mononucleotide- (GDP and ATP) or phosphate (PO_4)-binding sites (Figure 7B) confirms the usefulness of the former ligand in inferring the binding of the latter ligands, as often practiced by crystallographers. We note that multiple SCCS may be assigned to a single motif if it contains multiple fold types or its member sites are located at an interface of multiple domains. In order to cover all possible fold pairs, we did not exclude motifs consisting of less than ten binding sites in this analysis. There were, in total, 4,035 pairs of structural motifs (52,709 pairs of binding sites) that exhibited significant similarities but did not share the same fold. The complete list of these pairs is available online (http://pdbjs6.pdbj.org/~akinjo/lbs/diffold.xml), and descriptions of some notable similarities are found in the legend of Figure 7.

As noted in the description of a network component (Figure 6F), protein kinases and transthyretins share similar binding sites that are located on a face of a β sheet (Figure 8E). Nevertheless, their ligand moieties also seem similar.

Also, as seen in the network component (Figure 6B), the phosphate-binding site of the P loop motif exhibits a significant similarity with the CoA-binding site of acetyltransferases (Figure 8F). A close examination showed that the phosphate bound to the P loop motif coincided with the phosphate group of CoA bound to the acetyltransferase.

The list of the crossfold similarities contained many other examples, including, but not limited to, those discussed in the context of the similarity network. Here, we give two other examples. Bacterial peptide deformylase 2 (SCOP: d.167) and human macrophage metalloelastase (SCOP: d.92) both act with peptides, and their ligand-binding sites exhibit high structural similarity (Figure 8G). DNA is one of the most abundant ligands found in crossfold similarities (Figure 7B). Not surprisingly, similarity between binding sites for DNA and RNA can also be found. One example is the KH1 domain of human poly(rC)-binding protein 2, which binds DNA and bacterial transcription



elongation protein NusA, which binds RNA (Figure 8H). These proteins have different variants of the KH domains (Grishin, 2001b).

It is natural to ask how many of these crossfold similarities are already known and how many are newly found. Since a systematic comparison is difficult due to the lack of a standardized database of structural motifs, we examined the overlapping similarities found in the present study and those in SitesBase (Gold and Jackson, 2006) for several binding sites, including the examples studied by Gold and Jackson (2006) and representative entries of connected components shown in Figure 6 (Table 1). With the P-value threshold of 10^{-15} ("Poisson Index" [Davies et al., 2007] in the case of SitesBase), the results of the present study mostly cover the SitesBase results (except for 1RP4, yeast Ero1p [Gross et al.,]). This tendency is attributed to the use of the updated database (SitesBase is based on a version of PDB in June 2005) as well as the inclusion of "trivial" ligands. To dissect these two effects, we also compared the results using PDB entries released by June 2005 (the numbers in parentheses in Table 1), and confirmed that including "trivial" ligands helps to find more significant similarities. When a more generous P-value threshold (10^{-3}) was used, however, the overlap was small (data not shown), indicating that weak similarities were detected differently by the two methods.

DISCUSSION

From the result of the exhaustive all-against-all comparison, we were able to obtain an extensive list of ligand-binding site similarities irrespective of sequence homology or global protein fold. The similarity network uncovered many crossfold similarities as well as well-known ones. Although it is still not clear how many of these similarities are functionally relevant, it was often observed that different folds were superimposable to a significant extent when the alignment was based on the ligand-binding sites (e.g., Figures 5 and 8). Aligning protein structures based on ligand-binding sites (or functional sites in general) irrespective of sequence similarity, sequence order, and protein fold (as currently defined) may be a useful approach to elucidating the evolutionary history of fold changes (Grishin, 2001a; Krishna and Grishin, 2004; Andreeva and Murzin, 2006; Taylor, 2007; Goldstein, 2008; Xie and Bourne, 2008).

As was seen in the similarity network (Figure 6), some links are based on the similarity of highly regular (secondary) structures that are found in many protein structures (e.g., Figures 6B and 6F). Although such similarities may not be directly related to any biochemical functions, they suggest that many ligand-binding sites are based on combinations of some regular local structures. It is known that a relatively small library of backbone fragments can accurately model tertiary structures of proteins (Kolodny et al., 2002). Consequently, the variety of contiguous fragments recurring in ligand-binding sites is also limited as far as backbone structure is concerned. Friedberg and Godzik (2005) found similarities across different protein folds, including those involved in various zinc-finger motifs and Rossmann-like folds, as shown in this study. They also showed significant correlations between the similarity of fragments and that of protein functions. This observation is consistent with the present results in that it suggests that specific combinations of fragments encode specific functions. To apply the GIRAF method to functional annotations, however, it is preferable to discriminate functionally relevant similarities from purely structural similarities.

Some of the short-comings of simple pairwise comparison may be overcome by the complete-linkage clustering analysis of similar binding sites, which allowed us to define precise structural motifs. It should be stressed that defining reliable

Figure 6. Networks of Structural Motifs of Ligand-Binding Sites

(A) Distribution of the size of the connected component of the similarity network with varying P-value thresholds. A transition is observed at $p = 10^{-15}$.

(B–F) The five largest connected components of the similarity network (P-value threshold = 10^{-15}). Some groups of structural motifs are marked by black circles annotated with ligand types and protein folds. To facilitate visualization, each node (shown as a sphere) is represented as a complete-linkage cluster (structural motif) of ligand-binding sites defined with $P = 10^{-15}$ (the sphere size is proportional to the cluster size). Nodes and edges are colored according to the values of their clustering coefficient (green, high; magenta, low) (Watts and Strogatz, 1998). (B) The largest connected component of the similarity network. Main constituents were mononucleotide- (ADP, GDP, etc.) or phosphate-binding (PO4) sites. Most notable were P-loop-containing NTH (SCOP: c.37) and PEP carboxykinases (SCOP: c.91), which formed a closely connected group, as they share similar phosphate-binding sites, i.e., the P loop motif (the term "group" used here indicates closely connected clusters in a network component colored in green in [B]–[F]). Directly connected with this group was the coenzyme A (CoA)-binding site of acetyl-CoA acetyltransferases. The magnesium ion (MG)-binding site of Ras-related proteins was also connected with the group of the P-loop-containing proteins since the magnesium ion is often located near the phosphate-binding site. Mononucleotide- or phosphate (AMP, U5P, PRP, PO4)-binding sites of various phosphoribosyltransferases and the flavin mononucleotide (FMN)-binding site of flavodoxins were also closely connected. The phosphate-binding site of tyrosine-protein phosphatases formed another group, which was weakly connected to the FMN-binding site of flavodoxins. (C) The second largest connected component. This component mainly consisted of mononucleotide- or dinucleotide-binding sites of the so-called Rossmann-like fold domains, which include, among others, NAD(P)-binding Rossmann-fold domains (SCOP: c.2), a FAD/NAD(P)-binding domain (SCOP: c.3), a nucleotide-binding domain (SCOP: c.4), SAM-dependent methyltransferases (SCOP: c.66), activating enzymes of the ubiquitin-like proteins (SCOP: c.111), and urocanase (SCOP: e.51). (D) The third largest component. Peptide (and inhibitor)-binding sites of trypsin-like and subtilisin-like proteases were found. These two proteases do not share a common fold, but were connected due to the similarity of the active site structures around the well-known catalytic triad. (E) The fourth largest component. The EF hand motif, a major calcium-binding motif, was found in addition to a variety of other calcium ion-binding sites. Although the main group in this component mostly consisted of the calcium ion-binding sites of various calmodulin-like proteins, it also contained similar sites of periplasmic-binding proteins (PBP). The ligands of these PBPs include sodium in addition to calcium ions. The main group was weakly connected to the calcium ion-binding sites of proteins of completely different folds such as galactose-binding domains (e.g., galactose oxidase, fucolectins), laminin G-like modules (e.g., laminin, agrin, etc.), α-amylases, annexins, and phospholipase A2. Due to its spatial proximity, the calcium-binding site of phospholipase A2 was also connected to its inhibitor-binding sites. (F) The fifth largest component. Most binding sites are associated with nucleotides. The main closely connected group consisted of the ATP (and inhibitors)-binding sites of protein kinase family proteins, next to which the ADP-binding sites of glutathione synthetase family proteins (including D-ala-D-ala ligases) were connected. Other closely connected groups included FAD-binding sites of ferredoxin reductase-like proteins and ATP, magnesium-binding sites of adenine nucleotide alpha hydrolases-like proteins, inhibitor-binding sites of nitric-oxide synthases, and NAD (analog)-binding sites of ADP-ribosylation proteins (e.g., T cell ecto-ADP-ribosyltransferase 2, iota toxin, etc.).

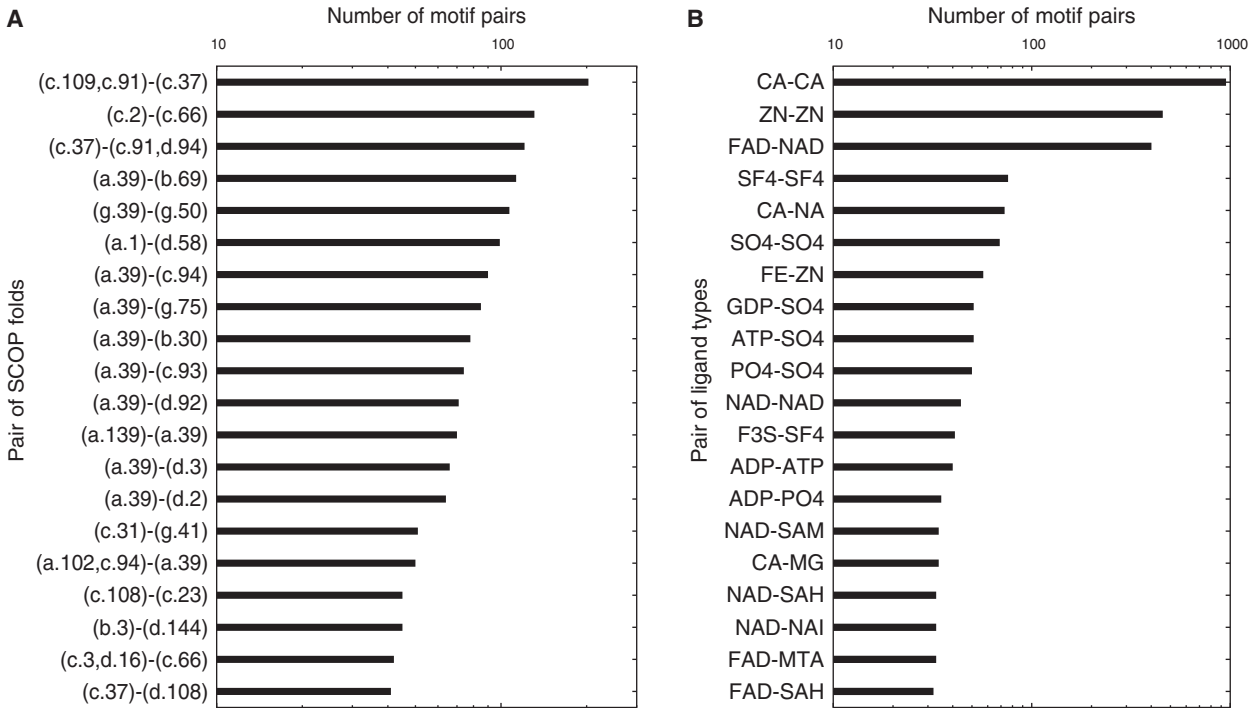


Figure 7. Ligand-Binding Sites Shared across Different Protein Folds

(A and B) (A) The 20 most common pairs of different folds sharing significant ligand-binding site similarities. (B) The 20 most common pairs of ligand types shared across different folds. Notes on some notable similarities follow: The most common crossfold similarity was found between the P-loop-containing NTH (SCOP: c.37) and the PEP-carboxykinase-like (SCOP: c.91) (c.f. Figure 5A). As described in the analysis of complete-linkage clusters, this corresponds to mononucleotide- or phosphate-binding sites. Mononucleotide- or dinucleotide-binding sites of various Rossmann-like folds (SCOP: c.2, c.3, c.4, c.66) also exhibited significant mutual similarities (e.g., Figures 5B and 8A). The calcium-binding sites of the EF hand-like fold (a.39) were found to be similar to the metal-binding sites of many folds, including β-propeller proteins (Figure 8B) and periplasmic-binding proteins (SCOP: c.93 [class I], c.94 [class II]), lysozyme-like (SCOP: d.2), Zincin-like (SCOP: d.92), and many others. Similar zinc-binding sites were found in many, mostly small, folds in addition to the DHS-like NAD/FAD-binding domain (SCOP: c.31) and Rubredoxin-like (g.41) (Figure 8C), the former of which may be regarded as an inserted zinc-finger motif. The similarity between globin-like (SCOP: a.1) and ferredoxin-like (SCOP: d.58) was due to the coordinated structures of the iron-sulfur clusters found in α-helical ferredoxins and ferredoxins, respectively. HAD-like fold proteins (SCOP: c.108) and CheY-like (flavodoxin fold) proteins (SCOP: c.23) often share similar binding sites (e.g., Figure 8D). Interestingly, although these proteins have very similar topologies, the orders of aligned secondary structure elements were different when the alignment was based on the ligand-binding site similarity.

motifs requires redundancy in the PDB (Wangikar et al., 2003), otherwise it would be more difficult to distinguish recurring structures from incidental matches. These motifs may be useful for defining structural templates for efficient motif matching (Wallace et al., 1997). Despite the diversity of binding sites and their similarities, most motifs were found to be confined within single families or superfamilies, and they were also found to be highly specific to particular ligands. Thus, these motifs may be helpful for annotating putative functions of proteins, especially of structural genomics targets.

However, we point out that a structural motif is defined as a set of mutually similar ligand-binding sites in the present study as well as in other studies (e.g., Brakoulias and Jackson, 2004). In other words, the format for expressing an abstract or idealized structural motif has not been developed, in contrast to the case of sequence motifs, which can be expressed as regular expressions. Defining a standard format for structural motifs would be useful for the fast retrieval and annotation of motifs, and for comparing different algorithms, but it is left for future studies. Perhaps a generalization of the 3D query template format of the TESS algorithm (Wallace et al., 1997) may be a good candidate for a structural motif format.

In conclusion, the development of an extremely efficient search method (GIRAF) to detect local structural similarities made it possible to conduct the first, to our knowledge, exhaustive all-against-all comparison of all ligand-binding sites in all known protein structures. We identified a number of well-defined structural motifs and enumerated many nontrivial similarities. Although exhaustive pairwise comparisons are useful for detecting weak and possibly partial similarities between ligand-binding sites, the significance of such matches may not be immediately obvious because some of them may be based on ubiquitous regular structures.

Meanwhile, complete-linkage clusters of ligand-binding sites are useful for identifying functionally relevant binding site structures, but they may neglect partial but significant matches. Therefore, these two approaches, exhaustive pairwise comparison and motif matching, are complementary to each other; hence, the combination thereof may be helpful for more reliable annotations of proteins with unknown functions. These approaches may be further supplemented by other existing

Table 1. Comparison between GIRAF and SitesBase

		GIRAF		SitesBase		Shared	
PDB	Ligand	Entry	Family	Entry	Family	Entry	Family
9LDT[a]	NAD	250 (221)	29 (26)	128	12	122 (122)	10 (10)
1RM8[a]	BAT	96 (75)	4 (4)	54	3	54 (54)	3 (3)
1M6Z[a]	HEC	21 (20)	7 (6)	5	2	5 (5)	2 (2)
1RP4[a]	FAD	2 (2)	1 (1)	3	1	1 (1)	0 (0)
1BYU[b]	GDP	556 (414)	35 (34)	357	23	303 (303)	23 (23)
1KYQ[c]	NAD	341 (297)	43 (38)	9	6	7 (7)	6 (6)
1AD8[d]	MDL	723 (611)	6 (6)	344	2	333 (333)	2 (2)
1PKD[e]	UCN	307 (208)	2 (1)	91	1	89 (89)	1 (1)

The table shows the number of PDB entries ("Entry") and SCOP families ("Family") detected by GIRAF and SitesBase, and those shared by both (with a P-value threshold of 10^{-15}). The numbers in the parentheses indicate the results with PDB entries released by June 2005 (same as SitesBase). References: 9LDT, Dunn et al. (1991); 1RM8, Lang et al. (2004); 1M6Z, A. Noergaard et al., personal communication; 1RP4, Gross et al. (2004); 1BYU, Stewart et al. (1998); 1KYQ, Schubert et al. (2002); 1AD8, Malikayil et al. (1997); 1PKD, L.N. Johnson et al., personal communication.
[a] Entries annotated in Gold and Jackson (2006).
[b] Other entries are the "centers" of connected components (entries with the greatest number of connections in each component): Figure 6B;
[c] Figure 6C;
[d] Figure 6D;
[e] Figure 6F. (Entries corresponding to Figure 6E were not treated since SitesBase does not include calcium ion-binding sites.)

fold and/or sequence-based methods (Standley et al., 2008; Xie and Bourne, 2008). Alternatively, sequence information (as well as optional sequence-order constraints) may be directly incorporated into the GIRAF method in a manner similar to that described by Jonassen et al. (2000). This might help to rescue some false-negative hits that are currently not detected due to structural deviations in spite of conserved sequence motifs.

The present method can also be applied to a whole protein structure (not limited to its predefined ligand-binding sites) to find potential ligand-binding sites (Kinjo and Nakamura, 2007). In this way, we are currently annotating all structural genomics targets (Chen et al., 2004). We also plan to make this method available as a web service so that structural biologists can routinely search for ligand-binding sites of their interest.

EXPERIMENTAL PROCEDURES

The GIRAF Method

The details of the original GIRAF method has been published elsewhere (Kinjo and Nakamura, 2007). Here, we provide a brief summary of the method. In this study, an improved version of GIRAF was used for conducting the all-against-all comparison. The improvement includes more sensitive geometric indexing with atomic composition around each reference set, simplified SQL expressions, and parallelization (A.R.K. and H.N., unpublished data). A protein structure is dissected into a set of Delaunay tetrahedra, each of which is characterized by its volume, edge lengths, and compositions of surrounding atoms in the direction of each face. These tetrahedra serve as reference sets ("refsets") for local coordinate systems. The atomic coordinates of template ligand-binding sites expressed in various local coordinate systems are saved in a relational table, with the corresponding refsets indexed by their characteristic values. A query structure is processed in the same manner as the templates, with its refsets saved in a temporary relational table, but local atomic coordinates are saved in a hash table. Query refsets matching with template refsets can be retrieved efficiently by a relational algebraic procedure, after which matching atomic coordinates are counted. After this procedure, promising candidate matches are subject to alignment refinement, which is carried out by iteratively applying a Hungarian algorithm (Lawler, 2001) and optimal superposition (Diamond, 1988) until convergence. Like the methods of Russell (1998) or of Brakoulias and Jackson (2004), the alignment is based solely on coordinates and the chemical identity of atoms, and it does not depend on sequence homology, sequence order, or protein fold. In this respect, GIRAF is in contrast with the method of Jonassen et al. (2000).

All-Against-All Comparison

Ligand-binding sites were extracted from PDBML files as described in Results. Here, ligands were defined as molecular entities satisfying the following criteria: (a) it is not annotated as "water," (b) if it is annotated as "polypeptide(L)," it contains less than 25 amino acid residues, (c) it is annotated neither as "water" nor as "polypeptide(L)". That is, a ligand can be a polypeptide shorter than 25 residues, DNA, RNA, polysaccharides (sugars), lipids, metal ions, iron-sulfur clusters, or any other small molecules. However, ligands with more than 1000 atoms were discarded. The all-against-all comparison was carried out on a cluster machine consisting of 20 nodes of 8-core processors (Intel Xeon 3.2 GHz). The whole computation was finished within ~60 hr.

Clusters of Similar Ligand-Binding Sites

To obtain complete-linkage clusters, we first constructed a single-linkage network based on a predefined P-value threshold. Then, this network was decomposed into connected components. Each component was then broken into finer components by imposing a more stringent P-value threshold. This decomposition was iterated until the P-value threshold reached 10^{-100}. Then, bottom-up complete linkage was iteratively applied to each connected component, the result of which was then combined into an upper component (previously determined with a higher P-value threshold). This bottom-up process was terminated when a P-value threshold of 10^{-15} was reached. Each (complete-linkage) cluster was defined as a structural motif for the ligand-binding sites. Note that, although every pair of binding sites in a single cluster (say, Cluster A) ia similar, with $P < 10^{-15}$ by definition, some (but not all) members of the cluster may also be related to some members of another cluster (say, Cluster B) with $P < 10^{-15}$. Such members of Cluster B are not included in Cluster A because they break the complete-linkage criterion within Cluster A. Nevertheless, this kind of relationship between different clusters indicates partial similarities between the clusters and serves as a basis of network analysis.

Analysis of Networks and Structural Motifs

To annotate thus obtained structural motifs with the SCOP (Murzin et al., 1995) codes, we used the parsable file of SCOP (version 1.73). When an analysis involved SCOP codes, those PDB entries whose SCOP classification has not yet been determined were ignored. Each SCOP SCCS code was assigned

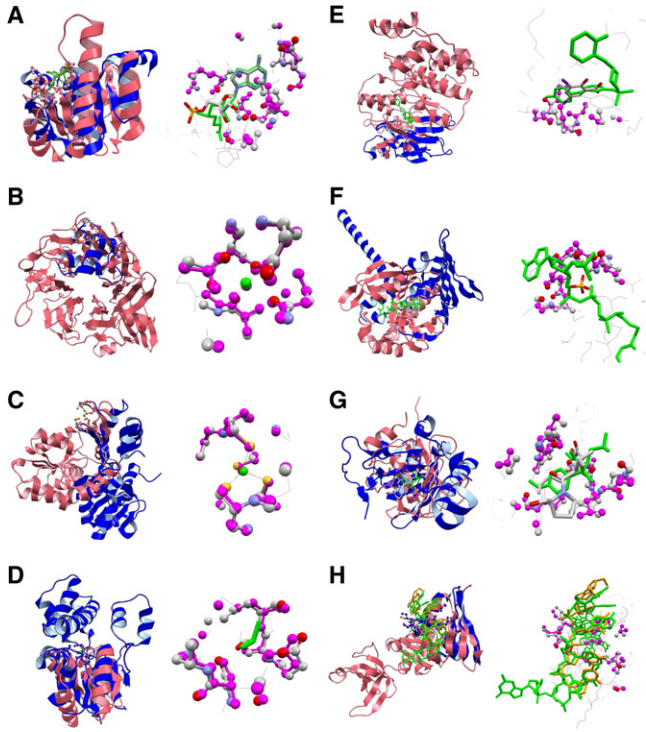


Figure 8. Examples of Ligand-Binding Sites Shared across Different Folds

(A–H) The color schemes are the same as in Figure 5. (A) AMP-binding site of *Thermotoga maritima* hypothetical protein tm1088a (PDB: 2G1U [Joint Center for Structural Genomics (2006)]; SCOP: c.2; blue/CPK colored) and the SAM-binding site of human putative ribosomal RNA methyltransferase 2 (PDB: 2NYU [Wu et al., personal communication]; SCOP: c.66; pink/protein in magenta, SAM in green). (B) Calcium-binding sites of *Clostridium thermocellum* cellulosomal scaffolding protein A (PDB: 2CCL [Carvalho et al., 2007]; SCOP: a.39; blue/CPK-colored) and human integrin α-IIb (PDB: 1TXV [Xiao et al., 2004]; SCOP: b.69; pink/protein in magenta, calcium in green). (C) Zinc-binding sites of human NAD-dependent deacetylase (PDB: 2H4H [Hoff et al., 2006]; SCOP: c.31 [inferred by SSM (Krissinel and Henrick, 2004)]; blue/CPK-colored) and *Bacillus stearothermophilus* adenylate kinase (PDB: 1ZIN [Berry and Phillips, 1998]; SCOP: g.41; pink/protein in magenta, zinc in green). (D) Formic acid-binding site of *Xanthobacter autotrophicus* L-2-haloacid dehalogenase (PDB: 1AQ6 [Ridder et al., 1997]; SCOP: c.108; blue/CPK-colored) and the BeF^{3-}-binding site of *Escherichia coli* PhoB (PDB: 1ZES [Bachhawat et al., 2005]; SCOP: c.23; pink/protein in magenta, BeF^{3-} in green). (E) 3,5-diiodosalicylic acid-binding site of human transthyretin (PDB: 3B56; SCOP: b.3; blue/CPK-colored) and the inhibitor (N-[3-(4-fluorophenoxy)phenyl]-4-[(2-hydroxybenzyl)amino]piperidine-1-sulfonamide)-binding site of human mitogen-activated protein kinase 14 (PDB: 1ZZ2; SCOP: d.144; pink/protein in magenta, inhibitor in green). (F) Phosphate-binding site of *Pyrococcus furiosus* Rad50 ABC-ATPase (PDB: 1II8; SCOP: c.37; blue/CPK-colored) and the coenzyme-A (CoA)-binding site of *Salmonella typhimurium* LT2 acetyl transferase (PDB: 1S7N; SCOP: d.108; pink/protein in magenta, CoA in green). (G) Actinonin-binding site of *B. stearothermophilus* peptide deformylase 2 (PDB: 1LQY [Guilloteau et al., 2002]; SCOP: d.167; blue/CPK-colored) and the NNGH-binding site of human macrophage metalloelastase (PDB: 1Z3J; SCOP: d.92; pink/protein in magenta, NNGH in green). (H) DNA-binding site of the KH1 domain of human poly(rC)-binding protein (PDB: 2AXY [Du et al., 2005]; SCOP: d.51; blue/CPK-colored, DNA in orange) and the RNA-binding site of *Mycobacterium tuberculosis* transcription elongation protein NusA (PDB: 2ATW [Beuth et al., 2005]; SCOP: d.52; pink/protein in magenta, RNA in green).

to a ligand-binding site, as described by others (Gold and Jackson, 2006). When a site resides at an interface of multiple domains, multiple SCCS codes were assigned to the site. Two or more binding sites are said to share the same fold (or family, superfamily, etc.) if the intersection of their SCCS code sets is not empty. The SCCS code assigned to a structural motif was defined as the union of all of the SCCS codes found in the corresponding cluster members. We used only the seven main SCOP classes (all-α [a], all-β [b], α/β [c], α+β [d], multidomain [e], membrane and cell surface proteins and peptides [f], and small proteins [g]). The figures of alignments (Figures 5 and 8) were created with jV version 3 (Kinoshita and Nakamura, 2004) by using the PDBML-extatom files produced by GIRAF. The network figures (Figures 6B–6F) were created with Tulip software (http://www.tulip-software.org/).

ACKNOWLEDGMENTS

The authors thank Kengo Kinoshita, Motonori Ota, and Hiroyuki Toh for helpful discussion, and Daron M. Standley for critically reading the manuscript. This work was supported by a grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (JST). H.N. was supported by Grant-in-Aid for Scientific Research (B) No. 20370061 from the Japan Society for the Promotion of Science (JSPS).

Received: August 19, 2008
Revised: November 10, 2008
Accepted: November 13, 2008
Published: February 12, 2009

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.L. (1997). Gapped blast and PSI-blast: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Andreeva, A., and Murzin, A.G. (2006). Evolution of protein fold in the presence of functional constraints. Curr. Opin. Struct. Biol. *16*, 399–408.

Bachhawat, P., Swapna, G.V., Montelione, G.T., and Stock, A.M. (2005). Mechanism of activation for transcription factor PhoB suggested by different modes of dimerization in the inactive and active states. Structure *13*, 1353–1363.

Barber, M.J., Neame, P.J., Lim, L.W., White, S., and Matthews, F.S. (1992). Correlation of X-ray deduced and experimental amino acid sequences of trimethylamine dehydrogenase. J. Biol. Chem. *267*, 6611–6619.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. *35*, D301–D303.

Berry, M.B., and Phillips, G.N., Jr. (1998). Crystal structures of Bacillus stearothermophilus adenylate kinase with bound Ap5A, Mg^{2+} Ap5A, and Mn^{2+} Ap5A reveal an intermediate lid position and six coordinate octahedral geometry for bound Mg^{2+} and Mn^{2+}. Proteins *32*, 276–288.

Beuth, B., Pennell, S., Arnvig, K.B., Martin, S.R., and Taylor, I.A. (2005). Structure of a *Mycobacterium tuberculosis* NusA-RNA complex. EMBO J. *24*, 3576–3587.

Brakoulias, A., and Jackson, R.M. (2004). Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins *56*, 250–260.

Carvalho, A.L., Dias, F.M.V., Nagy, T., Prates, J.A.M., Proctor, M.R., Smith, N., Bayer, E.A., Davies, G.J., Ferreira, L.M.A., Romao, M.J., et al. (2007). Evidence for a dual binding mode of dockerin modules to cohesins. Proc. Natl. Acad. Sci. USA *104*, 3089–3094.

Chen, L., Oughtred, R., Berman, H.M., and Westbrook, J. (2004). TargetDB: a target registration database for structural genomics projects. Bioinformatics *20*, 2860–2862.

Davies, J.R., Jackson, R.M., Mardia, K.V., and Taylor, C.C. (2007). The poisson index: a new probabilistic model for protein ligand binding site similarity. Bioinformatics *23*, 3001–3008.

Diamond, R. (1988). A note on the rotational superposition problem. Acta Crystallogr. A *44*, 211–216.

Dias, M.V., Faim, L.M., Vasconcelos, I.B., de Oliveira, J.S., Basso, L.A., Santos, D.S., and de Azevedo, W.F. (2007). Effects of the magnesium and chloride ions and shikimate on the structure of shikimate kinase from *Mycobacterium tuberculosis*. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *63*, 1–6.

Du, Z., Lee, J.K., Tjhen, R., Li, S., Pan, H., Stroud, R.M., and James, T.L. (2005). Crystal structure of the first kh domain of human poly(C)-binding protein-2 in complex with a C-rich strand of human telomeric DNA at 1.7 Å. J. Biol. Chem. *280*, 38823–38830.

Dunn, C.R., Wilks, H.M., Halsall, D.J., Atkinson, T., Clarke, A.R., Muirhead, H., and Holbrook, J.J. (1991). Design and synthesis of new enzymes based on the lactate dehydrogenase framework. Philos. Trans. R. Soc. Lond. B Biol. Sci. *332*, 177–184.

Friedberg, I., and Godzik, A. (2005). Connecting the protein structure universe by using sparse recurring fragments. Structure *13*, 1213–1224.

Garcia-Molina, H., Ullman, J.D., and Widom, J. (2002). Database Systems: The Complete Book (Upper Saddle River, NJ: Prentice Hall).

Gold, N.D., and Jackson, R.M. (2006). Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J. Mol. Biol. *355*, 1112–1124.

Goldstein, R.A. (2008). The structure of protein evolution and the evolution of protein structure. Curr. Opin. Struct. Biol. *18*, 170–177.

Gonzalez, B., Schell, M.J., Letcher, A.J., Veprintsev, D.B., Irvine, R.F., and Williams, R.L. (2004). Structure of a human inositol 1,4,5-trisphosphate 3-kinase: substrate binding reveals why it is not a phosphoinositide 3-kinase. Mol. Cell *15*, 689–701.

Grishin, N.V. (2001a). Fold change in evolution of protein structures. J. Struct. Biol. *134*, 167–185.

Grishin, N.V. (2001b). KH domain: one motif, two folds. Nucleic Acids Res. *29*, 638–643.

Gross, E., Kastner, D.B., Kaiser, C., and Fass, D. (2004). Structure of ero1p, source of disulfide bonds for oxidative protein folding in the cell. Cell *117*, 601–610.

Guilloteau, J.P., Mathieu, M., Giglione, C., Blanc, V., Dupuy, A., Chevrier, M., Gil, P., Famechon, A., Meinnel, T., and Mikol, V. (2002). The crystal structures of four peptide deformylases bound to the antibiotic actinonin reveal two distinct types: a platform for the structure-based design of antibacterial agents. J. Mol. Biol. *320*, 951–962.

Gutteridge, A., and Thornton, J.M. (2005). Understanding nature's catalytic toolkit. Trends Biochem. Sci. *30*, 622–629.

Herter, S., Piper, D.E., Aaron, W., Gabriele, T., Cutler, G., Cao, P., Bhatt, A.S., Choe, Y., Craik, C.S., Walker, N., et al. (2005). Hepatocyte growth factor is a preferred in vitro substrate for human hepsin, a membrane-anchored serine protease implicated in prostate and ovarian cancers. Biochem. J. *390*, 125–136.

Hoff, K.G., Avalos, J.L., Sens, K., and Wolberger, C. (2006). Insights into the sirtuin mechanism from ternary complexes containing NAD^+ and acetylated peptide. Structure *14*, 1231–1240.

Ikura, T., Kinoshita, K., and Ito, N. (2008). A cavity with an appropriate size is the basis of the ppiase activity. Protein Eng. Des. Sel. *21*, 83–89.

Jonassen, I., Eidhammer, I., Grindhaug, S.H., and Taylor, W.R. (2000). Searching the protein structure databank with weak sequence patterns and structural constraints. J. Mol. Biol. *304*, 599–619.

Jones, S., and Thornton, J.M. (2004). Searching for functional sites in protein structures. Curr. Opin. Chem. Biol. *8*, 3–7.

Kawabata, T. (2003). Matras: a program for protein 3D structure comparison. Nucleic Acids Res. *31*, 3367–3369.

Kawabata, T., and Nishikawa, K. (2000). Protein tertiary structure comparison using the markov transition model of evolution. Proteins *41*, 108–122.

Kinjo, A.R., and Nakamura, H. (2007). Similarity search for local protein structures at atomic resolution by exploiting a database management system. BIOPHYSICS *3*, 75–84. Published online December 28, 2007. 10.2142/biophysics.3.75.

Kinoshita, K., and Nakamura, H. (2004). eF-site and PDBjViewer: database and viewer for protein functional sites. Bioinformatics *20*, 1329–1330.

Kinoshita, K., Sadanami, K., Kidera, A., and Go, N. (1999). Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. Protein Eng. *12*, 11–14.

Kobayashi, N., and Go, N. (1997). ATP binding proteins with different folds share a common ATP-binding structural motif. Nat. Struct. Biol. *4*, 6–7.

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol. *323*, 297–307.

Krishna, S.S., and Grishin, N.V. (2004). Structurally analogous proteins do exist! Structure *12*, 1125–1127.

Krishna, S.S., Majumdar, I., and Grishin, N.V. (2003). Structural classification of zinc fingers: survey and summary. Nucleic Acids Res. *31*, 532–550.

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*, 2256–2268.

Lang, R., Braun, M., Sounni, N.E., Noel, A., Frankenne, F., Foidart, J.M., Bode, W., and Maskos, K. (2004). Crystal structure of the catalytic domain of MMP-16/MT3-MMP: characterization of MT-MMP specific features. J. Mol. Biol. *336*, 213–225.

Laronde-Leblanc, N., Guszczynski, T., Copeland, T., and Wlodawer, A. (2005). Autophosphorylation of *Archaeoglobus fulgidus* Rio2 and crystal structures of its nucleotide-metal ion complexes. FEBS J. *272*, 2800–2810.

Lawler, E. (2001). Combinatorial Optimization: Networks and Matroids (New York: Dover) (Originally published in 1976).

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. *8*, 995–1005.

Malikayil, J.A., Burkhart, J.P., Schreuder, H.A., Broersma, R.J., Jr., Tardif, C., Kutcher, L., 3rd, Mehdi, S., Schatzman, G.L., Neises, B., and Peet, N.P. (1997). Molecular design and characterization of an α-thrombin inhibitor containing a novel P1 moiety. Biochemistry *36*, 1034–1040.

Minai, R., Matsuo, Y., Onuki, H., and Hirota, H. (2008). Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. Proteins *72*, 367–381.

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.

Nagano, N., Orengo, C.A., and Thornton, J.M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J. Mol. Biol. *321*, 741–765.

Orengo, C.A., Jones, D.T., and Thornton, J.M. (1994). Protein superfamilies and domain superfolds. Nature *372*, 631–634.

Pattabhi, V., Syed Ibrahim, B., and Shamaladevi, N. (2004). Trypsin activity reduced by an autocatalytically produced nonapeptide. J. Biomol. Struct. Dyn. *21*, 737–744.

Polacco, B.J., and Babbitt, P.C. (2006). Automated discovery of 3D motifs for protein function annotation. Bioinformatics *22*, 723–730.

Porter, C.T., Bartlett, G.J., and Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. *32*, D129–D133.

Ridder, I.S., Rozeboom, H.J., Kalk, K.H., Janssen, D.B., and Dijkstra, B.W. (1997). Three-dimensional structure of L-2-haloacid dehalogenase from *Xanthobacter autotrophicus* GJ10 complexed with the substrate-analogue formate. J. Biol. Chem. *272*, 33015–33022.

Rognan, D. (2007). Chemogenomic approaches to rational drug design. Br. J. Pharmacol. *152*, 38–52.

Russell, R.B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J. Mol. Biol. *279*, 1211–1227.

Schubert, H.L., Raux, E., Brindley, A.A., Leech, H.K., Wilson, K.S., Hill, C.P., and Warren, M.J. (2002). The structure of *Saccharomyces cerevisiae* Met8p, a bifunctional dehydrogenase and ferrochelatase. EMBO J. *21*, 2068–2075.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. (2004). Recognition of functional sites in protein structures. J. Mol. Biol. *339*, 607–633.

Standley, D.M., Toh, H., and Nakamura, H. (2008). Functional annotation by sequence-weighted structure alignments: Statistical analysis and case studies from the Protein 3000 structural genomics project in Japan. Proteins *72*, 1333–1351.

Stark, A., Sunyaev, S., and Russell, R.B. (2003). A model for statistical significance of local similarities in structure. J. Mol. Biol. *326*, 1307–1316.

Stewart, M., Kent, H.M., and McCoy, A.J. (1998). The structure of the Q69L mutant of GDP-Ran shows a major conformational change in the switch II loop that accounts for its failure to bind nuclear transport factor 2 (NTF2). J. Mol. Biol. *284*, 1517–1527.

Stoll, V.S., Simpson, S.J., Krauth-Siegel, R.L., Walsh, C.T., and Pai, E.F. (1997). Glutathione reductase turned into trypanothione reductase: structural analysis of an engineered change in substrate specificity. Biochemistry *36*, 6437–6447.

Tari, L.W., Mattte, A., Pugazhenthi, U., Goldie, H., and Delbaere, L.T. (1996). Snashot of an enzyme reaction intermediate in the structure of the ATP-Mg^{2+}-oxalate ternary complex of Escherichia coli PEP carboxykinase. Nat. Struct. Biol. *3*, 355–363.

Tari, L.W., Matte, A., Goldie, H., and Delbaere, L.T. (1997). Mg^{2+}-Mn^{2+} clusters in enzyme-catalyzed phosphoryl-transfer reactions. Nat. Struct. Biol. *4*, 990–994.

Taylor, W.R. (2007). Evolutionary transitions in protein fold space. Curr. Opin. Struct. Biol. *17*, 354–361.

Wallace, A.C., Borkakoti, N., and Thornton, J.M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. application to enzyme active sites. Protein Sci. *6*, 2308–2323.

Wangikar, P.P., Tendulkar, A., Ramya, S., Mali, D.N., and Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J. Mol. Biol. *326*, 955–978.

Watts, D.J., and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. Nature *393*, 440–442.

Westbrook, J., Ito, N., Nakamura, H., Henrick, K., and Berman, H.M. (2005). PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics *21*, 988–992.

Whitlow, M., Arnaiz, D.O., Buckman, B.O., Davey, D.D., Griedel, B., Guilford, W.J., Koovakkat, S.K., Liang, A., Mohan, R., Phillips, G.B., et al. (1999). Crystallographic analysis of potent and selective factor Xa inhibitors complexed to bovine trypsin. Acta Crystallogr. D Biol. Crystallogr. *55*, 1395–1404.

Wolfson, H.J., and Rigoutsos, I. (1997). Geometric hashing: an overview. IEEE Comput. Sci. Eng. *4*, 10–21.

Xiao, T., Takagi, J., Coller, B.S., Wang, J.H., and Springer, T.A. (2004). Structural basis for allostery in integrins and binding to fibrinogen-mimetic therapeutics. Nature *432*, 59–67.

Xie, L., and Bourne, P.E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc. Natl. Acad. Sci. USA *105*, 5441–5446.

