PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

PDB-scale analysis of known and putative ligand-binding sites with structural sketches

Jun-Ichi Ito,[1,2] Yasuo Tabei,[3] Kana Shimizu,[2] Kentaro Tomii,[1,2]* and Koji Tsuda[2,3]

[1] Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8568, Japan

[2] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan

[3] Minato Discrete Structure Manipulation System Project, ERATO, Japan Science and Technology Agency, Sapporo 060-0814, Japan

ABSTRACT

Computational investigation of protein functions is one of the most urgent and demanding tasks in the field of structural bioinformatics. Exhaustive pairwise comparison of known and putative ligand-binding sites, across protein families and folds, is essential in elucidating the biological functions and evolutionary relationships of proteins. Given the vast amounts of data available now, existing 3D structural comparison methods are not adequate due to their computation time complexity. In this article, we propose a new bit string representation of binding sites called structural sketches, which is obtained by random projections of triplet descriptors. It allows us to use ultra-fast all-pair similarity search methods for strings with strictly controlled error rates. Exhaustive comparison of 1.2 million known and putative binding sites finished in ~30 h on a single core to yield 88 million similar binding site pairs. Careful investigation of 3.5 million pairs verified by TM-align revealed several notable analogous sites across distinct protein families or folds. In particular, we succeeded in finding highly plausible functions of several pockets via strong structural analogies. These results indicate that our method is a promising tool for functional annotation of binding sites derived from structural genomics projects.

Proteins 2011; 00:000–000.

Key words: structure and function; ligand-binding site; neighbor search algorithm; pocketome.

INTRODUCTION

As the number of known protein structures derived from structural genomics projects increases, structural comparison and classification have become essential processes for understanding the functional and evolutionary relationships of proteins. A number of tools for comparing protein 3D structures[1–3] have been developed, which can efficiently find global structural similarities between proteins. In databases such as SCOP[4] and CATH,[5] structures are divided into relatively large units, that is, domains, and then hierarchically classified according to their global structure and sequence. However, proteins that do not exhibit any overall sequence or structural similarity can share common functions. Well-known instances include the Ser-His-Asp catalytic triad found in serine proteases[6,7] and the P-loop containing nucleotide-binding proteins.[8] In these cases, only a few key residues close to the ligand are highly conserved, whereas their folds are distinct from one another. Given that most proteins exhibit their functions through interactions with other molecules (so-called ligands), most protein functions can be characterized by ligand-binding sites, that is, a set of residues directly involved in interaction with a ligand. Therefore, the pairwise comparison of ligand-binding sites, across different families or folds, is an appropriate approach to gaining functional and evolutionary knowledge about proteins.

Today, hundreds of thousands of ligand-binding sites can be found in the protein data bank (PDB).[9] In addition to these known sites, several computational search methods are available for predicting potential ligand-binding sites. They include structure-based methods,[10–17] sequence-based methods,[18,19] and hybrids of these.[20,21] All-pair similarity searches between such known and potential binding sites are computationally demanding, but useful in the prediction of protein functions. Furthermore, these may provide important insights into the rules of protein-ligand

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: JSPS KAKENHI; Grant numbers: 21680025 and 23500374; Grant sponsor: FIRST program.
*Correspondence to: Kentaro Tomii, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan. E-mail: k-tomii@aist.go.jp.
Received 30 December 2010; Revised 13 October 2011; Accepted 18 October 2011
Published online 29 October 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.23232

interactions, which could be exploited in fast screening of target proteins for drug discovery.

Comparisons of ligand-binding sites differ from the alignment of global structures, because a binding site consists of a small number of amino acids that are not necessarily continuous in the sequence and do not need to be located in the same domain or chain. Several algorithms have been proposed[22–29] in the last decade to enable such noncontinuous protein local regions to be compared. Typically, all of these methods follow three common steps. In the first step, binding sites are decomposed into simplified objects such as triangle meshes,[22] tetrahedrons,[27] chemically important representative points,[23,24,26] or all atoms coordinates[25,28,29] to reduce the computational cost of the following steps. The objects often have labels representing the geometric and/or physicochemical properties of amino acid residues. In the second step, 3D alignment algorithms are used to obtain the best possible superposition of objects. Examples include clique (i.e., maximum common sub-graph) detection,[22,23,26,29] geometric matching,[25,28] geometric hashing,[24,30] or geometric indexing.[27] Finally, a similarity score between two binding sites, such as a *P*-value or a Tanimoto-index, is calculated based on the number of matched objects.

Gold and Jackson[25] performed an all-against-all comparison of 33,168 known binding sites and detected several unexpected similarities across distinct proteins. Minai *et al.*[28] conducted an all-against-all comparisons of 48,347 potential ligand-binding sites that were predicted by a binding site prediction program PASS[12] and demonstrated the utility of comparing binding sites for drug design. More recently, Kinjo and Nakamura[31] conducted an all-against-all comparison of over 180,000 known ligand-binding sites found in the PDB and clustered them into about 3000 well-defined structural motifs. The results of clustering analysis suggested that the majority of chemical molecule binding sites are confined within the same protein families except for nucleotide-binding sites and ion-binding sites.

However, these methods can only be applied to small datasets such as representative protein chains or known ligand-binding sites in PDB, primarily because of time complexity. For example, Minai *et al.*[28] used a grid computing system for a comprehensive search, and it took 29 days; the total CPU time was about 2 years. The bottleneck of calculation with these methods lies in the second step: iterative optimization to obtain maximum 3D alignment. To overcome this problem, alignment-free methods of comparing fingerprints have been proposed,[32–35] in which binding sites are projected onto numerical fingerprints, that is, fixed-length high-dimensional vectors, and then similarities are measured by comparing their corresponding vector components. The main advantage of these methods is that they can be used to measure the similarities between binding sites without having to find 3D alignments in advance. These methods are, however, not sufficiently scalable for millions of data points, because the cost of brute-force pairwise similarity calculations in high-dimensional vector space is still computationally demanding.

We propose a fast alignment-free method for comparing huge amounts of protein-ligand binding sites, in which binding sites are mapped onto high-dimensional feature spaces based on their physicochemical properties and geometric features. The core to this algorithm involves structural sketches: bit strings created by random projections of triplet descriptors. It is well evidenced by, for example, BLAST[36] and BWA,[37] that string processing can be much more scalable than 3D structural objects. Once binding sites are converted to bit strings, an ultra-fast algorithm called multiple sorting is used to find all similar pairs. Because our algorithm involves random projection, a fraction of neighboring pairs remains undetected. However, one can theoretically bind the fraction and keep it below a negligibly small value, for example, 10^{-4}.

To demonstrate performance and scalability with our method, we conducted all-pair similarity searches for 1,201,461 known and potential binding sites. Using eight different types of structural sketches, we found about 88 million similar binding-site pairs in about 30 h on a single core. We applied TM-align[3] to all pairs other than pocket–pocket pairs, and selected 3.5 million well-aligned pairs. Of these, around 560,000 pairs were analogous, that is, CATH codes of two proteins were different. Here, we report several notable analogous pairs of ligand-binding sites shared by different superfamilies or folds, and demonstrate the versatility of similar binding sites. Even though manual investigation of all analogous pairs was inconceivable, our initial analysis focusing on nucleotide-containing ligands already identified several pockets whose biological functions were predicted via strong structural analogies. In addition, a global similarity network of binding sites could be drawn from our results, which contributed to our investigating the diversity of a protein family in terms of ligand-binding profiles and elucidating the relationships between ligands. The time efficiency of our algorithm is unprecedented, and such a fast algorithm will certainly contribute to enabling large amounts of data produced from structural genomics projects to be analyzed.

MATERIALS AND METHODS

Ligand-binding site dataset

We prepared two sets of ligand-binding sites (see Fig. 1): known binding sites obtained from protein-ligand complexes and potential binding sites predicted with a pocket detection program.

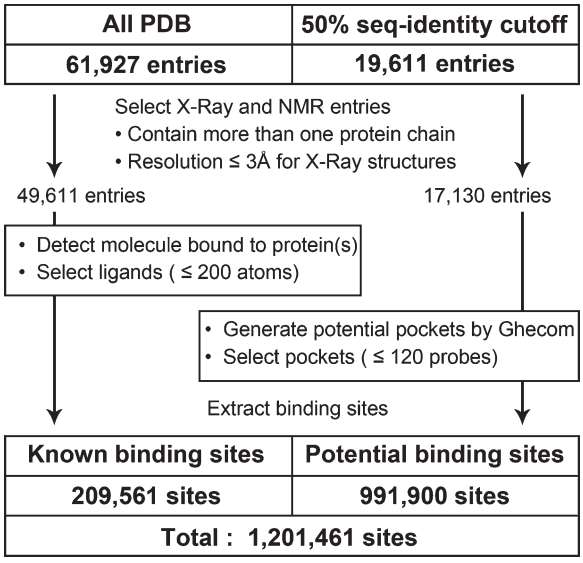

Figure 1

Overview of current binding site dataset. From known protein structures, 209,561 known binding sites were obtained, and 991,900 putative binding sites were predicted using a pocket detection program, *ghecom*. In total, we used 1,201,461 known and potential binding sites in this study.

Known ligand-binding site dataset

As of April 2010, we selected 49,611 entries (X-rays with resolution $\leq$3.0 Å and MODEL1 of all NMR) from the PDB that contained at least one protein chain. Out of the entries, we only focused on hetero (HET) molecules up to 200 atoms, excluding water molecules, nucleic acids (DNA/RNA), peptides, and large molecules of over 200 atoms. A binding site corresponding to a ligand was defined as a collection of all amino acids, each of which had at least one heavy atom lying within a distance of 5.0 Å from at least one heavy atom of the ligand. To exclude low-affinity and incidental ligand-binding sites, we discarded all binding sites with fewer than five residues. In total, 209,561 binding sites were obtained. In the following, we refer to them as known ligand-binding sites. We did not apply any sequence identity cutoff to create this dataset to exploit the data of the same protein bound to different ligands (or even the same ligands in different ways) and to demonstrate the capabilities of our method for PDB-scale analysis.

Potential ligand-binding site dataset

Potential regions are predicted by *ghecom*,[15] which can efficiently identify ligand-binding regions by detecting pockets on the protein surface. The main reason we chose this program from the various ones available was due to its high-speed and flexibility. We prepared a non-redundant PDB subset (50% sequence identity cutoff and resolution better than 3.0 Å for X-rays) that contained 17,130 entries, and for each of them, we iterated *ghecom* four times as changing the values [3 Å, 4 Å, 5 Å, and 6 Å] of the parameter, R_{large}. The size of each pocket in this program was measured by the number of small spheres with a radius of 1.87 Å filling up a pocket. We chose pockets that were smaller than 200 spheres, which approximately matched the largest ligands considered in the known sites. If a protein had more than 200 pockets, we selected the largest 200 pockets. The binding site corresponding to a pocket was defined as a collection of all amino acids, each of which had at least one heavy atom within a distance of 5.0 Å from one of the spheres. Discarding small binding sites containing at most four amino acids, 991,900 potential binding sites were obtained. The computational time required for processing all entries was around 10.3 h on a 4-core CPU machine (Intel Xeon 2.93 GHz). In total, we collected 1,201,461 known and potential ligand-binding sites.

Triplet descriptors of ligand-binding sites

The basic idea underlying our method was to represent binding sites by a vector of the weighted occurrence frequencies of descriptors and then map it to bit strings for the fast neighbor search algorithm. We used triplet descriptors, which are commonly used in binding site analysis and pharmacology.[28,30,34,38] We basically adopted the triplet descriptors of FUZCAV, but we added several extensions to their method. (1) FUZCAV uses a fixed set of descriptors. However, it is highly task-dependent on what kinds of properties should be represented in the descriptors. Thus, we prepared several kinds of descriptors to capture different aspects of binding sites. (2) Instead of simple frequencies of descriptors, we used weighted frequencies depending on the distance to the ligand. (3) FUZCAV uses a special similarity metric counting the number of exact matches of frequencies but is too sensitive against small frequency changes. Instead, we used cosine similarity, which is more robust. The procedure for constructing feature space is described as follows and schematically outlined in Figure 2.

Given a binding site, we encode each amino acid to a set of labels. To diversify our descriptors, we use multiple label sets for encoding, each of which lead to a different feature space. As summarized in Table I, we used eight sets of size 4 out of the following 10 labels: Char+, Char−, Hydro, Aroma, Small, H-d, H-a, SSE(α), SSE(β), and SSE(turn). The first 7 labels reflect the intrinsic physicochemical properties summarized in Table II. The last three labels indicate that the residue is included in α-helix, β-strand, and hydrogen-bonded turn, respectively, where the secondary structure is determined by the DSSP program[39] using the complete structures.

In a specific encoding, an amino acid is labeled by a subset of four uppercase letters (A, B, C, and D). We

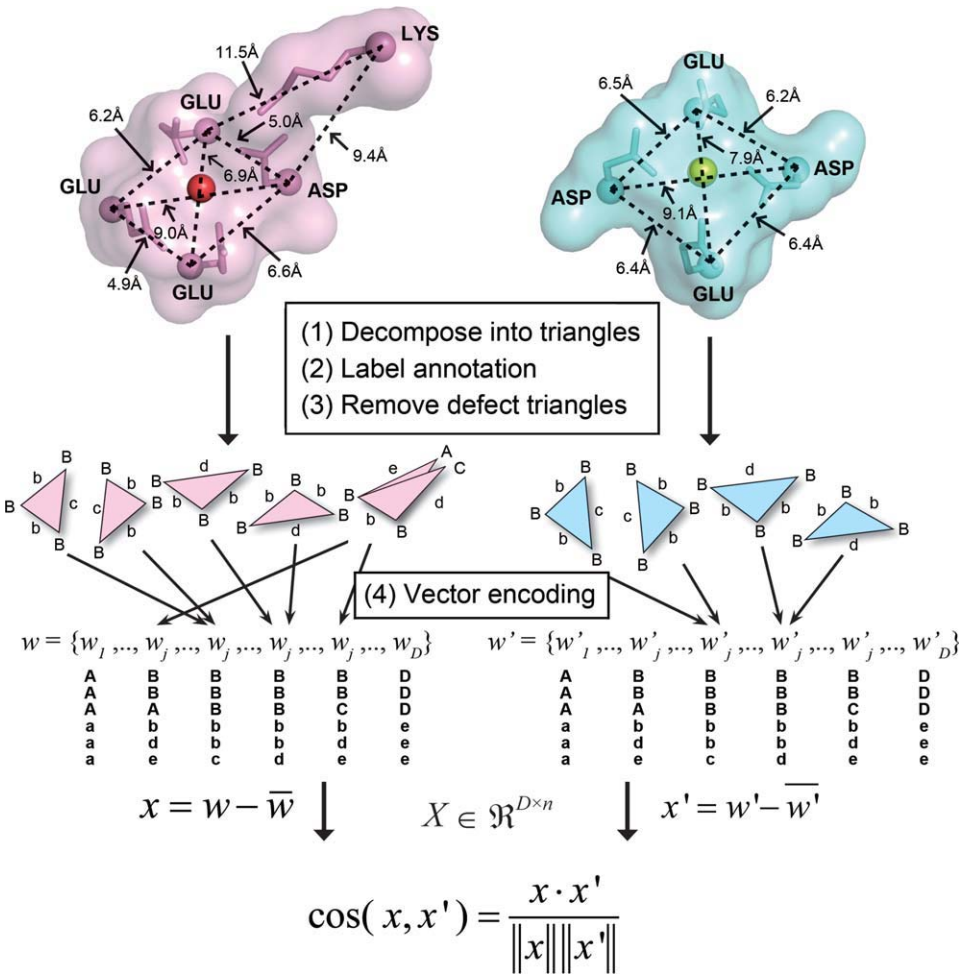

Figure 2

Schematic of our method of comparing binding sites. In our method, (1) each binding site is decomposed into all possible triangles comprising three amino acids, (2) and labels are assigned to the triangles according to their physicochemical properties and geometrical properties. (3) Defect triangles that contain at least one nonlabeled vertex or edge are removed. (4) Then, one binding site can be encoded as a high-dimensional feature vector that describes the frequency of the labeled triangles. Finally, the similarity between two binding sites can be measured as the cosine of two vectors. The two metal ion binding sites, an Mg-binding site and a Fe-binding site, are shown in cyan and pink, respectively. The dashed lines indicate the C_α—C_α atomic distances between two residues. In this instance, vertices were labeled based on the amino acid encoding set 1.

consider a graph of amino acids where residue pairs within 13.6 Å are connected by edges. Each edge is labeled according to the C_α—C_α distance: Interval [1 Å, 13.6 Å] is divided into five subintervals, each of which is assigned a lowercase label, (a, b, c, d, or e; see Table III).

In the graph of each binding site, we detect all cliques of size 3, each of which corresponds to a triangle of amino acids. If at least one node in a clique has multiple labels, it is further unfolded to multiple cliques. For

Table I
Table of Vertex Labels and Physicochemical Properties for Eight Encoding Sets

Vertex label	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
A	Char+	Char+	Char+	Char+	Char+	Hydro	Hydro	Hydro
B	Char−	Char−	Char−	Char−	Char−	Aroma	Small	Small
C	Hydro	Hydro	Aroma	H-d	SSE(α)	SSE(α)	SSE(α)	SSE(α)
D	Aroma	Small	Small	H-a	SSE(β)	SSE(β)	SSE(β)	SSE(T)

SSE(α), SSE(β), and SSE(T) indicate α-helix and β-strand, and turn, respectively.

Table II
Seven Physicochemical Properties and Classified Amino Acids into the Group(s) Based on Their Side-Chain Characteristics

Amino acids	Physicochemical property
K, R	Positive charge (Char+)
D, E	Negative charge (Char−)
A, F, I, L, M, P, V, C, K, T	Hydrophobicity (Hydro)
F, Y, W, H	Aromatic (Aroma)
G, S, C, A, T, N	Small (Small)
N, R, Q, H, K, S, T, W, Y	H-bond donor (H-d)
N, D, Q, E, H, S, T, Y	H-bond acceptor (H-a)

Table III
Edge Labels for Each $C_\alpha - C_\alpha$ Atomic Distance Interval

Edge label	$C_\alpha - C_\alpha$ atomic distance interval (Å)
a	1.0–4.8
b	4.8–7.0
c	7.0–9.2
d	9.2–11.4
e	11.4–13.6

The range of all intervals, except for the first interval, were set to 2.2 Å.

example, if the nodes have labels ([A, B], C, [A, D]), it is replaced with four cliques with labels (A, C, A), (A, C, D), (B, C, A), and (B, C, D). Taking into account edge labels as well, we have 1540 distinct cliques after resolving rotation invariance.

Finally, a feature vector of a binding site is obtained as a 1540 dimensional vector of weighted occurrence frequencies of cliques in the graph. The weight of each clique is determined via the distance to the ligand. For a known site, the distance from an amino acid to the ligand is defined as the distance to the nearest heavy atom in the ligand. Then, the distance from the clique to the ligand is defined as the average of the three distances. For a putative site, we substitute the heavy atoms by the central points of the spheres filled in by *ghecom*. Let j denote the index of the clique, $j = 1,\ldots,1540$, and p_j denote the number of occurrences. The jth element of the feature vector is calculated with the following Gaussian kernel:

$$w_j = \sum_{i=1}^{p_j} \exp\left[-\frac{\overline{d}_i^2}{2\sigma^2}\right],$$

where $\overline{d}_i$ represents the distance between the clique and the ligand. Parameter σ is the standard deviation of the Gaussian kernel, which is set to 4 Å. This Gaussian kernel is used to emphasize the contribution of close cliques. Note that a similar weighting scheme was used in another study.[21] Centralizing each element to have a zero mean (i.e., $x = w - \overline{w}$), the similarity between two binding sites can be defined as a cosine between two centralized vectors:

$$\cos(x, x') = \frac{x \cdot x'}{||x||||x'||}, \quad x \in \Re^D.$$

Feature extraction for 1.2 million sites was finished in ~1 h on a single core.

Quality of similarity metrics

Our method was compared with FUZCAV based on the different amino acid encodings in Table I. The performance was evaluated with a benchmark set mimicking "set 3" in a paper on FUZCAV.[34] We collected 871 binding sites of protein kinases with EC numbers, 2.7.1[0-3].* and 2.7.99.* from an up-to-date sc-PDB dataset that contained 8187 ligand-binding sites. The collected 871 sites are designated as the positive class, and the rest is defined as the negative class. Taking one site in the positive class as a query, we computed the similarity metric from the query to all 8187 binding sites. The ranking of binding sites and their associated positive/negative labels produces one receiver operating characteristic (ROC) curve. Ideally, all positive sites should appear before all negative ones. We computed an average ROC curve taking all positive sites as queries. Figure 3 plots the average ROC curves for the eight encodings and the two similarity metrics, our metric and FUZCAVs metric. In all cases, our method outperformed FUZCAV by a large margin, demonstrating the substantial advantages of our extensions.

Our metric was also evaluated with another dataset with a larger variety in the task of detecting physicochemical- and geometrical-similar binding-site pairs in structurally dissimilar proteins. We sampled a small set of binding sites from the reduced list of PDB proteins with 30% sequence identity and a TM-score with < 0.4 cutoff. Of the binding sites for the following 13 ligands [calcium ion (CA), iron/sulfur cluster (SF4), α-D-mannose (MAN), α-D-glucose (GLC), β-D-galactose (GLA), adenosine-5′-triphosphate (ATP), adenosine-5′-diphosphate (ADP), guanosine-5′-triphosphate (GTP), guanosine-5′-diphosphate (GDP), nicotinamide-adenine-dinucleotide (NAD), NADP nicotinamide-adenine-dinucleotide phosphate (NAP), flavin-adenine dinucleotide (FAD), and protoporphyrin ix containing Fe (HEM)], we randomly sampled up to 20 sites per ligand. The ground-truth pair sets were obtained at three cutoffs, that is, the match score ≥ 35, ≥ 40, and ≥ 45, by using

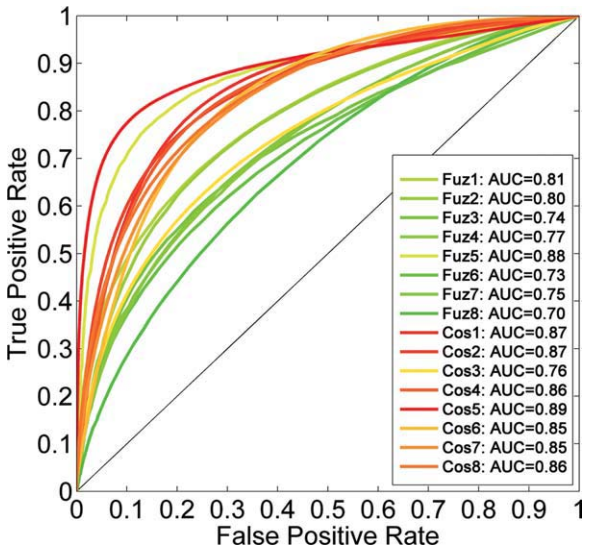

Figure 3
Average of ROC curves for searching binding sites with ligand-binding sites of sc-PDB.

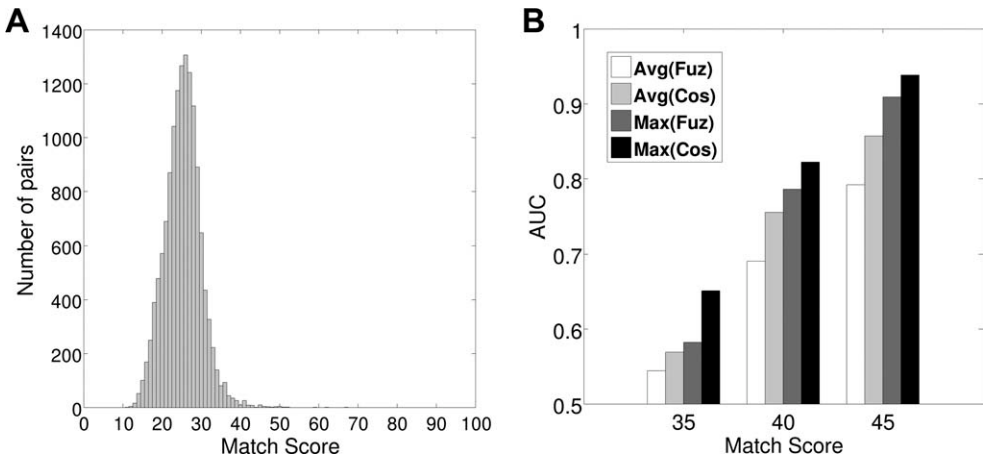

Figure 4

A: Distribution of match scores calculated by SiteEngine for all-pairs of 167 binding sites. **B**: Average and maximum AUC scores of cosine metric and FUZCAVs metric with different match score cutoffs.

SiteEngine,[24] which is a well-known method of finding similarity in binding sites as well as in catalytic sites. Similarities between any two sites are represented by the match score, ranging from 0 to 100 (a higher score indicates a closer match), in SiteEngine. The distribution of scores for the dataset is given in Figure 4(A). As there were fewer than 20 available binding sites for certain ligands and SiteEngine was not able to handle some small sites such as CA or SF4 binding sites, the total number of binding sites turned out to be 167 (Supporting Information Table SI). Our cosine metric and FUZCAVs metric were computed for all 13,861 pairs, and it was matched against the ground-truth sets to yield the ROC curves. We calculated average and maximum area under the curve (AUC) scores for the eight encodings and the two metrics [Fig. 4(B)]. Our metric was fairly successful, where the maximum AUC score was around 0.94 for the tighter ground-truth set (match score $\geq$ 45). This indicates that our metric is sufficiently sensitive to capture physicochemical- and geometrical-similar binding-site pairs.

Our metric was further evaluated with the above dataset in the task of detecting similar site pairs that bind the same ligand in dissimilar proteins. Of the 13,861 pairs, 1130 pairs with the same ligands, that is, the same HETATM recode, were treated as the ground-truth pairs. As well as FUZCAVs metric, our method was also compared with SiteEngine in this benchmark (Fig. 5). Similarly, our method outperformed FUZCAV in all encodings, and in most encoding types, the results for our cosine metric were comparable with that of SiteEngine. Note that our metric demonstrated the best performance (AUC = 0.62) in encoding type 3. Our performance and that of FUZCAV generally decreased compared with those of the previous benchmark. This is because the different ligands in the benchmark could share similar sites, such as the adenine dinucleotide-binding regions, shown below, that are common to FAD and NAD. The results indicate that our metric is more sensitive in detecting similar sites that bind the same ligand in dissimilar proteins than SiteEngine.

Furthermore, our method was compared with Geometric Indexing with Refined Alignment Finder (GIRAF),[27] an accurate geometric hashing-like algorithm for finding similar binding sites. GIRAF has been applied to an exhaustive comparison of all small-molecule binding sites

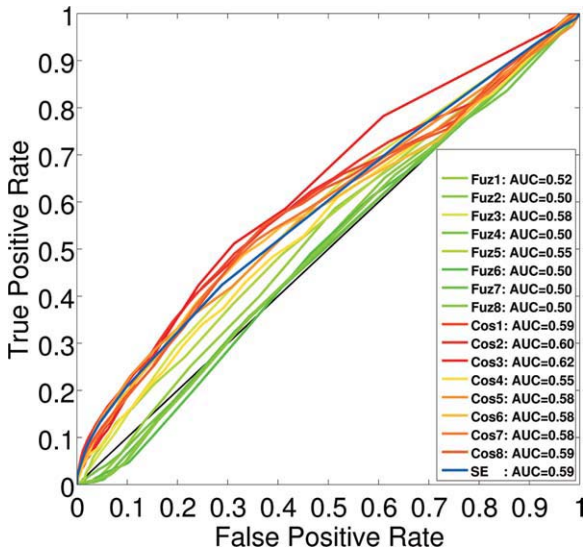

Figure 5

ROC curves for all-against-all comparison of 167 binding sites.

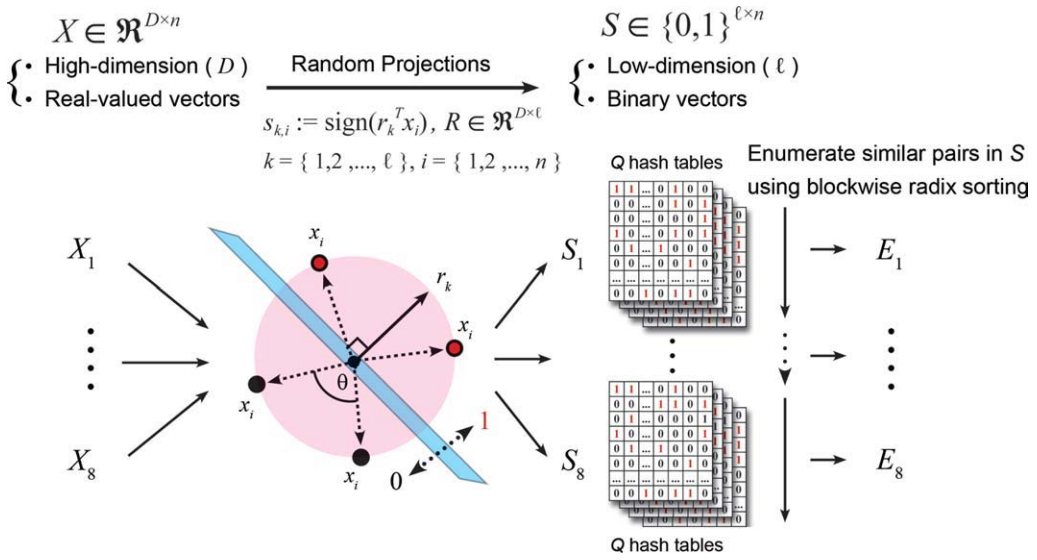

Figure 6

Schematic of SketchSort algorithm. In this algorithm, high-dimensional real-valued vectors are converted to low-dimensional binary strings of discrete symbols by locality sensitive hashing. The algorithm then finds close pairs using a method of blockwise masked radix sorting. These procedures are repeated for individual feature spaces. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

observed in PDB, which is the largest scale study for binding site comparisons as far as we know. The results of comparisons have been compiled as the GIRAF-database.[31] From the studies by Shulman-Peleg *et al.*,[24] we obtained a diverse dataset containing 126 binding sites. Out of these, we discarded protein–protein interfaces and empty binding sites in apo-form, and then selected 81 known small-molecule binding sites. Each of them was searched as a query against the GIRAF-database, and in total, 60,134 hits with a GIRAF-score of $\geq$ 10 were detected. Then, for each of the 81 sites as a query, the cosine similarities were also calculated against the up-to-date PDB (as of January, 2011). The pair coverage for the GIRAF-database was calculated with varying cutoffs of both methods (Supporting Information Fig. S1). Most of the hits with high GIRAF-scores were successfully detected with high cosine values. However, we found many similar site pairs were only detected by our method in comparison with the GIRAF-database. Basically, the GIRAF-database returned hits with a GIRAF-score of $\geq$ 10, which can be considered to be the lower bound for similar sites on GIRAF. Therefore, sites detected with high cosine similarities and not included in the GIRAF results for each query can be regarded as unique hits with our method. We found 4457 such hits with a cosine similarity of $\geq$ 0.85 as unique hits. For example, the cosine similarity for a similar pair of an ATP-binding site of 1AYL (query) and a GNP-binding site of 2ATX (hit) was 0.88 (Supporting Information Fig. S2A). The cosine for another similar pair of a FAD-binding site of 1B4V (query) and a FAD-binding site of 1JNR (hit) was also 0.88 (Supporting Information Fig. S2B).

Structural sketches

Our goal is to enumerate all similar pairs whose cosine is ε or larger,

$$E^* = \{(i,j) | \cos(x_i, x_j) \geq \varepsilon, i < j\}.$$

However, a brute-force pairwise cosine calculation takes $O(n^2)$ time, which is prohibitively slow. We used a novel neighbor search algorithm called SketchSort,[40] which first converts high-dimensional real-valued vectors to binary strings called "sketches" (strings of discrete symbols) through locality sensitive hashing (LSH),[41] and then finds close pairs using multiple sorting (Fig. 6).

LSH maps high-dimensional data to a binary string, that is, $\Re^D \rightarrow {}^{\ell}$, as preserving the original neighborhood information. LSH is implemented via random projections.[42] Let $R \in \Re^{D \times \ell}$ be a random matrix consisting of i.i.d. samples from the standard normal distribution of mean 0 and standard deviation 1. The projection is defined as:

$$s_{ik} := \mathrm{sign}(r_k^T\ x_i),$$

where s_{ik} is the kth bit of the ith string, r_k is the kth column of R and sign(t) produces 1 if $t > 0$ and 0 otherwise. The cosine similarity is approximately preserved as the Hamming distance due to the relationship:

$$\Pr(s_{ik} \neq s_{jk}) = \frac{\theta_{ij}}{\pi},\ \forall k,$$

where θ_{ij} is the angle between two feature vectors x_i and x_j being compared:

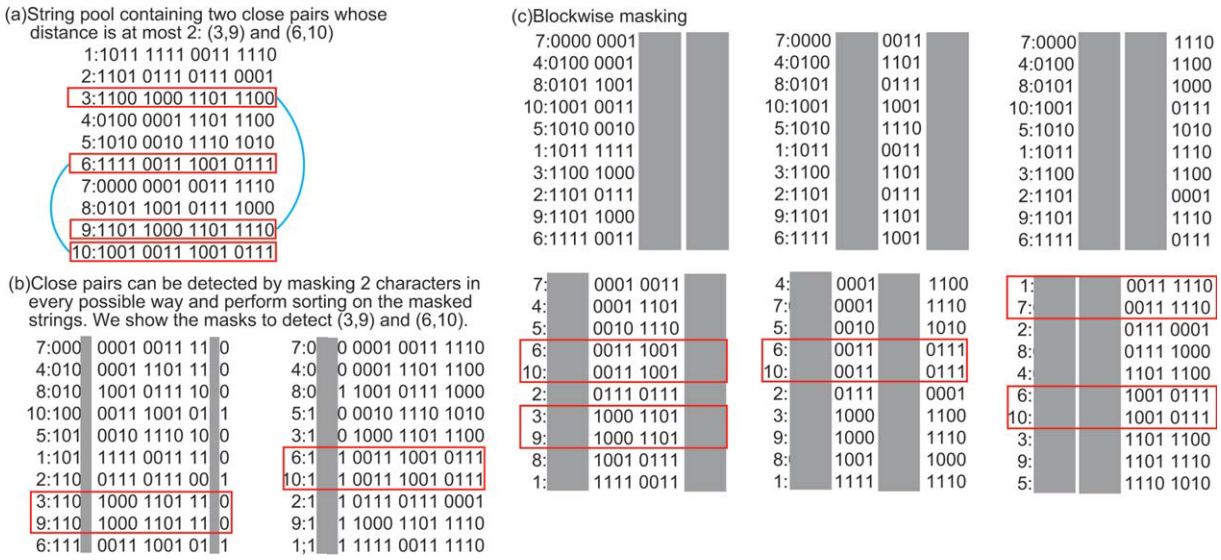

Figure 7

Schematic of multiple sorting. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\theta_{ij} = \arccos\left(\frac{x_i^T\, x_j}{||x_i||||x_j||}\right).$$

This relationship guarantees that the expected value of Hamming distance HamDist(s_i, s_j) is a monotonically decreasing function of cos(x_i, x_j). The larger the number of hash bits (ℓ), the more faithfully the original cosine similarity is preserved. In other words, in the limit $\ell \to \infty$, the Hamming distance divided by ℓ converges to the angle divided by π. Bit strings derived from our feature vectors are called structural sketches in the rest of this article.

Detection of neighboring pairs by multiple sorting

Once the data points are represented by bit strings, we can use ultra-fast methods such as the multiple sorting method (MSM)[43] to find all similar pairs in terms of the Hamming distance. Let us quickly review MSM. First, the strings are divided into b blocks. Then, d blocks are masked and radix sort is applied to the remaining part of strings to find exactly matching pairs in the remaining part (Fig. 7). By applying masks in all possible ways, every close string pair within distance d can be found in at least one masking pattern. As false positives (string pairs whose distances are larger than d) are also detected by sorting, all detected pairs are verified by calculating the actual Hamming distances, and pairs with distances longer than d are discarded. A crucial point is that the sorting operations can be done in linear time. Distance calculation is limited to the strings qualifying for sorting-based filtering.

Pipelining the two methods of LSH and MSM, one can basically find similar binding-site pairs. Nevertheless, fundamental questions including the following remain unsolved: (1) How does the Hamming distance threshold correspond to the cosine distance? (2) Because of random projection, some fraction of neighboring pairs is left undetected. What is the fraction of undetected pairs? A mathematical explanation is necessary to use our methodology with a theoretical guarantee. In our previous theoretical paper, we proposed a method called SketchSort that creates multiple short bit strings with LSH, and MSM is applied multiple times. We thoroughly analyzed the mathematical properties of SketchSort including the relation between the cosine and Hamming distances and the fraction of missing pairs. Essential aspects are explained below.

Missing pair ratio

SketchSort proceeds as follows: For relatively small ℓ, strings of length ℓ are created from the input data with LSH. This is repeated Q times, yielding Q individual string pools. Then, similar pairs within distance d are independently found in each pool and merged. For the merged set of pairs, actual cosine similarities are computed and those beyond ε are reported.

As SketchSort uses random projections, it misses a certain number of neighboring pairs. However, one can strictly control the fraction of missing pairs (missing pair ratio) under a small value as follows. Let us describe the pairs found in the qth string pool as

$$E_q = \left\{(i,j)|\mathrm{HamDist}(s_i^q, s_j^q) \le d,\ i<j\right\}.$$

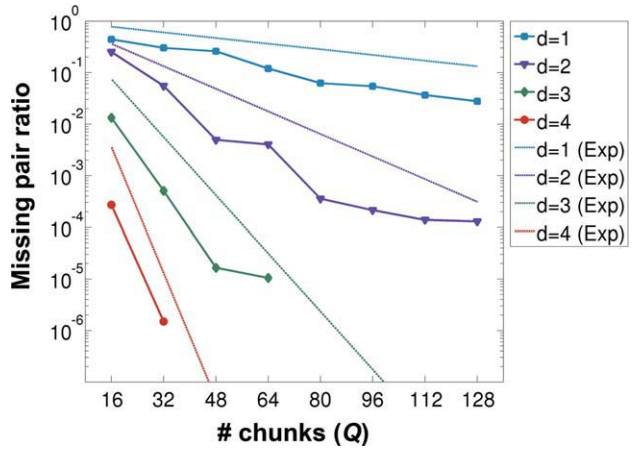

Figure 8

Empirical ratio of missing edges. Solid lines indicate observed values of the missing edge ratio and broken lines indicate the expected values (see text). Blue, green, and red lines indicate the results with $d = 1$, $d = 2$, and $d = 3$, respectively.

The output sets are merged into one, $E = E_1 \cup E_2 \cup \ldots \cup E_q$. Given the true pair set E^* and our solution E, the set of false negative pairs F is defined as

$$F = \{(i,j) | (i,j) \notin E,\ (i,j) \in E^*\},$$

where each pair in F has a Hamming distance larger than d in all string pools. The fraction of missing pairs is defined as $|F|/|E^*|$, whose expectation can be bounded as

$$E\left[\frac{F}{E^*}\right] \leq \left(1 - \sum_{k=0}^{d} \binom{1}{k} p^k (1-p)^{1-k}\right)^Q,$$

where p is an upper bound of the noncollision probability for neighbors. For cosine LSH, p is described as

$$p = \frac{\arccos(\varepsilon)}{\pi}.$$

How to use SketchSort

To use SketchSort, one specifies the cosine threshold, ε, and the required missing edge ratio. The internal parameters (d, b, ℓ, Q) are set such that the theoretical bound meets the required ratio. In the following experiments, we determined $\ell = 32$, $d = 3$, and $b = 6$, unless stated otherwise. The number of chunks Q was set to the smallest possible value keeping the theoretical bound below the required value.

RESULTS AND DISCUSSION

Benchmarking SketchSort

To observe the tightness of the bound, we compared the actual values of the missing pair ratio with the bound using a small dataset ($n = 50{,}000$) of ligand-binding sites (amino acid encoding type 1) by random sampling. As a result of a brute-force search (i.e., all-against-all cosine calculation), 665,053 pairs with cosine similarity ≥ 0.85 were obtained as the set of true pairs E^*. Next, we applied SketchSort to the same dataset with different values of $d = [1, 2, 3, 4]$ and $Q = [16, 32, 48, 64, 80, 96, 112, 128]$. The actual values of the missing pair ratio were compared with the theoretical bound in Figure 8. We verified that the fraction of missing pairs was consistently smaller than the theoretical bound by a small margin.

Next, we measured the computational time for SketchSort against the number of vectors n. In this experiment, n was set to various values [800, 1600, ..., 25,600, 50,000, 100,000], and the vectors were randomly selected. We tested two sets of parameters of $[d = 3, Q = 64]$ and $[d = 4, Q = 32]$, so that the missing pair ratio was smaller than 10^{-4}. As seen in Figure 9, SketchSort has much better scalability than brute-force. At $n = 100{,}000$, SketchSort is already over 10 times faster than the brute-force search. This time difference will be much larger for greater amounts of data.

Enumeration of similar binding site pairs

SketchSort was applied to 1.2 million binding sites with the cosine threshold set to 0.85 and the missing pair ratio set to 10^{-4}. Computation was performed with a single core Intel Xeon 2.93 GHz processor. The computational time and the number of discovered pairs for each amino acid encoding type are listed in Table IV. According to encoding types, the computational time varied from 1.45 to 5.80 h. Merging all paired sets across encodings and removing duplications, we obtained a total of 88,194,290 pairs.

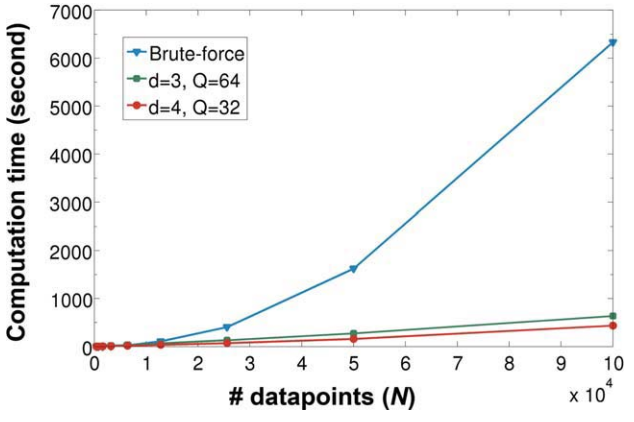

Figure 9

Evolution of execution time as a function of number of data points. Blue, green, and red lines indicate the Brute-force, SketchSort $[d = 3, Q = 64]$, and SkechSort $[d = 4, Q = 32]$, respectively.

Table IV
Number of Enumerated Binding Site Pairs and Execution Time for Each Feature Space

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E
No. of detected pairs	12,257,660	8,835,622	7,489,147	15,106,785	18,902,089	33,405,625	34,072,315	29,431,681	88,194,290
Execution time	13870.8 (s)	13470.1 (s)	5224.9 (s)	17466.7 (s)	20912.4 (s)	11615.8 (s)	6165.66 (s)	19341.6 (s)	30.01 (h)

About 47% of similar pairs were putative-putative pocket pairs that were only comprised of potential sites, which were not analyzed further in this study (Fig. 10). In the remaining pairs, we focused on 16,775,641 pairs of binding sites, both of which resided in the domains annotated with CATH code.[5] Then, we applied TM-align to every site pair to remove totally unalignable pairs (TM-score of < 0.3 or aligned residue length of < 5), yielding 3,469,036 qualified pairs. Although TM-align was originally developed for detecting similarities in protein domains, it has been applied to align two small sets of noncontinuous amino acid residues of protein–protein binding sites.[44] We applied TM-align to align sets of binding site residues that were concatenated in sequential order from the N- to C-terminus. We chose the TM-score threshold (TM-score of < 0.3) according to both the preceding study[44] and the results of our own empirical analysis, where we observed the distribution of TM-scores based on 1,000,000 randomly selected binding site pairs (Supporting Information Fig. S3). A large amount of similar site pairs of homologous proteins was distributed at a TM-score of ≥ 0.3. The minimum threshold 5 for the aligned residue lengths was set to avoid an accidental match when relatively larger sites are aligned.

The filtering by TM-align might result in biologically meaningful pairs being lost, especially those with sequence-order independent structural similarity. Nevertheless, we used it to narrow down the scope of manual inspection. We used TM-align for practical reasons, because this program is fast enough to obtain structural pairwise alignments for a large number of enumerated pairs. Other advantageous methods, such as sequence-order independent structural alignment methods are too slow for this task. For instance, SiteEngine takes around 10 s per pairwise comparison and needs 1941 days ([16,775,641 × 10]/[24 × 3600]) to obtain all superpositions of the 16,775,641 pairs on a single core CPU.

Ligand-binding sites shared by different superfamilies or folds

To further analyze ligand-binding sites, we classified the qualified pairs into two categories, homologous and analogous, based on their CATH codes. Homologous pairs were those with high similarity in sequences, whereas analogous pairs did not have clues on function sharing in terms of sequences. A binding site could have multiple CATH codes. A pair was regarded as homologous, if at least one pair of the CATH codes exactly matched at all four levels, and as analogous otherwise. As a result, about 17% (597,404) of the pairs were analogous (Fig. 10).

Figure 11 depicts the fraction of homologous pairs against the number of aligned residues obtained by TM-align. The fraction of homologous pairs is relatively low for the smallest binding sites. Most analogous pairs here involve metal ion-binding sites. For example, Figure 12(A) illustrates two proteins that do not have any global structural similarities: an EF-hand containing protein (PDB ID: 1XO5)[45] and a transferase (PDB ID:

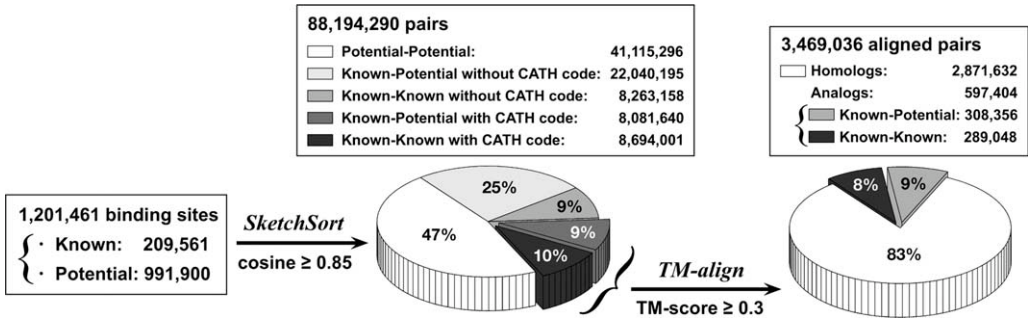

Figure 10
Groupings of obtained pairs. The left indicates the 1.2 million binding sites used in this study. The center indicates the pairs with cosine similarity ≥ 0.85 that were enumerated by SketchSort. The right indicates the pairs with TM-score ≥ 0.3 after structural alignment by TM-align.

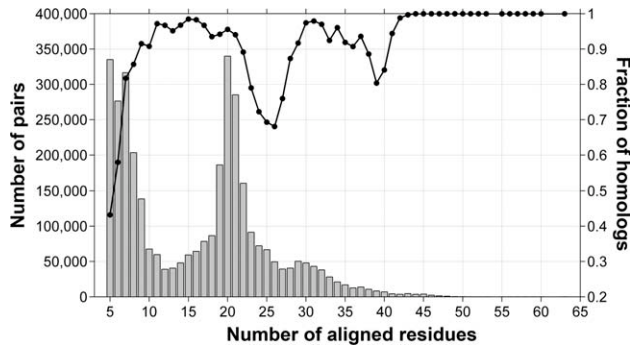

Figure 11

Distribution of structurally aligned residue lengths and fraction of homologous proteins. Bars represent the number of aligned residues by TM-align and the line indicates the fraction of site pairs that originated from two homologous proteins.

1NUD).[46] Both proteins have similar Ca^{2+} binding sites. Interestingly, the calcium-binding region in the former structure is a typical EF-hand motif whereas that in the later one is a loop between two β strands. It is known that binding/releasing of Ca^{2+} ion(s) plays important roles in the functional expression of various proteins, and thus their binding sites are structurally diverse.[57] Figure 12(B) shows an example of significantly similar zinc-binding sites shared by two different DNA primases: the bacteriophage T7 DNA primase (PDB ID: 1NUI)[47] and the *Bacillus stearothermophilus* DNA primase (PDB ID: 1D0Q).[48] Both structures have a zinc-binding domain (ZBD), which is known to be essential for RNA primer synthesis. Interestingly, the ZBD, including a zinc ribbon motif,[58] of T7 primase is separated from other domains by a long linker and is classified into mainly beta class in CATH (2.20.25.10), whereas the ZBD of *Bacillus* primase also has a zinc ribbon motif sandwiched between α-helices, and is classified into mixed alpha-beta class (3.90.580.10). Our method could detect such similar regions irrespective of the arrangement of secondary structures or fold topologies. As well as this salient example, we observed large numbers of other pairs of metal ion-binding sites shared across distinct proteins, such as similar iron (+2) cation (Fe^{2+}) binding sites between ribonucleotide-triphosphate reductase (PDB ID: 1H7A)[59] and rubredoxin (PDB ID: 1SMU)[60] and similar manganese (+2) cation (Mn^{2+}) binding sites between glutamine synthetase (PDB ID: 1LGR)[61] and mRNA triphosphatase (PDB ID: 1D8H).[62] We also observed a number of similar binding sites with metal substitutions, such as (Zn^{2+}, Cd^{2+}), (Zn^{2+}, Fe^{2+}), (Mg^{2+}, Ca^{2+}), and (Mn^{2+}, Fe^{2+}). These observations suggest that similar metal ion-binding sites are frequently conserved across distinct proteins possibly because of the significant roles they play in protein functions, and they suggest the substantial existence of promiscuous metal ion-binding sites.[63]

In the range between 19 and 21 in Figure 11, there is a small drop in the fraction of homologous pairs. This corresponds to analogous heme-binding site pairs. The fraction of analogous pairs is not large: Out of the 768,192 heme-binding site pairs detected in this study, only 1375 pairs were analogous. A case of such analogous pairs is shown in Figure 12(C): similar heme-binding sites shared between a succinate dehydrogenase (PDB ID: 1YQ4)[49] and a formate dehydrogenase (PDB ID: 1KQF).[50] Interestingly, the former heme-binding site is surrounded by four-helix bundles that originated from two different chains (two helices from the C-chain: residues 34–61 and 81–110 and two helices from the D-chain: residues 5–29 and 33–58), whereas the latter heme-binding site is also surrounded by four-helix bundles where all four helices (residues 12–36, 50–78, 55–134, and 145–175) reside in the same C-chain. This example demonstrates the unique ability of our method to detect similar sites in the interface region between multichains or multidomains.

There are two exceptional drops in the range between 20 and 40 in Figure 11. This is mostly due to analogous nucleotide-binding sites, such as the binding sites of FAD, ADP, NAD, and their cognates that are abundant in the current PDB. Figure 12(D) shows superpositions of the FAD-binding sites of dihydrolipoyl dehydrogenase (PDB ID: 2F5Z)[51] and the NAD-binding sites of malate dehydrogenase (PDB ID: 1GUY).[52] Although the nicotinamide- and flavin-binding regions of two structures have different conformations, the adenine dinucleotide-binding regions that are common to FAD and NAD can be well superimposed. Figure 12(E) shows similar ADP-binding sites of a bifunctional PAPS synthetase (PDB ID: 2OFX)[53] and a kinesin-like protein KIF11 (PDB ID: 1Q0B).[54] These two analogous pairs are well known and have also been found by other research groups.[30,31] In addition to these fully similar sites, we detected partially similar sites as shown in Figure 12(F): an ADP-binding site of a hypothetical ABC transporter ATP-binding protein (PDB ID: 1F3O)[55] and a phosphate (PO_4) binding site of an ATP-dependent phosphoenolpyruvate carboxykinase (PDB ID: 1J3B).[56] Although the overall topologies and the types of bounded ligands of the two structures are different, both share a part of the structural P-loop,[30] GK[S/T] signature.

Similarity networks

We generated a similarity network of known binding sites of nucleotide-containing ligands and SO_4/PO_4 to investigate their diversity with respect to superfamilies and folds (Supporting Information Fig. S4). We obtained several distinct components.

Of these, the largest connected component was comprised of the 2444 nodes with 448,342 edges shown in Supporting Information Figure S4A. Most nodes in the component are binding sites of FAD, NAD, NAP, and NDP, which are distributed in two different architectures,

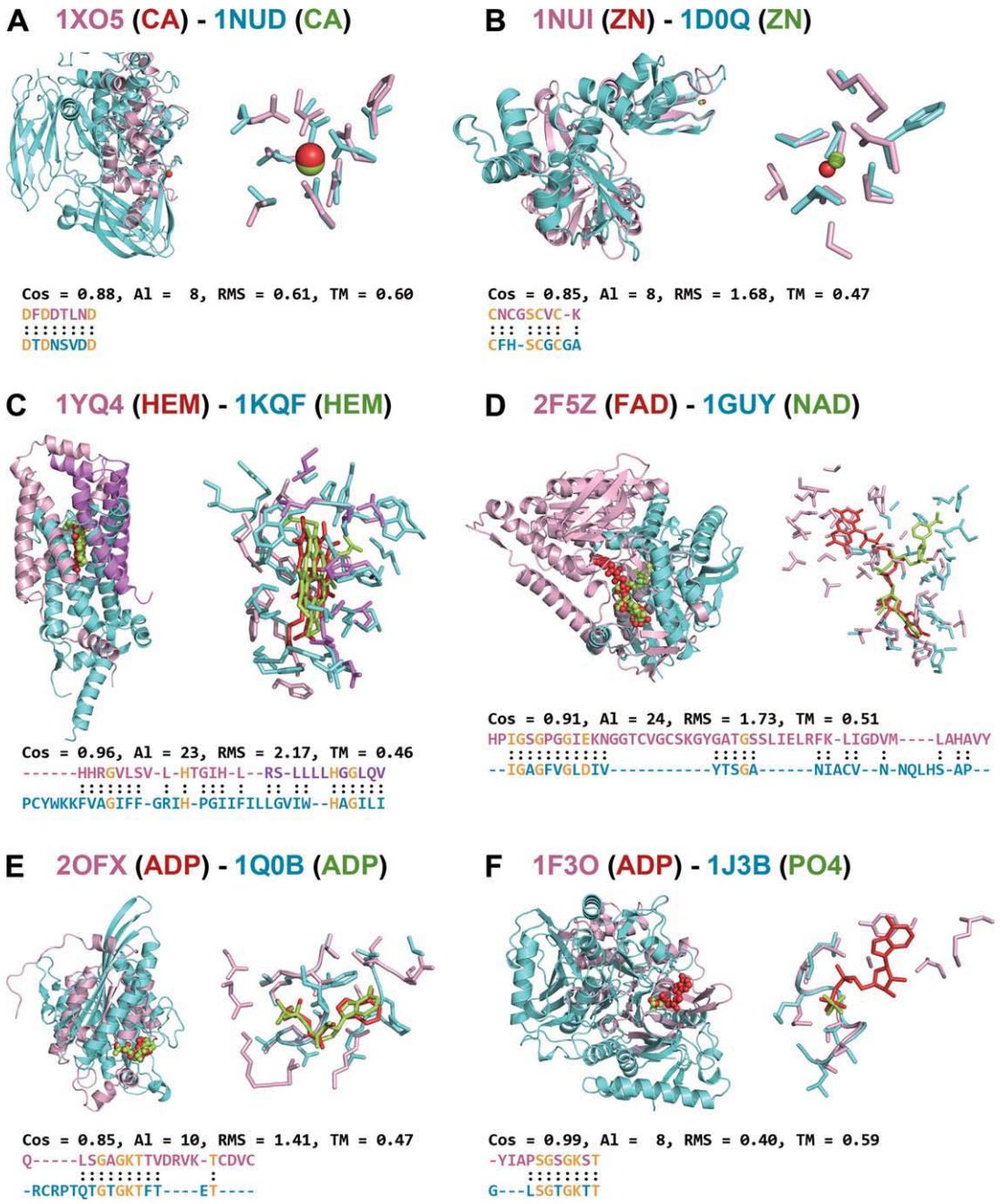

Figure 12

Similar binding sites shared between proteins with different topologies. The left side shows the two global structures (colored with pink and cyan) superimposed according to their binding sites shown in the right side. The ligand bound to each structure is shown in red and lime colored spheres or sticks. Cos, AL, RMS, and TM mean the cosine similarity, aligned length with TM-align, RMSD, and TM-score, respectively. ":" means a pair of aligned residues and "-" means a gap. **A**: Calcium-binding sites shared between a calcium binding protein (PDB ID: 1XO5[45]; CATH Code: 1.10.238.10) and a transglutaminase 3 (1NUD[46]; 3.90.260.10). **B**: Zinc-binding sites shared between bacteriophage T7 DNA primase (1NUI[47]; 2.20.25.10) and *Bacillus stearothermophilus* DNA primase (1D0Q[48]; 3.90.580.10). **C**: Similar heme-binding sites shared between a succinate dehydrogenase (1YQ4[49]; 1.20.1300.10) and a formate dehydrogenase (1KQF[50]; 1.20.950.20). The heme-binding site of succinate dehydrogenase is in the interface between two chain A and chain B colored with pink and violet. **D**: FAD- and NAD-binding sites shared between a dihydrolipoyl dehydrogenase (2F5Z[51]; 3.50.50.60) and a malate dehydrogenase (1GUY[52]; 3.40.50.720). **E**: ADP-binding sites shared between a bifunctional PAPS synthetase (2OFX[53]; 3.40.50.300) and a kinesin-like protein KIF11 (1Q0B[54]; 3.40.850.10). **F**: ADP- and PO_4-binding sites shared between a hypothetical ABC transporter ATP-binding protein (1F3O[55]; 3.40.50.300) and an ATP-dependent phosphoenolpyruvate carboxykinase (1J3B[56]; 3.90.228.20).

four different topologies and six different homologous superfamilies of the CATH classification. Of the topologies, the Rossman fold (3.40.50) and the FAD/NAD(P) binding domain (3.50.50) are connected most strongly. The two topologies contain a large number of FAD- or NAD-binding sites, which form a large share of the analogous pairs in this study. The other pairs well connected to the two topologies are aldehyde dehydrogenase (3.40.605) and deoxyhypusine synthase (3.40.910). As seen in Figure 12(D), 53% of the sites in the connected component contain the well-studied GxGxxG motif.[64]

The second largest component was comprised of 592 nodes with 16,904 edges (Supporting Information Fig. S4B). This component mainly consists of ATP/ADP- and GTP/GDP-binding sites, which are distributed in two different architectures, four different topologies and four different homologous superfamilies. The three different topologies, adenylosuccinate synthetase (3.40.440), kinesin (3.40.850), and phosphoenolpyruvate carboxykinase (3.90.228), are connected to the central superfamily of the P-loop containing nucleotide triphosphate hydrolases (3.40.50.300). Like those in Figure 12(E), 97% of the sites in the component contain a GK[S/T] signature. It can be speculated that the nucleotide-containing ligands are ubiquitous in nature and play critical roles in a number of biological processes and thus various unrelated proteins share the similar nucleotide-binding sites, which might divergently have evolved from a common ancestor or emerged through convergent evolution.[65] As discussed in another section (relationships between ligands), SO_4 and PO_4 often bind to similar sites of nucleotide-containing ligands.

Of the other components (Supporting Information Fig. S4C), we found the eighth largest component (86 nodes) particularly interesting. In this component mainly consisting of SO_4- and PO_4-binding sites, one node (SO_4-binding site) from kinesin (3.40.850) and four nodes (one SO_4-binding site and three PO_4-binding sites) from phosphoenolpyruvate carboxykinase (3.90.228) are connected to the nodes from the P-loop containing nucleotide triphosphate hydrolases (3.40.50.300). Most (97%) of these SO_4/PO_4-binding sites also contain a GK[S/T] signature, as seen in Figure 12(F). Although none of the P-loop containing nucleotide triphosphate hydrolases in the second largest component is SO_4/PO_4-binding, this component contains as many as 32 SO_4/PO_4-binding variants.

Pairs of known and putative binding sites

Out of 597,404 analogous pairs, 308,356 pairs contain an empty pocket found by *ghecom* (Fig. 10). These pairs are particularly interesting, because they can be used to speculate on the type of binding ligand and its binding mode for uncharacterized regions of proteins. Nevertheless, there are far too many pairs for exhaustive manual inspection. Therefore, to narrow down to the relatively large and potentially biologically relevant site pairs, we selected 17,806 pairs whose aligned residues were ≥ 15. Major ligands were nucleotide-containing ligands such as FAD, NAD, ADP, and their analogs, but were not limited to these (Table SII). In this study, we focused on pairs involving nucleotide-containing ligands. It is highly plausible that our results contain undiscovered pairs that are biologically important. We intend to make our pairs available to the public on the Web to allow other scientists to scrutinize them.

The first example we found interesting was a putative binding site of the ThiS-ThiF protein complex (PDB ID: 1ZUD)[66] that is involved in the synthesizing the thiazole moiety of thiamin. Although we found many known FAD binding sites that are similar to the pocket of chain 1 (or 3) in 1ZUD, we were also able to detect similarities in the pocket to the ADP binding site of domain 1 in uncharacterized protein DR_0571 (PDB ID: 3C4N) with a cosine similarity of 0.91. This is a typical example of the relationship between two different proteins, that is, CATH code 3.40.50.720 for the chain 1 of 1ZUD and CATH code 3.50.50.60 for the domain 1 of 3C4N, which share the similar nucleotides-binding sites shown in Supporting Information Figure S4A. In fact, 20 residues, including almost all residues involving ADP-binding in 3C4N are aligned to those in 1ZUD with the C_α RMSD of 1.51 Å by TM-align [Fig. 13(A)]. According to this alignment, three residues of I42, G43, and G45 on the nucleotide-binding loop of 3C4N are conserved (I34, G35, and G37) in 1ZUD. Although the glycine rich motif is not completely conserved, the similarity between the two sites was detected with our method. We also found three other residues of V67, E68, and E69 of 3C4N, located near the adenosine substituted similar residues (A58, D59, and D60) in 1ZUD. The correspondence between two polar residues, T79 in the short helix 2 of 3C4N and N67 in the 3_{10}-helix 5 of 1ZUD might give clues to the implication of a catalytic mechanism.[66] Thus, the results from comparing this pair supports the ATP-binding mode obtained from modeling studies by Lehmann *et al.*[66] and may also shed light on the uncharacterized mechanisms of the molecular function of 3C4N.

The second example is a putative binding site of ribosomal protein S6 kinase (PDB ID: 2WNT), which is a serine/threonine kinase, although its ATP-bound form has not been obtained. One of similar sites to the putative binding site was an ATP analog (ANP) binding site of other serine/threonine kinase (chain F from PDB ID: 2WTK).[67] Twenty-three residue pairs between the two sites were aligned with the C_α RMSD = 1.47 Å [Fig. 13(B)]. Despite their distant relationships, the two proteins were classified into an analogous pair in this study, because domain assignments for the two proteins in CATH are slightly different, that is, CATH code has not been assigned to the discontinuous N-terminal (01) domain of 2WNT yet, while CATH code 1.10.510.10 has been assigned to the discontinuous C-terminal (02) domain. Also, CATH code 3.30.200.20 has been assigned

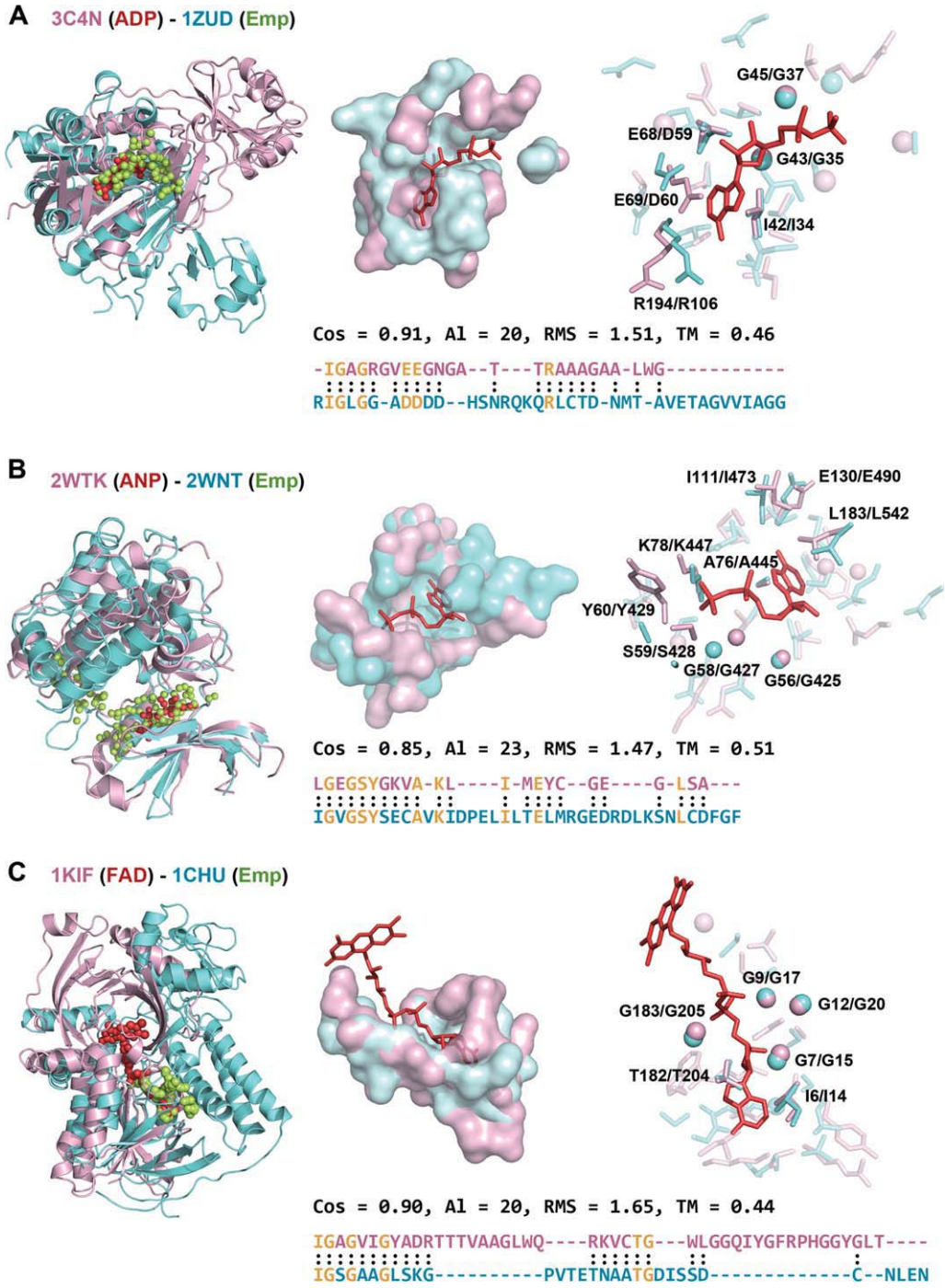

Figure 13

Example pairs of known and probable ligand-binding sites. Superimposed global structures based on their binding sites are shown at the left. Superimposed binding sites in the surface model and in the ball-and-stick model are shown at the center and the right, respectively. Alignment information is shown under the superimposed binding sites. The red spheres and sticks indicate a ligand bound to the protein colored in pink and the lime spheres indicate the potential ligand-binding region of the protein colored in cyan. The residues in yellow in the alignment indicate the well conserved residues in the pair. **A**: A pair of a pocket of ThiS-ThiF complex (PDB ID: 1ZUD[66]; CATH code: 3.40.50.720) and an ADP-binding site of DR_0571 protein (3C4N; 3.50.50.60). **B**: A pair of a pocket of ribosomal protein S6 kinase (2WNT; 1.10.510.10) and an ANP-binding site of Serine/Threonine-Protein kinase (2WTK[67]; 3.30.200.20). **C**: A pair of a pocket of L-aspartate oxidase (1CHU[70]; 3.50.50.60) and a FAD-binding site from D-amino acid oxidase (1KIF[71]; 3.40.50.720).

to the N-terminal (01) domain of chain F in 2WTK, whereas CATH code has not been assigned to the continuous C-terminal (02) domain. Interestingly, nine amino acid residues, including part of the GxGxxG kinase motif, are identical between the two sites, although the last Gly of the motif is substituted by Ser in 2WNT. Apart from previous analyses,[68,69] the results may suggest the importance of additional residues for the ATP-binding mode in the putative binding site of ribosomal protein S6 kinase without the structural determination of its ATP-bound form, and thus demonstrates the usefulness of our method for inferring the binding ligand to protein structures whose binding mechanism is unclear.

The third example is a putative binding site of apo L-aspartate oxidase (PDB ID: 1CHU)[70] that catalyzes the conversion of L-Asp to iminoaspartate in bacteria. There were dozens of known FAD-binding sites that matched to the potential pocket with cosine $\geq$ 0.9 and TM-score $\geq$ 0.4. Of these, the example of a 1KIF[71] and 1CHU pair is shown in Figure 13(C). Similarly, this is another example of the relationship between CATH code 3.40.50.720 for the domain 1 of 1KIF and CATH code 3.50.50.60 for the domain 1 of 1CHU. Twenty residues between the two sites were well aligned with the C_α RMSD of 1.65 Å, including those around the ribose and adenosine of FAD in 1KIF. This indicates that the pocket of the 1CHU is bound with FAD in the region shown in Figure 13(C), which is consistent with the FAD-bound form solved by Bossi *et al.*[72] This example indicates that our method is also useful for inferring probable binding sites using its apo structure.

Relationships between ligands

It is known that different ligands, for example, manganese and magnesium,[73] can be accommodated in the same or similar binding sites. To systematically investigate such relationships, we construct a ligand distance (LD) matrix using the pairs of known sites we obtain. The known sites are categorized according to their ligands. Denote by $|e_{s,t}|$ the number of similar site pairs across two categories corresponding to ligand s and t. In addition, $|e_s|$ denotes the number of similar pairs inside the category corresponding to ligand s. The distance between two ligands $ld_{s,t}$ is defined as

$$ld_{s,t} = -\log\left(\frac{|e_{s,t}|}{\sqrt{|e_s|}\sqrt{|e_t|}}\right).$$

If $ld_{s,t} \geq 12$, it is set to 12. A short distance implies that similar sites are widely shared by different ligands s and *t*.

Figure 14 shows the distance matrix for 36 frequently observed and biologically relevant ligands in PDB as well as the results from hierarchical clustering analysis. Solution additives such as glycerol were excluded from this analysis. As expected, similar types of ligands such as GTP, GDP, and GNP tend to cluster together. The biggest cluster consists of metal ions, indicating that they can often be substituted for one another. Ca^{2+} demonstrated a high proximity to sugars, such as MAN and GAL and not only to metal ions. This is because sugar-binding proteins, such as lectins, exhibit their sugar-binding affinity depending on the calcium-binding, and both a sugar and Ca^{2+} are accommodated in the same binding site (cf. PDB ID: 1RDK[74] and 1KZC[75]). Nucleotide-containing ligands such as ADP, GDP, S-adenosylmethionine (SAM), NAD, and FAD form another large cluster, because all these ligands have a common nucleotide portion and their binding regions often exhibit similarities. Interestingly, sulfate ions (SO_4) and phosphate ions (PO_4) seem to have proximity to various types of ligands including metal ions, sugars, and nucleotide-containing ligands. Taken together, the distance matrix seems to reflect the exchangeability of ligands quite well. Even though this analysis was concentrated on common ligands, the extended distance matrix including drug leads would be helpful in multitarget drug design.

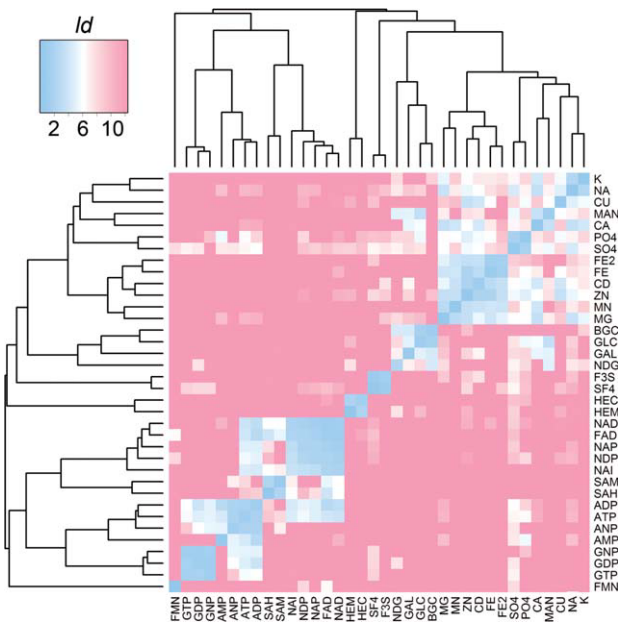

Figure 14

Distance matrix constructed based on 36 frequently observed ligands in PDB. The smaller the value of *ld* (i.e., the closer the color to blue), the more similar the two ligands. For hierarchical clustering, the distance between clusters was defined as the averaged distance between cluster members.

CONCLUSIONS

We proposed a tremendously fast algorithm that enables the enumeration of similar pairs in a huge number of protein-ligand binding sites. The execution time for enumerating all similar pairs is extremely fast: about 30 h on single-core processor for 1.2 million known and

potential binding sites. The superb efficiency allowed us to try different amino acid encodings to cover different aspects of binding sites. We obtained about 3.5 million structurally similar binding-site pairs with the assigned CATH code. It turned out that most of these (over 83%) were conserved within the homologous proteins according to the CATH classification. Our analysis of the remaining analogous pairs shed light on the hidden relationships between distinct protein families or folds. We believe that our method is a powerful tool for comprehensively discovering similar binding sites before they are verified by accurate but slow structural alignment. In combination with other methods, ours would be useful for finding new structural motifs relevant to protein functions. We plan to build a Web-based interface, through which all pairs detected with our method will be available for researchers. The source code for SketchSort is available on our Web site (http://sites.google.com/site/yasuotabei/).

ACKNOWLEDGMENTS

We are grateful to Prof. Takeshi Kawabata for supplying the *ghecom* program. Useful discussions with Takeaki Uno, Junichi Higo, Paul Horton, Akira Kinjo, and Ryoko Morioka are gratefully acknowledged.

REFERENCES

1. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
2. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
3. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
4. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
5. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
6. Brady L, Brzozowski AM, Derewenda ZS, Dodson E, Dodson G, Tolley S, Turkenburg JP, Christiansen L, Huge-Jensen B, Norskov L, Thim L, Menge. A serine protease triad forms the catalytic centre of a triacylglycerol lipase. Nature 1990;343:767–770.
7. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 1996;5:1001–1013.
8. Via A, Ferre F, Brannetti B, Valencia A, Helmer-Citterich M. Three-dimensional view of the surface motif associated with the P-loop structure: cis and trans cases of convergent evolution. J Mol Biol 2000;303:455–465.
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
10. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330, 307–328.
11. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15:359–363, 389.
12. Brady GP, Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14:383–401.
13. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 2005;4:752–761.
14. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 2006;34(Web Server issue):W116–W118.
15. Kawabata T. Detection of multi-scale pockets on protein surfaces using mathematical morphology. Proteins 2010;78:1195–1211.
16. Yu J, Zhou Y, Tanaka I, Yao M. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics 2010;26:46–52.
17. Kasahara K, Kinoshita K, Takagi T. Ligand-binding site prediction of proteins based on known fragment-fragment interactions. Bioinformatics 2010;26:1493–1499.
18. Wang B, Chen P, Huang DS, Li JJ, Lok TM, Lyu MR. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 2006;580:380–384.
19. Chen XW, Jeong JC. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 2009;25:585–591.
20. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 2006;6:19–29.
21. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 2009;5:e1000585.
22. Kinoshita K, Furui J, Nakamura H. Identification of protein functions from a molecular surface database, eF-site. J Struct Funct Genomics 2002;2:9–22.
23. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
24. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
25. Gold ND, Jackson RM. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol 2006;355:1112–1124.
26. Park K, Kim D. A method to detect important residues using protein binding site comparison. Genome Inform 2006;17:216–225.
27. Kinjo AR, Nakamura H. Similarity search for local protein structures at atomic resolution by exploiting a database management system. Biophysics 2007;3:75–84.
28. Minai R, Matsuo Y, Onuki H, Hirota H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. Proteins 2008;72:367–381.
29. Najmanovich R, Kurbatova N, Thornton J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. Bioinformatics 2008;24:i105–i111.
30. Brakoulias A, Jackson RM. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins 2004;56:250–260.
31. Kinjo AR, Nakamura H. Comprehensive structural classification of ligand-binding motifs in proteins. Structure 2009;17:234–246.
32. Schalon C, Surgand JS, Kellenberger E, Rognan D. A simple and fuzzy method to align and compare druggable ligand-binding sites. Proteins 2008;71:1755–1778.
33. Yin S, Proctor EA, Lugovskoy AA, Dokholyan NV. Fast screening of protein surfaces using geometric invariant fingerprints. Proc Natl Acad Sci USA 2009;106:16622–16626.
34. Weill N, Rognan D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. J Chem Inf Model 2010;50:123–135.

35. Chikhi R, Sael L, Kihara D. Real-time ligand binding pocket database search using local surface descriptors. Proteins 2010;78:2007–2028.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
37. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–595.
38. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
39. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
40. Tabei Y, Uno T, Sugiyama M, Tsuda K. Single versus multiple sorting for all pairs similarity search. In the 2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, 2010;13:145–160.
41. Indyk P, Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the 30th Annual ACM Symposium on Theory of Computing, Dallas, TX, 1998. pp. 604–613.
42. Goemans M, Williamson D. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J Assoc Comput Machine 1995;42:1115–1145.
43. Uno T. Multi-sorting algorithm for finding pairs of similar short substrings from large-scale string data. Knowl Inf Syst 2010;25:229–251.
44. Sommer I, Muller O, Domingues FS, Sander O, Weickert J, Lengauer T. Moment invariants as shape recognition technique for comparing protein binding sites. Bioinformatics 2007;23:3139–3146.
45. Gentry HR, Singer AU, Betts L, Yang C, Ferrara JD, Sondek J, Parise LV. Structural and biochemical characterization of CIB1 delineates a new family of EF-hand-containing proteins. J Biol Chem 2005;280: 8407–8415.
46. Ahvazi B, Boeshans KM, Idler W, Baxa U, Steinert PM. Roles of calcium ions in the activation and activity of the transglutaminase 3 enzyme. J Biol Chem 2003;278:23834–23841.
47. Kato M, Ito T, Wagner G, Richardson CC, Ellenberger T. Modular architecture of the bacteriophage T7 primase couples RNA primer synthesis to DNA synthesis. Mol Cell 2003;11:1349–1360.
48. Pan H, Wigley DB. Structure of the zinc-binding domain of *Bacillus stearothermophilus* DNA primase. Structure 2000;8:231–239.
49. Huang L, Sun G, Cobessi D, Wang A, Shen J, Tung E, Anderson V, Berry E. 3-Nitropropionic acid is a suicide inhibitor of mitochondrial respiration that, upon oxidation by complex II, forms a covalent adduct with a catalytic base arginine in the active site of the enzyme. J Biol Chem 2005;281:5965–5972.
50. Jormakka M, Tornroth S, Byrne B, Iwata S. Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. Science 2002;295:1863–1868.
51. Brautigam C, Wynn R, Chuang J, Machius M, Tomchick D, Chuang D. Structural insight into interactions between dihydrolipoamide dehydrogenase (E3) and E3 binding protein of human pyruvate dehydrogenase complex. Structure 2006;14:611–621.
52. Dalhus B, Saarinen M, Sauer UH, Eklund P, Johansson K, Karlsson A, Ramaswamy S, Bjork A, Synstad B, Naterstad K, Sirevag R, Eklund H. Structural basis for thermophilic protein stability: structures of thermophilic and mesophilic malate dehydrogenases. J Mol Biol 2002;318:707–721.
53. Sekulic N, Dietrich K, Paarmann I, Ort S, Konrad M, Lavie A. Elucidation of the active conformation of the APS-kinase domain of human PAPS synthetase 1. J Mol Biol 2007;367:488–500.
54. Yan Y, Sardana V, Xu B, Homnick C, Halczenko W, Buser CA, Schaber M, Hartman GD, Huber HE, Kuo LC. Inhibition of a mitotic motor protein: where, how, and conformational consequences. J Mol Biol 2004;335:547–554.
55. Yuan YR, Blecker S, Martsinkevich O, Millen L, Thomas PJ, Hunt JF. The crystal structure of the MJ0796 ATP-binding cassette. Implications for the structural consequences of ATP hydrolysis in the active site of an ABC transporter. J Biol Chem 2001;276:32313–32321.
56. Sugahara M, Ohshima N, Ukita Y, Kunishima N. Structure of ATP-dependent phosphoenolpyruvate carboxykinase from *Thermus thermophilus* HB8 showing the structural basis of induced fit and thermostability. Acta Crystallogr D Biol Crystallogr 2005;61(Part 11):1500–1507.
57. Grabarek Z. Structural basis for diversity of the EF-hand calcium-binding proteins. J Mol Biol 2006;359:509–525.
58. Qian X, Jeon C, Yoon H, Agarwal K, Weiss MA. Structure of a new nucleic-acid-binding motif in eukaryotic transcriptional elongation factor TFIIS. Nature 1993;365:277–279.
59. Larsson KM, Andersson J, Sjoberg BM, Nordlund P, Logan DT. Structural basis for allosteric substrate specificity regulation in anaerobic ribonucleotide reductases. Structure 2001;9:739–750.
60. Park IY, Youn B, Harley JL, Eidsness MK, Smith E, Ichiye T, Kang C. The unique hydrogen bonded water in the reduced form of *Clostridium pasteurianum* rubredoxin and its possible role in electron transfer. J Biol Inorg Chem 2004;9:423–428.
61. Liaw SH, Jun G, Eisenberg D. Interactions of nucleotides with fully unadenylylated glutamine synthetase from *Salmonella typhimurium*. Biochemistry 1994;33:11184–11188.
62. Lima CD, Wang LK, Shuman S. Structure and mechanism of yeast RNA triphosphatase: an essential component of the mRNA capping apparatus. Cell 1999;99:533–543.
63. Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M. Prediction of transition metal-binding sites from apo protein structures. Proteins 2008;70:208–217.
64. Dym O, Eisenberg D. Sequence-structure analysis of FAD-containing proteins. Protein Sci 2001;10:1712–1728.
65. Kinoshita K, Sadanami K, Kidera A, Go N. Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. Protein Eng 1999;12:11–14.
66. Lehmann C, Begley TP, Ealick SE. Structure of the *Escherichia coli* ThiS-ThiF complex, a key component of the sulfur transfer system in thiamin biosynthesis. Biochemistry 2006;45:11–19.
67. Zeqiraj E, Filippi BM, Deak M, Alessi DR, van Aalten DM. Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. Science 2009;326:1707–1711.
68. Leader DP. Identification of protein kinases by computer. Nature 1988;333:308.
69. Malakhova M, Tereshko V, Lee SY, Yao K, Cho YY, Bode A, Dong Z. Structural basis for activation of the autoinhibitory C-terminal kinase domain of p90 RSK2. Nat Struct Mol Biol 2008;15:112–113.
70. Mattevi A, Tedeschi G, Bacchella L, Coda A, Negri A, Ronchi S. Structure of L-aspartate oxidase: implications for the succinate dehydrogenase/fumarate reductase oxidoreductase family. Structure 1999;7:745–756.
71. Mattevi A, Vanoni MA, Todone F, Rizzi M, Teplyakov A, Coda A, Bolognesi M, Curti B. Crystal structure of D-amino acid oxidase: a case of active site mirror-image convergent evolution with flavocytochrome b2. Proc Natl Acad Sci USA 1996;93:7496–7501.
72. Bossi RT, Negri A, Tedeschi G, Mattevi A. Structure of FAD-bound L-aspartate oxidase: insight into substrate specificity and catalysis. Biochemistry 2002;41:3018–3024.
73. Bock C, Kaufman-Katz A, Markham G, Glusker J. Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions. J Am Chem Soc 1999;121:7360–7372.
74. Ng KK, Drickamer K, Weis WI. Structural analysis of monosaccharide recognition by rat liver mannose-binding protein. J Biol Chem 1996;271:663–674.
75. Ng KK, Kolatkar AR, Park-Snyder S, Feinberg H, Clark DA, Drickamer K, Weis WI. Orientation of bound ligands in mannose-binding proteins. Implications for multivalent ligand recognition. J Biol Chem 2002;277:16088–16095.