BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 10 2012, pages 1345–1352
doi:10.1093/bioinformatics/bts138

Structural bioinformatics

Advance Access publication April 6, 2012

Fast protein binding site comparisons using visual words representation

Bin Pang[1], Nan Zhao[1], Dmitry Korkin[1,2] and Chi-Ren Shyu[1,2,*]

[1]Informatics Institute and [2]Department of Computer Science, University of Missouri, Columbia, MO 65211, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Finding geometrically similar protein binding sites is crucial for understanding protein functions and can provide valuable information for protein–protein docking and drug discovery. As the number of known protein–protein interaction structures has dramatically increased, a high-throughput and accurate protein binding site comparison method is essential. Traditional alignment-based methods can provide accurate correspondence between the binding sites but are computationally expensive.

Results: In this article, we present a novel method for the comparisons of protein binding sites using a 'visual words' representation (PBSword). We first extract geometric features of binding site surfaces and build a vocabulary of visual words by clustering a large set of feature descriptors. We then describe a binding site surface with a high-dimensional vector that encodes the frequency of visual words, enhanced by the spatial relationships among them. Finally, we measure the similarity of binding sites by utilizing metric space operations, which provide speedy comparisons between protein binding sites. Our experimental results show that PBSword achieves a comparable classification accuracy to an alignment-based method and improves accuracy of a feature-based method by 36% on a non-redundant dataset. PBSword also exhibits a significant efficiency improvement over an alignment-based method.

Availability: PBSword is available at http://proteindbs.rnet.missouri.edu/pbsword/pbsword.html

Contact: shyuc@missouri.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 3, 2011; revised on February 8, 2012; accepted on March 18, 2012

1 INTRODUCTION

Protein–protein interaction, as an essential aspect of biological processes, plays a significant role in determining protein functions and forming protein complexes. The protein complexes consist of multiple subunits (domains or chains) that interact via protein–protein interfaces (PPIs) Bahadur and Zacharias, 2008; Kim *et al.*, 2006). The PPIs are composed of geometrically and physicochemically complementary binding sites from the respective subunits. In a protein complex, each binding site corresponds to a region of residues that are in close spatial proximity and can interact with residues from another subunit.

With the development of high-throughput experimental techniques (e.g. yeast-two-hybrid screens), the size of data repositories of protein–protein interactions has dramatically increased, which makes a comprehensive understanding of the protein interaction network feasible (Aloy and Russell, 2006). Several protein–protein interactionrelated databases have been constructed, such as SCOPPI (Winter *et al.*, 2006), PIBASE (Davis and Sali, 2005), IntAct (Kerrien *et al.*, 2012), iPfam (Finn *et al.*, 2005) and DOMMINO (Kuang *et al.*, 2012). Particularly, the SCOPPI database provides an evolutionary and structural classification of PPIs based on SCOP family (Murzin *et al.*, 1995), sequence similarity and geometric features of binding sites. To analyze the protein–protein interaction mechanisms, an efficient method for comparing protein binding sites is an indispensible tool, with a potential impact on protein function prediction, protein–protein docking, drug discovery and evolutionary studies (Bradford *et al.*, 2006; Henschel *et al.*, 2006; Tuncbag *et al.*, 2008; Wu *et al.*, 2004; Zhao *et al.*, 2011).

Methods for comparing the protein binding sites can be generally classified into two categories: (i) alignment-based and (ii) feature-based. Early research works on alignment-based methods focused on comparing binding sites by aligning the corresponding C_α atoms from different protein complexes using global structural alignment tools (Keskin and Nussinov, 2007; Keskin *et al.*, 2004; Tsai *et al.*, 1996). This group of methods works well when sequence and structure are well conserved. However, as shown in an earlier work (Nagano *et al.*, 2002), the limitation of global protein structure alignment is that it cannot usually produce accurate interface alignment since the true interface may not be the first priority for a global structure alignment method. To overcome this issue, iAlign was developed to align the residues from PPIs (Gao and Skolnick, 2010). Another method, I2ISiteEngine, supports alignment of two binding sites based on the similarity of physicochemical properties and the shapes of surfaces (Shulman-Peleg *et al.*, 2004). The alignment-based methods can provide accurate correspondence between two binding sites but are usually computationally expensive, which makes them infeasible when working with very large datasets.

Feature-based methods, on the other hand, make comparison based on feature descriptor, which may be structural and/or geometric properties of the binding site. Typically, the structural descriptors are distributions of distances between different types of functional atoms of the binding sites (Sander *et al.*, 2008), whereas the geometric descriptors include spin-image (Merelli

*To whom correspondence should be addressed.

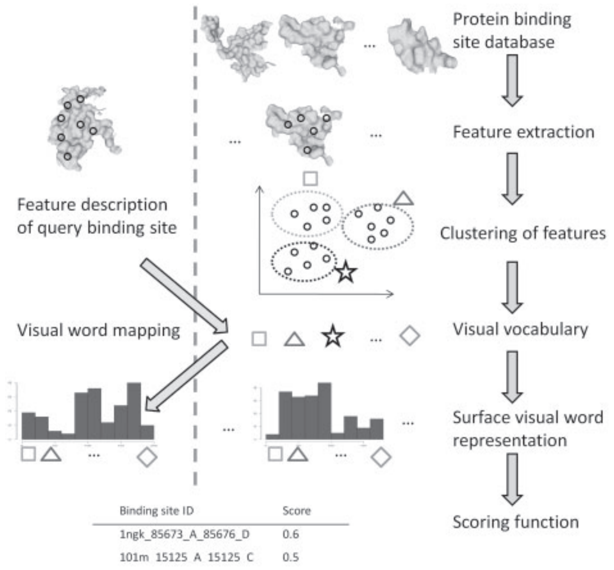

Fig. 1. Framework of PBSword. The inputs include a query binding site and a database of binding sites. The outputs are similarity scores between the query and the database binding sites. For each binding site, geometric features are extracted and assigned to the nearest visual word from a vocabulary which is generated by a huge set of features collected from the entire database. Thus, each binding site can be represented by a histogram of occurrence of visual words. Two binding sites can be compared by the similarity of two visual words vectors. A color version of this figure is available as Supplementary Material.

et al., 2011), 3D-Zernike (Sael *et al.*, 2008), shape distribution (Das *et al.*, 2009), moment invariants (Sommer *et al.*, 2007) and many others (Liu *et al.*, 2009; Yin *et al.*, 2009). After generating surface features, two binding sites can be compared using various metric functions in the feature space to perform high-throughput surface comparisons without explicit alignments. Existing feature-based methods mainly focus on protein–ligand binding sites, which have not been extensively applied or evaluated on protein–protein binding site classification and related applications. This is important as protein–protein binding sites are known to have some unique characteristics, such as relatively large areas and flat binding site surfaces (Bahadur and Zacharias, 2008).

To deal with very large datasets, one can employ the classic method from the information retrieval area of comparing the similarity between word frequency profiles of two documents, which has been successfully utilized in web search engines. Recently, this idea has been extended to the matching and retrieval of 3D objects (Bronstein *et al.*, 2011), in which a word is represented by a cluster of similar features. In the bioinformatics area, this technique has been applied in the alignment-free genome sequence comparisons (Sims *et al.*, 2009) and in the retrieval of similar protein structures (Budowski-Tal *et al.*, 2010).

In this article, we further extend the text comparison method and propose a novel approach, PBSword, for characterizing protein binding sites with a collection of 'visual words' such that similar binding sites from the database can be efficiently and accurately identified. We conduct experiments to evaluate the performance of PBSword-based classification and retrieval of protein binding sites. The experimental results show that PBSword can achieve comparable performance to an alignment-based method, iAlign, and significantly outperforms a feature-based method using moment invariant descriptors (Sommer *et al.*, 2007). More importantly, comparing PBSword vectors is orders of magnitude faster than the alignment-based methods. Thus, PBSword can be used to quickly identify geometrically similar candidates of a query protein binding site from a large dataset.

2 METHODS

The framework of PBSword is shown in Figure 1. The inputs include a query binding site and a database of protein binding sites. The outputs are similarity scores between the query and database binding sites. The workflow of PBSword consists of the following four steps. First, we select feature points of each database binding site surface and extract corresponding geometric features. Second, a visual vocabulary is built by clustering a huge number of feature point descriptors collected from the entire database of binding sites. Third, according to its descriptor, each feature point is associated with the nearest visual word from the visual vocabulary. This allows each binding site to be represented by the corresponding distribution of visual words. The above processes for the database binding sites are performed off-line. For the query binding site, we follow similar steps to generate its visual word representation. Finally, pairs of binding sites are compared by a scoring function which calculates the similarity between the two visual word vectors.

Comparing protein binding sites is a more challenging problem than comparing protein sequences or structures since the residues of a binding site are not always sequential in nature (Tsai *et al.*, 1996). Hence, a binding site cannot be represented by a string of residues from the N- to C-terminus of a protein chain. To solve this problem, we first use the surface features to represent the whole binding site. Then, by projecting the feature vectors into a visual word from the vocabulary, we can use the occurrence of visual words to describe the surface, which alleviates the need for sequential relationships between two feature points. As a feature-based method, PBSword shares some attractive properties as other similar methods in this category, such as being sequence- and structure- independent and alignment-free. The uniqueness of PBSword is in the way it represents the binding sites as a collection of visual words, which can be used not only in the development of a compact representation for a family of protein binding sites but also in the construction of an inverted index for fast retrieval of geometrically similar binding sites from a large dataset.

2.1 Surface generation

For a protein complex, we use the MSMS program (Sanner *et al.*, 1996) to generate a triangulated mesh for each of its interacting subunits and set the density and probe radius to 2.0 points/Å^2 and 1.4 Å, respectively. Since we are only interested in the binding regions, for each protein mesh, we retain only those surface points that are within a distance of 4Å from the surface of its binding partner (Shulman-Peleg *et al.*, 2004). A triangle is selected when its three vertices are all retained in the interaction region. In this article, we define a face, as used in the computer graphics community, to represent the surface of a triangle.

2.2 Feature descriptor

In our approach, feature points are defined as the centers of faces (or triangles), and a number of sample faces are selected using the procedure proposed in Osada *et al.* (2001) to improve computational efficiency. In this method, we first calculate the area of each face on the surface and store it as an array of cumulative areas. Then, we generate a random number ranging from 0 to the total area, and select the face corresponding to that value in the array of cumulative areas.

We further develop an approach based on the representation of shape contexts, originally introduced in Belongie *et al.* (2002) for 2D object matching, as a feature descriptor for each feature point. Before calculating

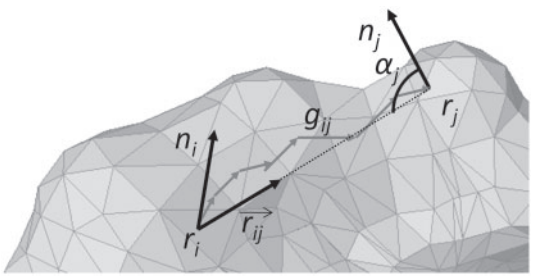

Fig. 2. Feature descriptors for a protein binding site. Two sample points (i and j) are shown on the binding site surface. n_i (or n_j) and r_i (or r_j) represent the normal vector and 3D coordinate of the i-th (or j-th) face center, respectively. The geodesic distance g_{ij} is calculated between r_i and r_j and the angle α_j is calculated between n_j and a unit vector pointing from r_i to r_j. A color version of this figure is available as Supplementary Material.

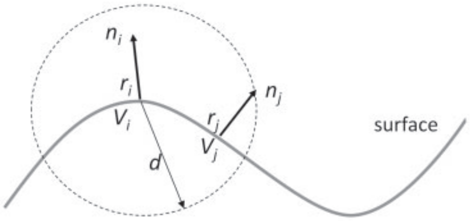

Fig. 3. Calculation of a pair-word frequency profile (PFP). Two sample points (i and j) are shown on the binding site surface. n_i (or n_j) and r_i (or r_j) represent the normal vector and 3D coordinate of the i-th (or j-th) face center, respectively. Suppose the visual word index for the sample point i (or j) is V_i (or V_j). To calculate the PFP for the sample point i, a ball with radius d is first placed at r_i. Then, the occurrence of word pairs with index (V_i, V_j) (j=1,…,k) within the ball is counted for different directions θ (θ = 0°, 45°, 90°, 135°, and 180°) separately. Here, the direction θ is defined as the angle between two corresponding normal vectors.

features, we normalize the scale of binding site mesh by the maximum geodesic distance from all pair-wise distances within each binding site. Assuming binding site A has N faces and M ($M << N$) sample faces, for face F_i, we use n_i and r_i (Fig. 2) to represent the normal vector and 3D coordinate of the i-th face center, respectively. For a given sample face F_i (i= 1,…,M) and another face F_j ($j \neq i, j$= 1,…,N) on the binding surface, we first calculate the geodesic distance g_{ij} between r_i and r_j using Dijkstra's shortest path algorithm (Dijkstra, 1959) and angle α_j between n_j and a unit vector pointing from r_i to r_j. Then, a feature descriptor W_i (or shape context) for the sample face F_i is calculated to quantitatively measure the distributions between F_i and the remaining $N-1$ face centers in logarithmic geodesic distances bins and angles bins, which are set to 10 and 8, respectively. By construction, the feature vector at a given face center is invariant under translation and scaling. In total, we extract M features $\{W_i^A, i \in \{1,\ldots,M\}\}$, and say that binding site A is characterized by its feature descriptor W^A. In the experiments, M is emperically set to 200.

2.3 Visual vocabulary construction

The feature descriptors from different binding sites are clustered to construct a vocabulary of visual words, where each word in the vocabulary is quantitatively described by the corresponding cluster. We empirically identify a number of clusters based on the standard square-error partitioning method, k-means (Lloyd, 1982). In this step, a vocabulary V= {V_1, …,V_k} of size k is obtained. Descriptors sharing the same center will be represented by the same word.

Unlike the vocabulary of a text corpus whose size is relatively fixed, the size of a vocabulary for protein binding sites is controlled by the number of feature clusters generated by k-means. With a small vocabulary, the visual word is not very discriminative because dissimilar feature points can be mapped to the same visual word. As the vocabulary size increases, the feature is expected to be more discriminative; however, it becomes possible for similar feature points to be mapped to different visual words. As there is no standard way to select the appropriate size of a vocabulary, we empirically select k based on the experiment for various vocabulary sizes (see Section 3.3).

2.4 Surface visual words representation

The surface visual words representation consists of two parts: (i) single-word frequency profile (SFP) and (ii) pair-word frequency profile (PFP).

We first compute the frequency of occurrence of each individual visual word for a binding site A. Given a vocabulary V, for each sample face $i \in$A with descriptor W_i^A, we calculate the Euclidean distance between W_i^A (i= 1,…,M) and each visual word V_j (j= 1,…,k). We then associate face i with the index of the visual word with the smallest distance from the vocabulary V.

The SFP of binding site A is a $k \times$ 1 vector $P^A = (p_1^A,\ldots,p_k^A)$, which is defined as the frequency of each word in A. An example of the SFP for three binding sites with k = 400 is shown in Supplementary Figure S1. As these binding sites are from the database SCOPPI, we use the same identifier as (Sommer *et al.*, 2007), to name each binding site: <PDB-ID>_<SCOP-domain of the binding site>_<Chain-ID of the binding site>_<SCOP-domain of the binding partner>_<Chain-ID of the binding partner>. The group ID for each binding site is labeled as: <SCOP-family of the binding site >_<cluster ID within the SCOP family>. In Supplementary Figure S1, the binding sites 1m3d_78535_B_78538_C and 1t60_106525_V_106528_W are from same SCOPPI group (ID: d.169.1.6_1) whereas the binding site 1cer_29989_R_39901_O is from another SCOPPI group (ID: c.2.1.3_25). From the figure, we can see that the SFPs of 1m3d_78535_B_78538_C and 1t60_106525_V_106528_W are very similar whereas they are starkly different from the SFP of 1cer_29989_R_39901_O.

The pair-word spatial-sensitive frequency profile (PFP) of binding site A, extended from (Haralick *et al.*, 1973) for texture analysis of images, is based on the idea that the surface of a binding site can be characterized spatially by measuring the local distribution of pairs of visual words that are within a given distance d in a given direction θ for a sample point. Given a sample point with a visual word index V_i, the frequency $P_{d,\theta}$ (i, j) is calculated by accumulating the occurrences of a pair of visual words that have visual word indices of visual word (V_i, V_j) and are located within a distance d in direction θ (Fig. 3). The following measures [shown in Equations (1)–(7)] are computed for each binding site surface:

$$\text{Energy} = \sum_{(i,j)} P_{d,\theta}(i,j)^2 \tag{1}$$

$$\text{Entropy} = \sum_{(i,j)} -P_{d,\theta}(i,j) \log P_{d,\theta}(i,j) \tag{2}$$

$$\text{Homogeneity}_1 = \sum_{(i,j)} \frac{P_{d,\theta}(i,j)}{1+|i-j|} \tag{3}$$

$$\text{Homogeneity}_2 = \sum_{(i,j)} \frac{P_{d,\theta}(i,j)}{1+|i-j|^2} \tag{4}$$

$$\text{Contrast} = \sum_{(i,j)} |i-j|^2 P_{d,\theta}(i,j) \tag{5}$$

$$\text{Correlation} = \sum_{(i,j)} \frac{(i-\mu)(j-\mu)P_{d,\theta}(i,j)}{\sigma^2} \tag{6}$$

$$\text{ClusterTendency} = \sum_{(i,j)} (i+j-2\mu)^2 P_{d,\theta}(i,j) \tag{7}$$

where $\mu = \Sigma_{(i,j)} i P_{d,\theta}(i,j)$ and $\sigma = \Sigma_{(i,j)}(i-\mu)^2 P_{d,\theta}(i,j)$. Here, the distance d is defined as 50% of the maximum Euclidean distance between the coordinates of all visual words on the binding sites, and the direction θ is defined as the angle between the two corresponding normal vectors. As there is no particular purpose required to retain the θ dependence, we compute the above mentioned measures with θ = 0, 45, 90, 135 and 180° and take the average over these angles.

2.5 Scoring function

For a given query binding site A, the scoring function is used to assign a similarity score of a retrieved binding site B from the database. We first define the score of SFP as follow:

$$S_{\mathrm{SFP}}(A,B)=1-\mathrm{COS}^{-1}\left(\frac{<P^A,P^B>}{||P^A||\bullet||P^B||}\right) \tag{8}$$

where P^A and P^B are the SFPs for binding sites A and B, respectively. $<P^A, P^B>$ is the inner product of P^A and P^B, and $||P^A||$ and $||P^B||$ are the norms of P^A and P^B, respectively.

In addition, we define a score for PFP as follows:

$$S_{\mathrm{PFP}}(A,B)=1-\frac{||T^A-T^B||}{\underset{i}{\mathrm{argmax}}(||T^A-T^i||)} \tag{9}$$

where T^A and T^B are the PFPs for binding sites A and B, respectively. The distance $||T^A-T^B||$ is normalized by the maximum distance between A and the database binding sites T^i (i= 1,...,L) where L is the size of database.

Finally, we also look at the ratio of areas between two binding sites A and B. The intuition is that if two binding sites display a significant difference in surface area, they will not be geometrically similar. We define the scoring function regarding area of binding sites as follows:

$$S_{\mathrm{Area}}(A,B)=\frac{\min(R^A,R^B)}{\max(R^A,R^B)} \tag{10}$$

where R^A and R^B are areas of binding sites A and B, respectively.

An aggregated similarity function is defined to include the above three scoring functions to provide a final similarity score:

$$S_{\mathrm{BS}}(A,B)=w_1\times S_{\mathrm{SFP}}+w_2\times S_{\mathrm{PFP}}+w_3\times S_{\mathrm{Area}}, \tag{11}$$

where w_1, w_2 and w_3 are used to weight the contributions from the three similarity terms.

3 RESULTS

To demonstrate the significance of this work, we compare the performance to current methods in terms of accuracy and efficiency. The current methods include an alignment-based method, iAlign (Gao and Skolnick, 2010), which is designed for PPI comparison based on the local substructures of subunits, and a feature-based method, MI (Sommer *et al.*, 2007), which describes the geometric shape of a binding site using a feature vector based on moment invariants.

We apply PBSword, iAlign and MI to a non-redundant dataset from SCOPPI to evaluate classification and retrieval performance for protein–protein binding sites. During the experiment, a query dataset is selected and used to retrieve similar binding sites from the entire database.

To evaluate classification accuracy, we use a general metric, the correct classification rate (CCR), which is defined as follows:

$$\mathrm{CCR}=\frac{\text{The number of correctly classified binding sites}}{\text{The total number of test binding sites}} \tag{12}$$

In addition, for each query, we identify the top results whose SCOPPI group matches the query's group. The ranks of these 'top results' are then accumulated with summary statistics reported.

Table 1. CCR of PBSword and MI for Q_1 and D_1

Top	1	5	10	1%
PBSword	0.76	0.88	0.91	0.95
MI	0.40	–	–	–

Comparison of the CCR for PBSword and MI. The result of MI is from caption of Figure1 in Sommer *et al.* (2007). CCR of PBSword is calculated for top 1, 5, 10 and 1% ranked results.

We also use the AUC (area under curve) in the ROC (receiver operator characteristics) curve (Bradley, 1997) to measure how well each method identifies the other binding sites from the query's group. An AUC of 1.0 would correspond to perfect classification, which would rank the binding sites from the same group as the query's before all other binding sites, whereas an AUC of 0.5 would be expected for a random classifier.

3.1 Protein binding site classification and retrieval

The dataset used in this experiment, denoted as D_1, is a non-redundant dataset of protein binding sites extracted from SCOPPI 1.69 and has been used to evaluate the performance of MI. Dataset D_1 consists of 2819 protein binding sites clustered into 501 groups, as determined in Sommer *et al.* (2007). The query dataset, denoted as Q_1, includes 224 binding sites from 53 groups that are selected from D_1 by applying a structural alignment tool, TM-align (Zhang and Skolnick, 2005), to ensure the similarity score (i.e. TM-score) among the binding sites within one group was >0.45 [for a more detailed description of D_1, see (Sommer *et al.*, 2007)].

We test our method with a vocabulary size (k) of 400 on all protein binding sites from Q_1. As each binding site in D_1 belongs to a group of geometrically similar binding sites, our evaluation is to measure how well we can correctly classify members from the same group. For a query binding site, we rank each binding site in D_1 (with the query binding site excluded) based on similarity score [Equation (11)]. In a perfect scenario, the query binding site would be classified into the same group as the top-ranked binding site. Table 1 presents the CCR performance comparison of PBSword and MI for Q_1 and D_1. Intuitively, the optimal accuracy of classification is 100% CCR. Our classification results show that when using only the top rank, PBSword achieves a 76% CCR, which is a significant improvement over the 40% CCR reported by MI in its original paper (Sommer *et al.*, 2007). We also investigate the performance of PBSword by examining up to the top 1% (=28) of ranked results. In these situations, if one protein binding site from the same group as query can be found in a certain range, the query binding site is regarded as correctly classified. As shown in Table 1, CCRs are 88 and 91% when the top 5 and top 10 ranks are considered, respectively.

Table 2 shows summary statistics for the best hits from PBSword and MI for Q_1 and D_1. Here, the quartiles are values that divide a set of data into four equal parts. From the table, we can see that the third quartile of PBSword is still 1, as opposed to 18.3 for MI, which represents a significant improvement. In the worst case, PBSword finds a binding site from the same group as the query at rank 587 (top 20%).

Table 3 shows summary statistics for the AUC of PBSword and MI for Q_1 and D_1. From the table, we can see PBSword also outperforms MI in terms of AUC. In the worst case, PBSword can achieve

Table 2. Summary statistics for best hits from PBSword and MI for Q_1 and D_1

	Min	1st Q	2nd Q	3rd Q	Max
PBSword	1	1	1	1	587
MI	1	1	2	18.3	1902

The results of MI are from Sommer *et al.* (2007) The columns 'Min' and 'Max' represent the minimum and maximum rank of the best hit, and the columns '1st Q', '2nd Q' and '3rd Q' correspond to the first, second and third quartile, respectively.

Table 3. Summary statistics for the AUC from PBSword and MI for Q_1 and D_1

Methods	Min	1st Q	2nd Q	3rd Q	Max
PBSword	0.66	0.94	0.99	1	1
MI	0.30	0.94	0.98	0.99	1

Comparison of the AUC for PBSword and MI. The results of MI are from Sommer *et al.* (2007). The columns 'Min' and 'Max' represent the minimum and maximum value of AUC, and the columns '1st Q', '2nd Q', and '3rd Q' correspond to the first, second and third quartile, respectively.

Table 4. CCR of PBSword and iAlign for Q'_1 and D_1

Top	1	5	10	1%
PBSword	0.77	0.88	0.92	0.95
iAlign	0.82	0.99	1	1

Comparison of the CCR for PBSword and iAlign. CCR is calculated for top 1, 5, 10 and 1% ranked results.

Table 5. Summary statistics for the best hits from PBSword and iAlign for Q'_1 and D_1

	Min	1st Q	2nd Q	3rd Q	Max
PBSword	1	1	1	1	587
iAlign	1	1	1	1	6

The same columns of Table 2 are shown.

Table 6. Summary statistics for the AUC from PBSword and iAlign for Q'_1 and D_1

	Min	1st Q	2nd Q	3rd Q	Max
PBSword	0.75	0.98	0.99	1	1
iAlign	0.69	1	1	1	1

The same columns of Table 3 are shown.

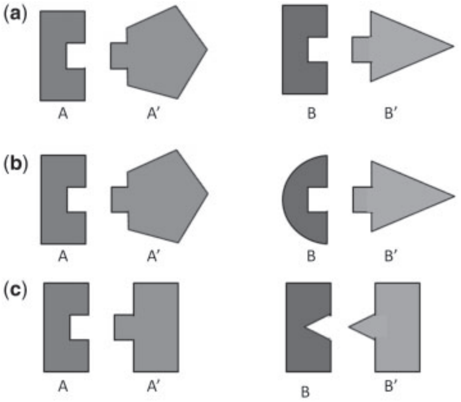

Fig. 4. Schematic representations of the three case studies. The protein chains A (green) and A' (magenta) are from the query protein while B (blue) and B' (orange) are from the database protein. (**a**) A and B are from same SCOP family whereas A' and B' are from different families. The binding site surface on A is geometrically similar as that on B. (**b**) A, A', B, and B' are all from different SCOP families but the two binding site surfaces are similar. (**c**) A, A', B, and B' are all from same SCOP family, but two interactions have dissimilar binding surfaces which correspond to the different functions carried out by each protein complex. A color version of this figure is available as Supplementary Material.

an AUC of 0.66, which is better than a random classifier (0.5). Supplementary Figure S2 presents the histogram of AUC values for PBSword. With our method, for 25% of the queries, we can identify all the members from the same group without any false positives, which significantly outperforms MI (5%) and improves by 20%.

In addition to comparing the feature-based method MI, we also perform the same experiments using the alignment-based method, iAlign, on Q_1 and D_1. Unfortunately, iAlign cannot always find all the alignments of binding sites from the dataset; hence, to facilitate the comparison between PBSword and iAlign, we construct a reduced query dataset, Q'_1, by removing those groups from the query dataset that contained the binding sites iAlign could not find. This results in the exclusion of 10 groups (ID: a.2.11.1_3, c.1.9.2_7, c.1.14.1_1, c.48.1.2_2, d.8.1.1_2, d.8.1.1_5, d.19.1.1_36, d.58.33.1_6, d.153.1.4_13 and f.21.1.2_15) from the query dataset Q_1, which left 188 binding sites clustered into 43 groups. For completeness, the performance of PBSword on the excluded binding sites $\{Q_1 - Q'_1\}$ is reported in Supplementary Tables S1 and S2.

Table 4 presents the CCR performance comparison between PBSword and iAlign for the reduced query dataset Q'_1. As can be seen, PBSword achieves a 77% CCR, which is relatively close to the 82% CCR of iAlign. We further investigate the CCR performance by examining up to the top 1% of ranked results. By using the top five results, PBSword can achieve a CCR of 88%, which is comparable to the CCR using top result from iAlign (82%). Thus, on this dataset, instead of performing one-against-all structural alignment using iAlign, we can achieve a similar classification accuracy by checking the top five results (0.1% of the entire database of binding sites) generated by PBSword. Table 5 presents the summary statistics for the best hits from these two methods. Though it is noted that the worst rank of the best hit from PBSword is 587, which is worse than iAlign, PBSword still filters out ~80% of the dissimilar binding sites. By performing iAlign on the remaining 20% of binding sites in the database, we cannot only achieve a similar classification accuracy as iAlign, but can also save ~80% in computing resources compared with a one-against-all alignment using iAlign. Table 6 shows the summary statistics for the AUC. In the worst case, PBSword obtains an AUC of 0.75, which is better than that of iAlign (0.69). From these experimental results, we can see that PBSword provides an efficient way to filter out geometrically dissimilar binding sites and obtain a short list of similar ones, which can be sent to the existing alignment-based methods for refinement.

We also evaluate performance of PBSword using a non-redundant dataset selected from SCOPPI 1.75, which is denoted as D_2. The experimental results of D_2 can be found in Supplementary Material S1.

3.2 Efficiency

We measure the average response time for 188 binding sites from Q'_1 to evaluate the efficiency of PBSword. The experiments are conducted on a Linux Fedora server with AMD Opteron dual-core 1000 series processors and 8 GB RAM. With PBSword, each query takes 0.31 s for one-against-all score calculations. For iAlign, however, each query takes 1016 s to scan the entire database D_1. Note that we have excluded the CPU time spent on generating the surface and calculating the visual words, as it can be performed off-line during the preprocessing stage. An efficiency comparison with I2ISiteEngine, another alignment-based method, was not performed in this article, though we note that I2ISiteEngine has been evaluated elsewhere (Gao and Skolnick, 2010) where experimental results showed that iAlign can achieve about an 89-fold speedup over I2ISiteEngine.

3.3 Parameter selection

The performance of our method is heavily dependent on the vocabulary size (k). To study the influence of k on the results, we carried out three experiments on D_1 with k= 200, 400 and 600. The corresponding CCR of top rank is 69, 76 and 74%, respectively. As such, we selected k= 400 as our default settings.

4 STUDYING PROTEIN STRUCTURES FROM A GEOMETRIC PERSPECTIVE

We have shown that PBSword can identify geometrically similar protein binding sites from the same SCOP family based on surface shape features. As the molecular shape has long been recognized as a key factor in protein–protein interactions, we further investigate whether PBSword can discover non-trivial biological connections among proteins from a geometric perspective. In this section, we first study whether our approach can help us to investigate the relationships between the geometrically similar shapes of protein binding sites participating in an interaction and the functions carried out by the interactions. Specifically, we use our approach to first retrieve geometrically similar binding sites for a ‘seed’ binding site A, and then select the top-ranked binding site B to analyze the functional similarity between the corresponding proteins. The binding partners of A and B are denoted as A′ and B′, respectively. We consider two cases: (i) A and B are from same SCOP family, whereas A′ and B′ are from different families; and (ii) A, A′, B and B′ are all from different SCOP families. We then study the relationships between the shapes of protein binding sites and functional diversity within a SCOP family. Intuitively, proteins from the same family are expected to be structurally similar and have related functions. Discovering a protein binding site from such a family would not be very biologically significant, since the binding sites from the structurally similar proteins are expected to be similar and clustered together. What would be more interesting would be the discovery of two protein binding sites which are from the same family, but have dissimilar geometric shapes and different molecular functions. For this study, we consider another case: (iii) A, A′, B and B′ are from the same SCOP family but belong to different functional groups. In this case, the binding site B is not the top-ranked, but the highest ranking result from the same family. The schematic representations for the three cases are shown in Figure 4.

To study the structure–function relationship based on the geometric feature of binding site, we select another non-redundant dataset of protein–protein interfaces generated by all-against-all interface comparisons of protein complexes from PDB using I2ISiteEngine (Mintz *et al.*, 2005). This dataset, denoted as D_3, consists of 604 protein–protein interfaces clustered into 59 groups. In each group, interface members share a sequence identity of <50%. As the binding site in this dataset is defined on the protein chain, we use <PDB-ID>_<Chain ID of the binding site><Chain ID of the binding site partner> to represent a protein binding site.

In the first case, the protein binding site 4sgb_EI is selected as a query to retrieve similar binding sites from D_3 with our method. The top result is 1sgd_EI from the same group (Fig. 5a). The protein chains 4sgb_E and 1sgd_E are all from the prokaryotic proteases family, whereas the chain 4sgb_I and 1sgd_I belong to the plant proteinase inhibitors family and ovomucoid domain III-like family, respectively. The alignment results of TM-align and iAlign are shown in Figure 5b and c, which show that two protein chain pairs can be well aligned. The TM-score of the two aligned protein structures is 0.99, and the IS-score of the aligned interfaces is 0.64, which are statistically significant (Gao and Skolnick, 2010; Xu and Zhang, 2010). The binding site surfaces with the chemical environments (i.e. electrostatic potential), shown in Figure 5d, are also similar. In this case, 4sgb_I and 1sgd_I belong to different families, but can bind to the equivalent sites of homologous 4sgb_E and 1sgd_E, respectively. This type of interactions, also known as convergently evolved interaction motifs, would be very biologically significant as it can be used to discover rules for interactions and design ligand (Henschel *et al.*, 2006).

In the second case, we select protein binding site 1b99_AD. In D2, 1b99_AD has been classified into a group including 1l0o_AB, 1l3b_AD, 1e7p_AD, 1gtt_BC and 1iun_AB (Mintz *et al.*, 2005). The top result from PBSword is 1tmk_BA, which is from another group. The proteins 1b99 and 1tmk are from the nucleoside diphosphate kinase family and nucleotide and nucleoside kinases family, respectively. However, both proteins are kinase and mainly composed of α-helices and β-strands (Supplementary Fig. S4A). The sequence alignment between the 1b99 and 1tmk binding chains shows low similarity. The global structure alignment using TM-align is given in Supplementary Figure S4B, which shows that the residues from the binding chains 1b99_D and 1tmk_A cannot be aligned together. Hence, we obtain a TM-score of 0.25, which is not statistically significant. In contrast, iAlign can find partial residue correspondences between the two interfaces (Supplementary Fig. S4C) with an IS-score of 0.24 (p-value = 0.42 × 10^{-2}). As a comparison, the binding site surfaces of 1b99_AD and 1tmk_BA are shown in Supplementary Figure S4D. Despite the sequence and global structure conservation being low, the binding site surfaces of 1b99_AD and 1tmk_BA are geometrically similar, which is effectively captured by PBSword.

In the third case, we select the family of C-type lectin domains (SCOP ID d.169.1.1). In 1993, Drickamer first classified proteins in

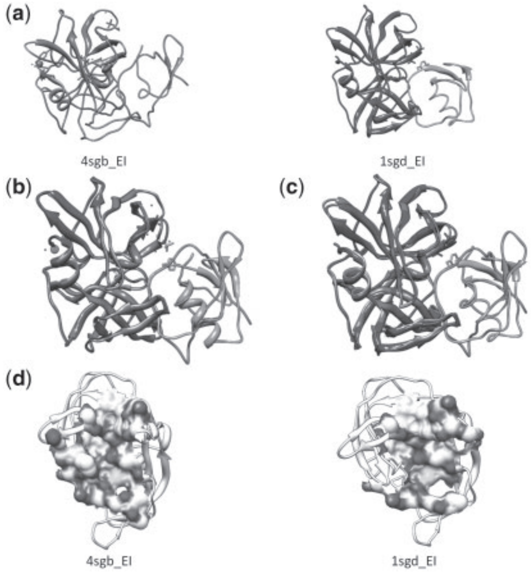

Fig. 5. (**a**) Protein structures of 4sgb_EI (chain E-green, chain I-magenta) and 1sgd_EI (chain E-blue, chain I-orange). The protein chains 4sgb_E and 1sgd_E are both from the prokaryotic proteases family, while the chains 4sgb_I and 1sgd_I belong to the plant proteinase inhibitors family and ovomucoid domain III-like family, respectively. (**b**) Global structure alignment of 4sgb_EI and 1sgd_EI using TM-align (TM-score = 0.99). (**c**) Local substructure alignment of 4sgb_EI and 1sgd_EI using iAlign (IS-score = 0.64). (**d**) Binding site surfaces with the chemical environments (i.e., electrostatic potential) of 4sgb_EI and 1sgd_EI. In this case, 4sgb_I and 1sgd_I belong to different families, but can bind to the equivalent sites of homologous 4sgb_E and 1sgd_E, respectively. This type of interaction is also known as convergently evolved interaction motifs. A color version of this figure is available as Supplementary Material.

this family into seven groups and showed that such classification can capture functional similarities between proteins (Drickamer, 1993). The classification was later revised (Drickamer and Fadden, 2002; Zelensky and Gready, 2005), and currently this family has 17 groups. The spatial relationships of protein binding sites in this family were analyzed in (Korkin *et al.*, 2005), and significant diversity was found. In this case, protein binding site 1k9i_68344_B68349_G from this family, which is not included in the query dataset Q_1, is selected and compared with the sites from D_1. In D_1, binding site 1bv4_42381_B42383_C (Supplementary Fig. S5A) from the same family is recognized as having a similar interface type and is classified into the same group (ID: d.169.1.1_21) according to the sequence and structure alignment methods employed by SCOPPI. With our approach, the binding site 1bv4_42381_B42383_C is ranked as 2334. The structure alignments of TM-align and iAlign are shown in Supplementary Figure S5B and C, respectively. As these two bind sites belong to same family, we can find significant similarity from TM-align (TM-score = 0.86). However, the IS-score from iAlign is only 0.11 (p-value = 0.84). The binding site surfaces, shown in Supplementary Figure S5D, are remarkably dissimilar. The functional groups of 1k9i and 1bv4 are also different. According to the classification of functional groups in (Zelensky and Gready, 2005), protein 1k9i belongs to the asialoglycoprotein and DC receptors group whereas protein 1bv4 belongs to the collectins group. From this case, we can see that the surface dissimilarity detected by PBSword can be potentially used to discover functional diversity between two proteins from the same family.

5 CONCLUSIONS

We have presented PBSword, a novel method for protein binding site characterization and comparison based on the distribution of visual words of surfaces. The proposed method complements existing alignment-based approaches in the analysis of protein–protein interactions. The method is applied to evaluate the classification and retrieval performance of protein–protein binding sites and is compared with an alignment-based method (iAlign) and to a feature-based method using moment invariants. The results show that PBSword can achieve comparable classification accuracy to the alignment-based methods with greatly improved efficiency.

We emphasize that PBSword, as a feature-based method for fast filtering of similar binding sites from a large dataset, is not designed to be a replacement of existing alignment-based methods. Instead, PBSword works as a complementary approach to the various structure comparison methods and offers an efficient way to classify and retrieve geometrically similar binding sites.

Our future work includes (i) the development of a more comprehensive scoring function for the surface comparison that takes into consideration physicochemical properties, and (ii) extension to the protein–ligand binding site comparison and retrieval.

ACKNOWLEDGEMENTS

We would like to thank the authors of iAlign, SCOPPI and MI who provided openly accessible tool and dataset for research use.

Funding: The Shumaker Endowment in Bioinformatics to B.P. and N.Z., and the National Science Foundation (DBI-0845196) to D.K.

Conflict of Interest: none declared.

REFERENCES

Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, **7**, 188–197.

Bahadur,R. and Zacharias,M. (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell. Mol. Life Sci.*, **65**, 1059–1072.

Belongie,S. *et al*. (2002) Shape matching and object recognition using shape contexts. *IEEE T Pattern Anal. Mach. Intell.*, **24**, 509–522.

Bradford,J.R. *et al.* (2006) Insights into protein-protein interfaces using a Bayesian network prediction method. *J. Mol. Biol.*, **362**, 365–386.

Bradley,A. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, **30**, 1145–1159.

Bronstein,A. *et al.* (2011) Shape google: geometric words and expressions for invariant shape retrieval. *ACM Trans. Graph.*, **30**, 1–20.

Budowski-Tal,I. *et al.* (2010) FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc. Natl Acad. Sci. USA*, **107**, 3481–3486.

Das,S. *et al.* (2009) Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model*, **49**, 2863–2872.

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

Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numer. Math.*, **1**, 269–271.

Drickamer,K. (1993) Evolution of Ca(2+)-dependent animal lectins. *Prog. Nucleic Acid Res. Mol. Biol.*, **45**, 207–232.

Drickamer,K. and Fadden,A.J. (2002) Genomic analysis of C-type lectins. *Biochem. Soc. Symp.*, 59–72.

Finn,R.D. *et al*. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Gao,M. and Skolnick,J. (2010) iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics*, **26**, 2259–2265.

Haralick,R. *et al*. (1973) Textural features for image classification. *IEEE T Syst. Man Cybern.*, **3**, 610–621.

Henschel,A. *et al*. (2006) Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics*, **22**, 550–555.

Kerrien,S. *et al.* (2012) The IntAct molecular interaction database in 2012, *Nucleic Acids Res.*, **40**, D841–D846.

Keskin,O. and Nussinov,R. (2007) Similar binding sites and different partners: implications to shared proteins in cellular pathways, *Structure*, **15**, 341–354.

Keskin,O. *et al.* (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Kim,W.K. *et al.* (2006) The many faces of protein-protein interactions: a compendium of interface geometry. *PLoS Comput. Biol.*, **2**, e124.

Korkin,D. *et al*. (2005) Localization of protein-binding sites within families of proteins. *Protein Sci.*, **14**, 2350–2360.

Kuang,X. *et al.* (2012) DOMMINO: a database of macromolecular interactions. *Nucleic Acids Res.*, **40**, D501–D506.

Liu,Y. *et al*. (2009) IDSS: deformation invariant signatures for molecular shape comparison. *BMC Bioinform.*, **10**, 157–170.

Lloyd,S. (1982) Least squares quantization in PCM. *IEEE T Inform. Theory*, **28**, 129–137.

Merelli,I. *et al.* (2011) Image-based surface matching algorithm oriented to structural biology. *IEEE/ACM T Comput. Biol. Bioinform.*, **8**, 1004–1016.

Mintz,S. *et al.* (2005) Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins*, **61**, 6–20.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nagano,N. *et al*. (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.

Osada,R. *et al.* (2001) Matching 3D models with shape distributions. In *Proceedings of the International Conference on Shape Modeling & Applications*. IEEE Computer Society, Genoa, Italy, p. 154.

Sael,L. *et al.* (2008) Rapid comparison of properties on protein surface. *Proteins*, **73**, 1–10.

Sander,O. *et al.* (2008) Structural descriptors of protein-protein binding sites. In *Proceedings of 6th Asia-Pacific Bioinformatics Conference*. Imperial College Press, London, pp. 79–88.

Sanner,M.F. *et al.* (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.

Shulman-Peleg,A. *et al.* (2004) Protein-protein interfaces: recognition of similar spatial and chemical organizations. In Jonassen,I. and Kim,J. (eds) *Algorithms in Bioinformatics*. Springer Berlin/Heidelberg, pp. 194–205.

Sims,G.E. *et al.* (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl Acad. Sci. USA*, **106**, 2677–2682.

Sommer,I. *et al.* (2007) Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*, **23**, 3139–3146.

Tsai,C.J. *et al.* (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.

Tuncbag,N. *et al.* (2008) Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.*, **381**, 785–802.

Winter,C. *et al.* (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.

Wu,C.H. *et al.* (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96.

Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.

Yin,S. *et al.* (2009) Fast screening of protein surfaces using geometric invariant fingerprints *Proc. Natl Acad. Sci. USA*, **106**, 16622–16626.

Zelensky,A.N. and Gready,J.E. (2005) The C-type lectin-like domain superfamily. *FEBS J.*, **272**, 6179–6217.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

Zhao,N., *et al.* (2011) Structural similarity and classification of protein interaction interfaces. *PLoS One*, **6**, e19554.