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.met.missouri.edu/pbsword/pbsword.html.

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 (PFIs) Bahadur and Zacharias, 2008; Kim et al., 2006). The PFIs 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 (Alloy and Russell, 2006). Several protein–protein interaction-related databases have been constructed, such as SCOPPI (Wintet et al., 2006), PBASE (Davis and Sali, 2005), IntAct (Kerrien et al., 2012), iPam (Finn et al., 2005) and DOMMINO (Kuang et al., 2012). Particularly, the SCOPPI database provides an evolutionary and structural classification of PFIs 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 indispensable 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 Cα 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 PFIs (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-Peled 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...
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/Å² 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 4 Å 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...
Assuming binding site A has $N_A$ color version of this figure is available as Supplementary Material.

and say that binding site A is characterized by its feature descriptor $F_i$ face geodesic distance from all pair-wise distances within each binding site.

translation and scaling. In total, we extract $M$ distances bins and $N$ angles bins, which are set to 10 and 8, respectively. By respectively. The geodesic distance $g_{ij}$ is defined as

$g_{ij} = \min_{\phi \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ\}} \sum_{k=1}^{n_i} d_{ik}$

$\theta_{i,j} = \frac{\sum_{k=1}^{n_i} d_{ik} \cdot \cos(\theta_{i,j})}{\sum_{k=1}^{n_i} d_{ik}^2}$

$\sigma = \sum_{i,j} d_{ij}^2$ (5)

$\mu = \sum_{i,j} d_{ij} \cdot \mu$ (6)

$\sigma = \sum_{i,j} d_{ij}^2$ (7)

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 $i$ with descriptor $\Psi_i^4$, we calculate the Euclidean distance between $\Psi_i^4$ 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 $P^4 = [p_1,...,p_k]$, 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>-<Chain-ID of the binding partner>-<SCOP-chain 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_7B5B_3B8C, 1h01_106525_V_/106528, and 1t60_106525_V_/106528 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 $\theta$ for a sample point. Given a sample point with a visual word index $i$ in $V$, the frequency $P_{ij}$ 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 $\theta$ (Fig. 3). The following measures shown in Equations (8)–(11) are computed for each binding site surface:

![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 $\alpha_{ij}$ is calculated between $n_i$ and a unit vector pointing from $r_i$ to $r_j$. A color version of this figure is available as Supplementary Material.](image1)

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 $\alpha_{ij}$ is calculated between $n_i$ 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 FFP for the sample point i, a ball with radius $d$ is first placed at $r_i$. Then, the occurrence of word pairs with index ($V_i$, $V_j$) ($j=1,...,M$) within the ball is counted for different directions $\theta$ ($\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ$) separately. Here, the direction $\theta$ is defined as the angle between two corresponding normal vectors.](image2)

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 FFP for the sample point i, a ball with radius $d$ is first placed at $r_i$. Then, the occurrence of word pairs with index ($V_i$, $V_j$) ($j=1,...,M$) within the ball is counted for different directions $\theta$ ($\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ$) separately. Here, the direction $\theta$ is defined as the angle between two corresponding normal vectors.
B. Pang et al.

coordinates of all visual words on the binding sites, and the direction \( \theta \) is defined as the angle between the two corresponding normal vectors. As there is no particular purpose required to retain the \( \theta \) dependence, we compute the above mentioned measures with \( \theta = 0, 45, 90, 135 \) and 180\(^\circ\) 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:

\[
S_{\text{SFP}}(A, B) = 1 - \text{COS}^{-1} \left( \frac{\langle P^A, P^B \rangle}{\|P^A\| \times \|P^B\|} \right)
\]

where \( P^A \) and \( P^B \) are the SFPs for binding sites \( A \) and \( B \), respectively. \( \langle P^A, P^B \rangle \) 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:

\[
S_{\text{PFP}}(A, B) = 1 - \frac{\|P^A - T^B\|}{\max_{i \in T} \|P^A - T^B\|}
\]

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

\[
S_{\text{area}}(A, B) = \frac{\text{min}(R^A, R^B)}{\text{max}(R^A, R^B)}
\]

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:

\[
S_{\text{aggr}}(A, B) = w_1 \times S_{\text{SFP}} + w_2 \times S_{\text{PFP}} + w_3 \times S_{\text{area}}
\]

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 and scoring functions to provide a final similarity score:

\[
(Gao \text{ and Skolnick}, 2010), \text{which is designed for PPI comparison}
\]

performance to current methods in terms of accuracy and efficiency.

entire database.

dataset is selected and used to retrieve similar binding sites from the
database. During the experiment, a query

to evaluate classification and retrieval performance
invariants.

We apply PBSword, iAlign and MI to 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 2810 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 1. CCR of PBSword and MI for \( Q_1 \) and \( D_1 \)

<table>
<thead>
<tr>
<th>Top</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSword</td>
<td>0.76</td>
<td>0.88</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>MI</td>
<td>0.40</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Comparison of the CCR for PBSword and MI. The result of MI is from caption of Figure 1 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. We then used to evaluate the performance of MI. Dataset \( D_1 \) consists of 2810 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...
Table 2. Summary statistics for best hits from PBSword and MI for Q1 and D1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Min</th>
<th>1st Q</th>
<th>2nd Q</th>
<th>3rd Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSword</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>587</td>
</tr>
<tr>
<td>MI</td>
<td>1</td>
<td>2</td>
<td>18.3</td>
<td>1902</td>
<td></td>
</tr>
</tbody>
</table>

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 Q1 and D1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Min</th>
<th>1st Q</th>
<th>2nd Q</th>
<th>3rd Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSword</td>
<td>0.66</td>
<td>0.94</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MI</td>
<td>0.30</td>
<td>0.94</td>
<td>0.98</td>
<td>0.99</td>
<td>1</td>
</tr>
</tbody>
</table>

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 Q1 and D1

<table>
<thead>
<tr>
<th>Top</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSword</td>
<td>0.77</td>
<td>0.88</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>iAlign</td>
<td>0.82</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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 D1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Min</th>
<th>1st Q</th>
<th>2nd Q</th>
<th>3rd Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSword</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>587</td>
</tr>
<tr>
<td>iAlign</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

The same columns of Table 2 are shown.

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 Q1 and D1. 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.1.11.1_36, d.58.3.1_6, d.153.1.4_13 and f.21.1.2_15) from the query dataset Q1, which left 188 binding sites clustered into 43 groups. For completeness, the performance of PBSword on the excluded binding sites {Q1 − Q′1} is reported in Supplementary Tables S1 and S2.

Table 6. Summary statistics for the AUC from PBSword and iAlign for Q′1 and D1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Min</th>
<th>1st Q</th>
<th>2nd Q</th>
<th>3rd Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSword</td>
<td>0.75</td>
<td>0.98</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>iAlign</td>
<td>0.69</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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.
The experimental results of I2ISiteEngine show that iAlign can achieve about an 89-fold speedup over elsewhere (Gao and Skolnick, 2010) where experimental results were carried out three experiments on D1 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 the same SCOP family, whereas A’ and B’ are from different families; and (ii) A’, B’ and A 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 D2, 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 4ggb_E is selected as a query to retrieve similar binding sites from D2 with our method. The top result is 1sgd_EI from the same group (Fig. 5a). The protein chains 4ggb_E and 1sgd_E are all from the prokaryotic proteases family, whereas the chain 4ggb_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, 4ggb_I and 1sgd_I belong to different families, but can bind to the equivalent sites of homologous 4ggb_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 1b99_AB, 1b99_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 = $0.42 \times 10^{-2}$). 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...
With our approach, the binding site 1bv4_42381_B42383_C is 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


B. Pang et al.


