An Integrated Approach to Sequence-Independent Local Alignment of Protein Binding Sites

Bin Pang, David Schlessman, Xingyan Kuang, Nan Zhao, Daniel Shyu, Dmitry Korkin, and Chi-Ren Shyu

Abstract—Accurate alignment of protein-protein binding sites can aid in protein docking studies and constructing templates for predicting structure of protein complexes, along with in-depth understanding of evolutionary and functional relationships. However, over the past three decades, structural alignment algorithms have focused predominantly on global alignments with little effort on the alignment of local interfaces. In this paper, we introduce the PBSalign (*P*rotein-protein *B*inding *S*ite *align*ment) method, which integrates techniques in graph theory, 3D localized shape analysis, geometric scoring, and utilization of physicochemical and geometrical properties. Computational results demonstrate that PBSalign is capable of identifying similar homologous and analogous binding sites accurately and performing alignments with better geometric match measures than existing protein-protein interface comparison tools. The proportion of better alignment quality generated by PBSalign is 46, 56, and 70 percent more than iAlign as judged by the average match index (MI), similarity index (SI), and structural alignment score (SAS), respectively. PBSalign provides the life science community an efficient and accurate solution to binding-site alignment while striking the balance between topological details and computational complexity.

Index Terms—Structural bioinformatics, binding site alignment

1 Introduction

Protein-protein binding sites consist of residues on the protein surface through which proteins can interact to form a complex and perform a specific function. Two geometrically and physicochemically complementary binding sites from different protein subunits (chains or domains) can form a protein-protein interaction (PPI) [1]. The main goal of binding site alignment, similar to sequence- and structure-based protein alignment, is to determine similarities between a pair of binding sites so that further functional and evolutionary relationships can be identified. Additionally, the binding site alignment can be employed to improve performance of protein-protein docking [2] and construct templates for protein complex structure prediction [3].

With the development of high-throughput experimental techniques, the size of data repositories of protein-protein interactions has dramatically increased [4]. With this trend arises the need to analyze, compare, and classify protein binding sites using three-dimensional (3D) structural information [5], [6], [7], [8]. Methods have been developed to compare and classify proteins using their overall sequence and structural similarities. However, it is known that the overall sequence and structure similarities of proteins do not necessarily imply similarities in protein binding sites and associated functions. Hence, there is great need for new methods that can compare such function-related local structural similarities of protein binding sites [9].

Comparing protein-binding sites at the three-dimensional level is challenging because the residues of a binding site are not always sequential in nature, resulting in a large search space for possible alignments. To date, only a handful of methods are available for binding site alignments, which is in sharp contrast to the overall structure alignment methods developed in last three decades [10], [11]. A typical binding site alignment method consists of two steps: (i) mapping potential correspondences of residue amongst two binding sites and (ii) searching for an optimal alignment of known residue correspondences. In these approaches, protein binding sites are normally represented using the three-dimensional coordinates of the C_α atoms from the protein backbone [11] or coordinates of functional sites on the local surface [12]. In the first step, residue or surface point correspondences can be established using the structure information [11] or surface features [12], [13]. After that, one of the binding sites is rotated, translated, and superimposed onto the other binding site to obtain possible alignments between them. In the second step, various heuristic algorithms have been employed to search all possible alignment combinations of residues between two binding sites and select the final alignment with

- *B. Pang and X. Kuang are with the Informatics Institute, University of Missouri, Columbia, MO 65211. E-mail: {bptwb, xkf2b}@mail.missouri.edu.*
- *D. Schlessman is with the Department of Computer Science, University of Missouri, Columbia, MO 65201. E-mail: dsdn7@mail.missouri.edu.*
- *N. Zhao is with the Department of Basic Sciences, College of Veterinary Medicine, Mississipi State University, Mississippi State, MS 39762. E-mail: nzhao@cvm.msstate.edu.*
- *D. Shyu is with the Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN 47907. E-mail: dshyu@purdue.edu.*
- *D. Korkin is with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609. E-mail: korkin@korkinlab.org.*
- *C.-R. Shyu is with the Informatics Institute, Departments of Electrical and Computer Engineering and Computer Science, University of Missouri, Columbia, MO 65211. E-mail: shyuc@missouri.edu.*

Manuscript received 9 Apr. 2014; accepted 29 May 2014. Date of publication 9 Sept. 2014; date of current version 3 Apr. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2355208

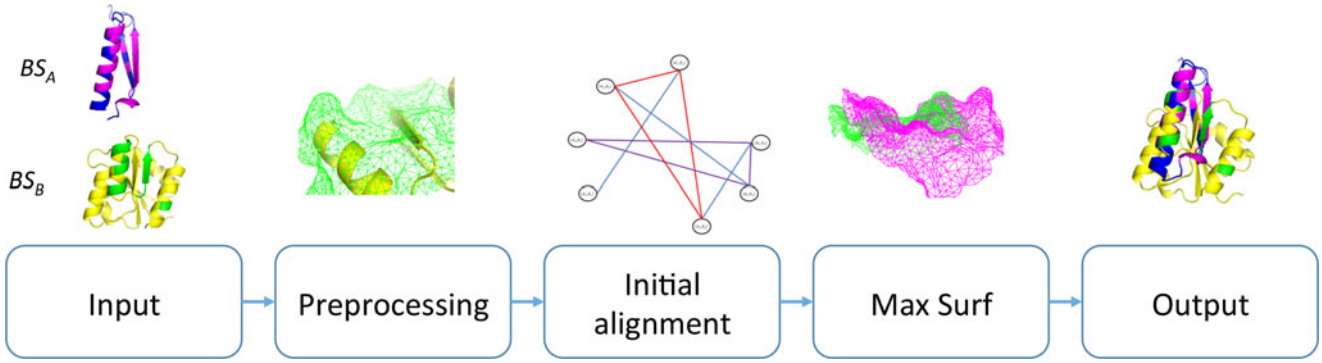


Fig. 1. The input of PBSalign is a pair of protein binding sites. In the preprocessing step, various surface properties are calculated for the binding sites. In the initial alignment step, similarity of properties is used to find potential correspondences between two binding site surfaces and a list of seed alignments is generated by the correspondences, which are refined using MaxSurf. Finally, the outputs include residue correspondences and similarity score of alignment.

the minimization of the root mean square deviation (RMSD). Most heuristic algorithms of binding site alignments are influenced by the global structural alignments [14], [15], [16] or computer vision methods [12], [17], [18].

In this paper, we present a new protein binding site alignment algorithm, namely PBSalign, for comparing a pair of sites based on the surface and structure properties of local regions. In PBSalign, geometrical and physicochemical properties from two binding sites are compared with each other, enabling exploration of different combinations of residue correspondences and grouping of the sets of correspondences into initial alignments. PBSalign employs various techniques to eliminate incompatible residue correspondences and reduce computational complexity. Alignment quality can be significantly improved by refining the initial seed alignments. We conduct computational experiments to compare the alignment quality of PBSalign and the existing method iAlign [11]. The experimental results illustrate PBSalign's superior performance over the existing methods.

2 Methods

The framework of PBSalign is shown in Fig. 1. The input of PBSalign is a pair of protein-protein interfaces: $IA = \{BS_A{}^1, BS_A{}^2\}$ and $IB = \{BS_B{}^1, BS_B{}^2\}$. Each protein-protein interface (PPI) consists of two binding sites (BS) which could come from the same or different fold(s). Without loss of generality, we assume $BS_A{}^1$ and $BS_B{}^1$ are binding site pairs for comparison and use BS_A and BS_B to represent these two sites in the following sections. During the alignment, BS_A will be fixed, and BS_B will be rotated and translated towards BS_A as a rigid body. The outputs include residue correspondences and similarity score of alignment. Similar as the global structure alignment, PBSalign aims to determine an alignment between residues of two given binding sites such that functional and evolutionary relationships between them can be identified.

PBSalign mainly consists of three steps (see Fig. 1): preprocessing, initial alignment, and MaxSurf. In the preprocessing step, a binding site surface is generated and various surface properties are calculated. In the initial alignment step, similarity of properties is used to find potential correspondences between two binding site surfaces and a list of seed alignments is generated by the correspondences. Finally, the seed alignments are refined using MaxSurf, an algorithm developed to find maximal overlapping surface of two binding sites while minimizing RMSD of alignment.

2.1 Preprocessing

The workflow of preprocessing is illustrated in Fig. 2, which includes surface generation, properties calculation, and feature point and region selection.

Surface generation. For a given protein complex, we use the MSMS program [19] to generate a triangulated mesh for each of its interacting subunits and set the density and probe radius to 1.0 point/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 cutoff from the surface of its binding partner. A triangle is selected when its three vertices are all retained in the interaction region. We represent the binding site surface as a connected and non-directed graph $G = (V, E)$ where each node $v \in V$ represents a surface point and E includes all and only the edges (v_i, v_j) such that point v_i and v_j are adjacent on the surface.

Properties calculation. Each node $v \in V$ is associated with a local value of three properties: shape index, *si(v)*, electrostatic potential, *esp(v)*, and hydrophobicity, *hyd(v)*. The shape index of *v*, *si(v)*, is defined using the maximum (k_1) and minimum (k_2) local curvature, which is given as follows [20]:

$$si(v) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{k_1(v) + k_2(v)}{k_1(v) - k_2(v)}\right). \quad (1)$$

It takes a value in the interval [0, 1] which is further divided into nine categories each corresponding to a well-known

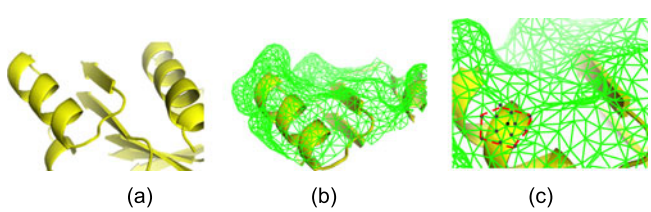


Fig. 2. Preprocessing step: (a) A protein complex is used to create a triangulated mesh (b) using the MSMS program and various properties are calculated. (c) Feature point is selected based on the distance to the binding site residue and feature region, which is growing around the feature point with radius $r = 4$Å.

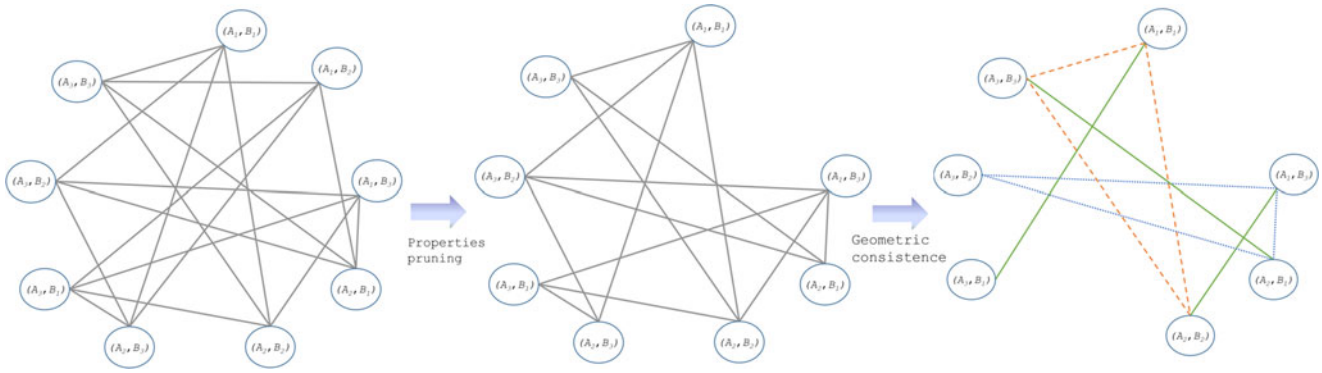


Fig. 3. Initial alignment step: A product graph is generated between feature points of two binding sites. After applying the properties pruning, vertices (A_1, B_2), whose region properties are incompatible, are removed from the graph. Geometric consistence is used to filter out invalid correspondences and maximum clique is detected. Three cliques are demonstrated: one by a solid line, one by a dotted line, and another by a dashed line. Residue correspondences, or seed alignments, are generated from these cliques.

shape, such as dome and saddle [20]. The initial values of potential in v, *esp*(v), and *hyd*(v) are calculated using the VASCo software package [21]. These values are standardized and discretized into six categories by PBSalign.

Feature point and region selection: To find potential correspondences between two binding sites, we should specify some feature points on the surfaces. In PBSalign, a feature point corresponds to a binding site residue and is defined as the surface point which has the smallest Euclidean distance to the C_α atom of the corresponding residue. In the following sections, we will use the feature point and binding site residue interchangeably. For a feature point v_i, we grow a feature region R_i which is centered at the feature point and covers surface points with Euclidean distance < 4Å (see Fig. 2). The feature region can be represented using the following set:

$$R_i = \{c(R_i), n(R_i), s(R_i)\}, \tag{2}$$

where $c(R_i)$ represents the coordinates of feature point, $n(R_i)$ is the normal vector of feature point, and a region signature set $s(R_i)$ contains three signatures of the region R_i for the shape index, electrostatic potential, and hydrophobicity properties, respectively. Each signature is represented by a histogram and a region signature set has a collection of all histograms. The number of histogram bins is set to the discretization categories for the characteristics of the three signatures, which are empirically set to nine, six, and six for *si*, *esp*, and *hyd*, respectively. The above process is repeated for each binding site residue.

2.2 Initial Alignment

We use the feature regions to match the binding sites BS_A and BS_B by seeking a set of common feature regions on two binding sites and an alignment transformation that brings BS_B close to BS_A. For each feature region on BS_A, we first find all corresponding feature regions on BS_B and then extract subsets of consistent region correspondences. The initial correspondence set is then filtered by several pruning algorithms, to be discussed shortly, and the final set of matching regions is found using a maximal clique detection algorithm [22]. In this step, we keep all the consistent sets of matching feature regions, which are then used as seed alignments for further refinement by the next step, MaxSurf. An overview of initial alignment is illustrated in Fig. 3.

Product graph. Given the two binding sites, BS_A and BS_B, and their graph representations, $G_A = (V_A, E_A)$ and $G_B = (V_B, E_B)$, the product graph of G_A and G_B, $G_P = (V_P, E_P)$, is constructed by inserting every pair of feature points, $v_i \in V_A$ and $v_j \in V_B$, from two binding sites into its vertices set V_P. Edges are drawn between every two vertices if they do not share a common feature point.

Properties pruning. For each node of V_P, correspondence between two feature points is established through their compatibility of associated regions. Two feature regions are compatible if a certain similarity is observed in their region properties. Let $p = (A, B)$ be a potential region correspondence (i.e., a node in the product graph G_P) between the binding sites BS_A and BS_B. We assess the compatibility of p by comparing the region signatures of A and B using the following similarity score of two feature regions:

$$S_R(p) = \frac{<s(R_A), s(R_B)>}{||s(R_A)|| \cdot ||s(R_B)||}, \tag{3}$$

where $s(R_A)$ and $s(R_B)$ are the signatures for regions A and B, respectively. $<s(R_A), s(R_B)>$ is the inner product of $s(R_A)$ and $s(R_B)$, and $||s(R_A)||$ and $||s(R_B)||$ are the second-norms of $s(R_A)$ and $s(R_B)$, respectively. A region correspondence p is considered further only if $S_R(p) > \varepsilon_R (= 0.6)$ holds true; otherwise we discard it from the product group G_P. In PBSalign, similarity scores of shape index, electrostatic potential, and hydrophobicity properties are applied squentially to filter out non-compatible feature points or regions correspondences.

Geometric consistency. Because the initial set of correspondences after the properties pruning might be quite large, we further examine the geometric consistency and relational constraints of two connected nodes in G_P to remove outliers. These processes make the next step of maximal clique detection run faster, but do not affect the correctness of the PBSalign algorithm. We call a pair of region correspondences, e.g., $p_1 = (A_1, B_1)$ and $p_2 = (A_2, B_2)$, geometrically consistent (see Fig. 4) if the Euclidean distance between the feature points' position $\delta_A = ||\mathbf{c}(A_1) - \mathbf{c}(A_2)||$ and $\delta_B = ||\mathbf{c}(B_1) - \mathbf{c}(B_2)||$ are of similar length, i.e., $|\delta_A - \delta_B| < \varepsilon_d$ (= 2Å based on empirical observations), and the angles between corresponding pairs of vectors $\angle(\mathbf{n}(A_1), \mathbf{n}(A_2))$ and $\angle(\mathbf{n}(B_1), \mathbf{n}(B_2))$ do not differ more than a certain

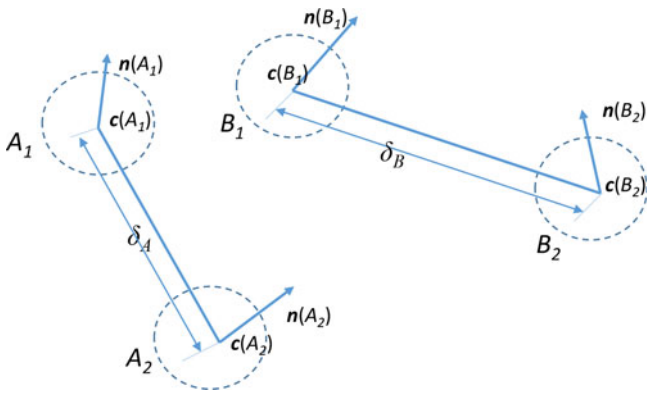


Fig. 4. Geometric consistency: A_1 (or A_2) and B_1 (or B_2) are feature regions from binding site A and B. ***c*** and ***n*** represents coordinate and normal vector of feature point.

threshold ε_α ($= 30$ degree based on empirical observations). In the product graph G_P, edges between p_1 and p_2 are removed if the geometric consistence does not hold.

Maximal clique detection. Having applied all the filters, the size of the potential correspondences is reduced so that it is possible to search for cliques in it. We use the algorithm of Bron and Kerbosch [22] to find all cliques with a minimum clique size of three. For each clique we generate a rigid body transformation based on all region correspondences, and we then calculate RMSD and the number of aligned surface feature points (N_p) of this transformation. All the cliques are sorted according to the ratio of RMDS/N_p in ascending order and up to N_{seed} ($= 20$) cliques are selected as initial alignments for further refinements. The intuition is that we select the cliques which have smaller RMSD and a larger number of aligned points. The binding site BS_B is then transformed and rotated towards BS_A using the results of the initial alignments.

2.3 MaxSurf

In this step, each seed alignment from the previous step is further refined by searching maximal surface overlapping through an iterative method, iterative closest point (ICP) [23]. A refinement process generates the final alignment.

Iterative closest point. The outputs of the initial alignments may not be the optimal transformation, but provide a coarse initialization for ICP to refine the rigid transformation. After applying ICP to the binding sites, the output is used to transform the binding site BS_B towards BS_A. As the ICP algorithm can only provide a local minimum of alignment error, the binding site pairs are sent to the next step for further refinement.

Refinement. In this step, we refine the matched binding site surfaces on the basis of the results recognized in the previous stage. For each seed alignment, we utilize the resulting product graph from the previous step. Second, we calculate the Euclidean distance between each residue on the binding site BS_A and its nearest neighbor from the binding site BS_B, then sort the distances in descending order and select 70th percentile of the observed distance as a cutoff to prune feature point correspondences (vertices from the graph). Secondary structure of a residue is optionally applied in this step, which is calculated using C_α coordinate of five neighbor residues [24]. Finally, we utilize maximum clique detection [25] to obtain a set of residue correspondences whose C_α atom distance is less than 4Å.

The output of maximum clique detection consists of a string of aligned residue pairs, which are sequentially ordered from their N to their C terminal. We further extend the alignments to include as many unaligned binding site residues as possible using a greedy approach. The fixed binding site, BS_A, is selected as the starting point. We define a fragment as a set of at least two adjacent residues in the binding site BS_A. For fragment F, let F_h denote the first residue and F_t denote the last residue of the fragment. The extension procedure for the fragment F is described as follows. First, we begin searching unaligned residues from F_h towards the N terminal. An unaligned residue is selected (if any), and the distance between its nearest unaligned residue from the BS_B is calculated. If the distance is < 4Å, the residue pairs are marked as aligned, and the next unaligned residue towards the N terminal is selected until an aligned residue or the beginning of binding sites is approached. Second, a similar searching process is started from F_t towards the C terminal. Finally, we use the q-Score [26] to rank each valid seed alignment and select the top one as the final output. The valid alignment of PBSalign means the residue correspondence is unique and sequential.

2.4 Time Complexity

For a clearer description, we define the number of feature points on the binding site as m. In the preprocessing step, the properties are calculated for m feature points, with a complexity of O(m). In the second step, we define the number of correspondences of each feature point as n for each feature point. Hence, we need to perform an n comparison of signatures for pruning, and the complexity is O(mn). We define g as the number of vertices of the product graph, and the time complexity for detecting maximal cliques is O($3^{g/3}$) [27]. In the third step, ICP is implemented using a k-d tree [28], and its time complexity is O($b_A log b_B$) where b_A and b_B are the numbers of vertices on the binding site surface BS_A and BS_B, respectively. The total time complexity of PBSalign is $O(mn) + O(3^{g/3}) + O(b_A log b_B)$.

3 Results

In this section, we compare the performance of PBSalign with current methods that provide publicly accessible software packages, including I2I-SiteEngine [12], Protein-protein interactions: methods for detection and analysis, iAlign [11], and Galinter [13]. Since the performance comparisons of Galinter/I2I-SiteEngine and I2I-SiteEngine/iAlign have been conducted in [11], [13] with iAlign reported as the best performing method, we selected iAlign to benchmarking the performance of PBSalign. The datasets consisted of homologous and non-homologous binding sites, as well as different types of protein complexes.

1) The first dataset, denoted as D_1, is composed of homologous PPIs, and was originally constructed to evaluate performance of iAlign [11].
2) The second dataset, denoted as D_2, is taken from Table 1 in [29], and was created to study structurally similar binding sites coming from different protein folds (i.e., non-homologous).

TABLE 1
Average (Standard Deviation) Values of SI, SAS, and MI with iAlign and PBSalign for Dataset D_1

Methods	Score		
	SI	*SAS*	*MI*
iALIGN	2.45 (0.88)	7.15 (3.25)	0.62 (0.09)
PBSALIGN	2.12 (0.76)	6.23 (3.04)	0.61 (0.10)

In the following sections, we used <PDB-ID><Chain ID of the binding site><Chain ID of the binding site partner> to represent a binding site.

Similar to the structural alignments, the performance comparison of binding sites alignments could be based on alignment quality or alignment accuracy. The alignment quality was evaluated using the geometric match measures, such as the RMSD of the superimposed structures. Alignment accuracy, a discrimination problem, is traditionally measured according to the accuracy with which an alignment method classifies a pair of protein structures into a similar/dissimilar category [30], [31] defined using the similarity of folds (e.g., SCOP [32]) or PPI (e.g., SCOPPI [33] and SCOWLP [34]). However, recent studies [29], [35] show that some proteins or binding sites from different folds may have very similar structures. Hence, we focus on the comparasion of different alignment methods based on the alignment quality.

When evaluated using the geometric match, the goal of structural alignment is to minimize the RMSD of the aligned region. However, as the RMSD depends on the number of aligned residues, the RMSD values associated with alignments of different lengths cannot be compared. To overcome this issue, we used the geometric match measures, which simultaneously consider various factors, such as similarity index (SI), match index (MI), and structural alignment score (SAS). These measures, which have also been used in a comprehensive evaluation of protein structure alignment method [36], are defined as follows:

$$SI = \frac{RMSD \times \min(L_A, L_B)}{N_e}, \tag{4}$$

$$SAS = \frac{RMSD \times 100}{N_e}, \tag{5}$$

$$MI = 1 - \frac{1 + N_e}{\left(1 + \frac{RMSD}{\omega_0}\right)(1 + \min(L_A, L_B))}, \tag{6}$$

where N_e is the number of aligned residues, ω_0 is a normalizing factor and set to 1.5 [36], and L_A and L_B are the lengths of binding site BS_A and BS_B. The units of SI and SAS are Å. MI takes values between 0 and 1. For these measures, lower values correspond to better alignments.

3.1 Experiments with Dataset D1

The dataset D_1 consists of biologically related PPIs. Here, a PPI pair is said to be biologically related if it is from same SCOP superfamily and shares a certain level of similarity. Hence, PPI pairs from this dataset are homologous and are expected, in general, to share similar function. For details about selecting PPI pairs, see [11].

For a given PPI pair, $I_A = \{BS_A{}^1, BS_A{}^2\}$ and $I_B = \{BS_B{}^1, BS_B{}^2\}$, iAlign first considers two possible ways of alignment. One is to align $BS_A{}^1$ to $BS_B{}^1$ and $BS_A{}^2$ to $BS_B{}^2$, and the other is to align $BS_A{}^1$ to $BS_B{}^2$ and $BS_A{}^2$ to $BS_B{}^1$. Then, iAlign selects the one whose score is the best. The final output of iAlign consists of two lists of residue correspondences, each of which is related to a pair of binding sites from different subunits. In contrast to iAlign, PBSalign is designed specifically for binding site comparison, which can designate binding site pairs for comparison. Hence, the problem is how to select one binding site pair aligned by iAlign and compare its alignment with that of PBSalign. Our solution is to pick the binding site pair which has lower SAS score to form a testing dataset for PBSalign. During the experiments, the geometric match measures are calculated on the same binding site pairs.

A comparison of the SI, SAS, and MI values obtained by iAlign and PBSalign is summarized in Table 1. As can be seen, PBSalign results in structural alignments with better SI, SAS, and MI values in comparison to iAlign. The detailed distributions of SI, SAS and MI values are depicted in Fig. 5. We also tested the statistical significance of the observed difference in the mean values of SI, SAS, and MI. The difference in the mean values of SI, SAS, and MI was found to be statistically significant (p-value $\ll 0.001$) using both a paired t-test and Wilcoxon test.

We further analyzed the relative improvement in various geometric match measures obtained by PBSalign or iAlign. The measure difference is defined as (measure(PBSalign)—measure(iAlign)). For example, the measure difference of SI is given as $\mathrm{dSI} = (\mathrm{SI(PBSalign)} - \mathrm{SI(iAlign)})$. Similarly, we defined the difference for SAS (dSAS) and MI (dMI). The detailed distribution of dSI, dSAS, and dMI values are shown in Fig. 6.

For these measure differences, it is difficult to identify the exact value which shows a significant change in the alignment quality. Based on the experimental results, we have empirically considered a $|\mathrm{dSI}| > 0.5$ as a significant improvement imposed by iAlign (i.e., $\mathrm{dSI} > 0.5$) or PBSalign (i.e., $\mathrm{dSI} < 0.5$) and a dSI between -0.5 and 0.5 as not so significant. Similarly, we define $|\mathrm{dSAS}| > 1.5$ and $|\mathrm{dMI}| > 0.05$ as significant improvement for SAS and MI, respectively. Table 2 shows summary of alignment improvement based on SI, SAS, and MI values. PBSalign results in alignments with better SI (∼30%), SAS (∼26%), and MI (∼14%) in comparison to iAlign.

3.2 Experiments with Dataset D2

The dataset D_2 consists of 69 binding sites from 10 groups. Members of each group have similar binding sites on one side of their interfaces, but the partner proteins are different. Members of this structurally non-redundant set illustrate bindings of partners with different geometries, sizes and composition. Whereas one can expect similar functionality of proteins in D_1, D_2 provides similar architectures where different functions of the interaction are expected.

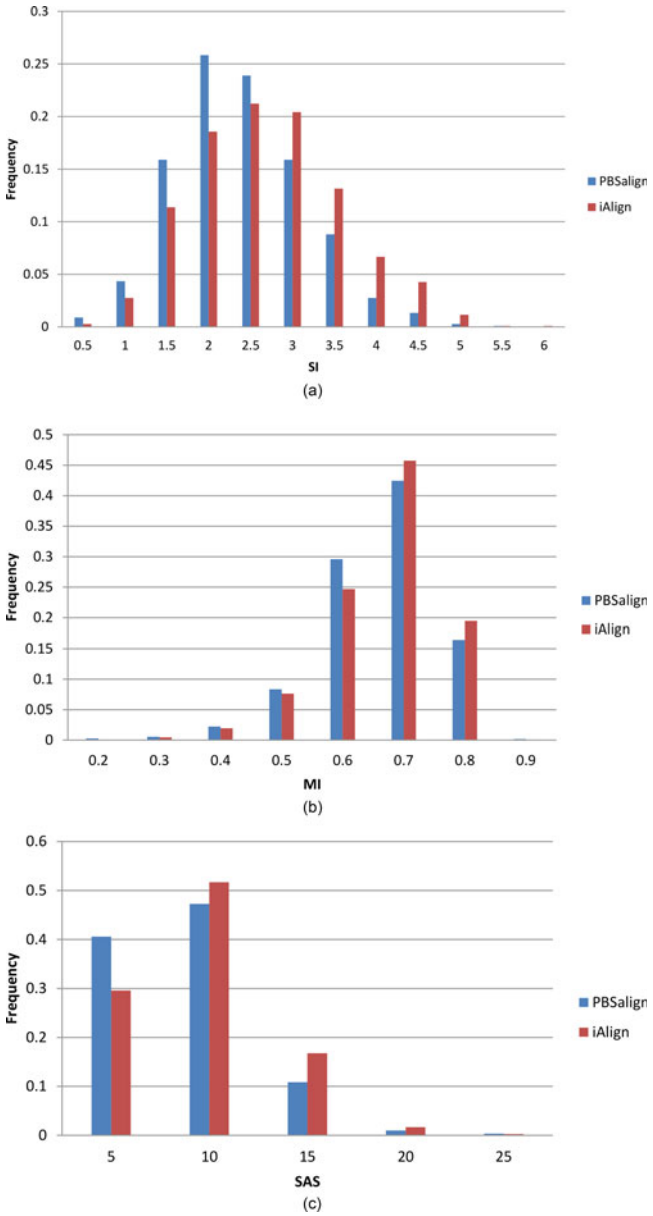


Fig. 5. Histograms of (a) SI, (b) SAS, and (c) MI for dataset D_1.

During the experiments, we performed pair-wise alignments among members from the same group. Hence, totally $k \times (k-1)/2$ times of alignments are needed for each group where k is the size of group.

Different from the dataset D_1, the dataset D_2 has designated similar binding site pairs. For some cases, iAlign cannot find the same matching binding site pairs as that of the datasets because of different definitions of similarity score and strategies to select paired binding sites. Hence, we simply excluded those pairs from the dataset D_2 according to the alignment results of iAlign. Finally, we obtained 54 binding site pairs, which are structurally similar but non-homologous.

We first calculated average and standard values of SI, SAS, and MI for the alignments generated by iAlign and PBSalign, which are shown in Table 3. From this table, we can see that PBSalign achieved better geometric match measures compared with iAlign. The detailed distributions of the SI, SAS and MI values are shown in Fig. 7.

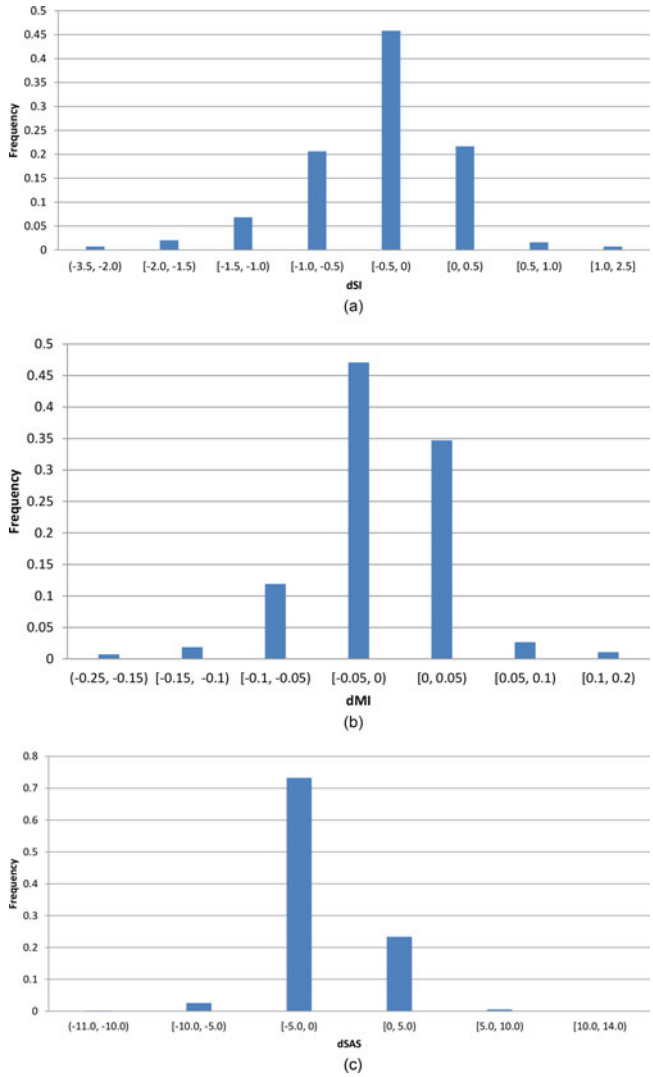


Fig. 6. Histograms of (a) dSI, (b) dSAS, and (c) dMI for dataset D_1.

Next, we tested the statistical significance of the observed difference in the mean values of SI, SAS, and MI, using a paired t-test and Wilcoxon test. The difference in the mean SI, SAS, and MI was found to be statistically significant (p-value $<< 0.001$).

We further analyzed the relative improvement in various geometric match measures by PBSalign. Similar to the dataset D_1, we defined $|\text{dSI}| > 0.5$, $|\text{dSAS}| > 1.5$, and $|\text{dMI}| > 0.05$ as a significant improvement for SI, SAS, and MI, respectively. Table 4 shows the summary of measure differences. From this table, we can see that PBSalign results in alignments with better SI (by ~56 percent), SAS (by ~76

TABLE 2
Comparison of iAlign with PBSalign for Dataset D_1 Using SI, SAS, and MI Measures

Measures	% Binding Site Pairs where	
	PBSalign is Better	iAlign is Better
SI	30	2
SAS	26	3
MI	14	4

TABLE 3
Average (Standard Deviation) Values of SI, SAS, and MI with iAlign and PBSalign for Dataset D_2

Methods	Score		
	SI	*SAS*	*MI*
iAlign	2.95 (1.39)	19.20 (9.89)	0.68 (0.10)
PBSalign	2.06 (1.07)	13.50 (8.46)	0.63 (0.10)

TABLE 4
Comparison of iAlign with PBSalign for Dataset D_2 Using SI, SAS, and MI Measures

Measures	% Binding Site Pairs where	
	PBSalign is Better	IAlign is Better
SI	56	15
SAS	76	6
MI	46	9

percent), and MI (by ~46 percent) in comparison to iAlign. The detailed distribution of dSI, dSAS, and dMI values are shown in Fig. 8.

4 Case Study

One important finding based on the dataset D_2 is that one or more binding site residues and their positions are conserved across proteins that have similar binding-site motifs, irrespective of global similarity or the similarity of binding partners. An implication of this is that local conserved residues between proteins with similar binding sites can be closely superimposed onto each other [29]. Based on this finding, it is informative to provide case studies that can visually demonstrate the capability of PBSalign as reported in Table 4 by comparing the degree to which PBSalign and iAlign can align proteins with similar binding sites in a previous study. Fig. 5 of [29] desmonstrates three protein complexes, 1f95_BA, 1otf_EA, and 1d5w_CB, which contain similar binding sites and are shown to have a close superimposition on their evolutionarily conserved local residues. Each of these protein complexes is characteristic of D_2 in that each shares similar binding-site motifs, contains a single-chain interface, and elicits multi-functions with different binding partners. PDB access no. 1f95 refers

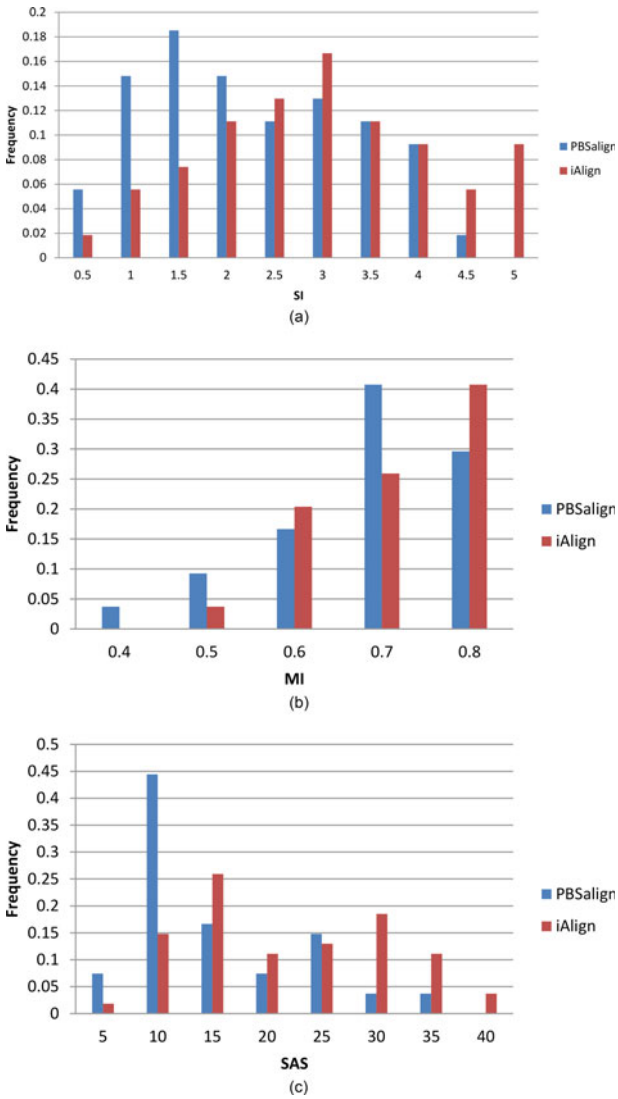


Fig. 7. Histogram of (a) SI, (b) SAS, and (c) MI for dataset D_2.

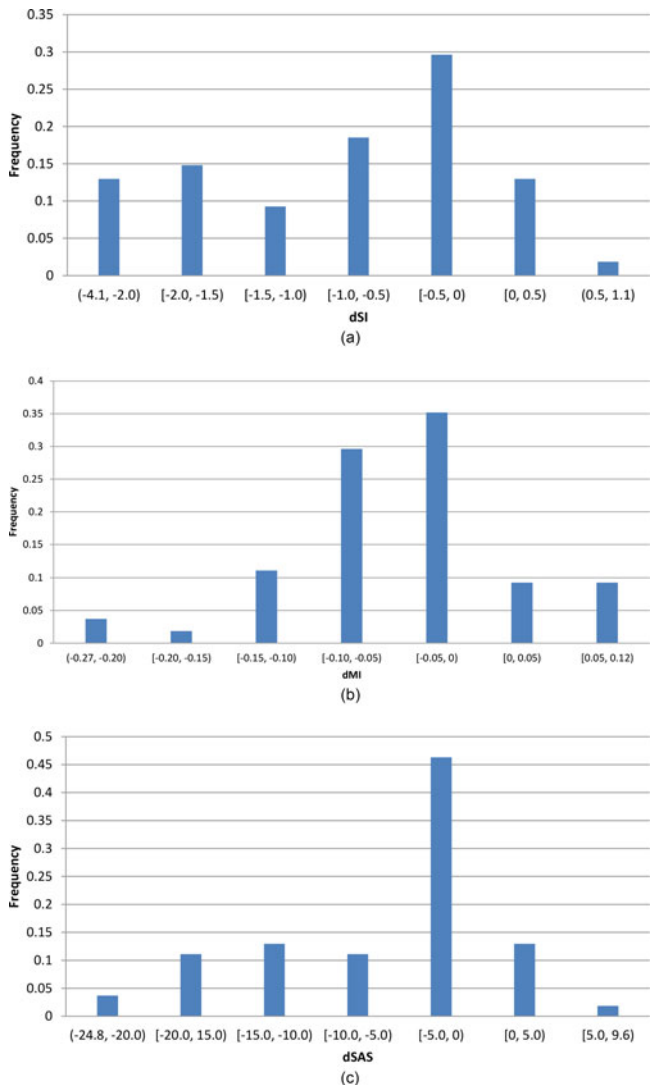


Fig. 8. Histogram of (a) dSI, (b) dSAS, and (c) dMI for dataset D_2.

Fig. 9. Alignment results of PBSalign: (a) 1f95_BA versus 1otf_EA, (b) 1f95_BA versus 1d5w_CB, and (c) 1otf_EA versus 1d5w_CB.

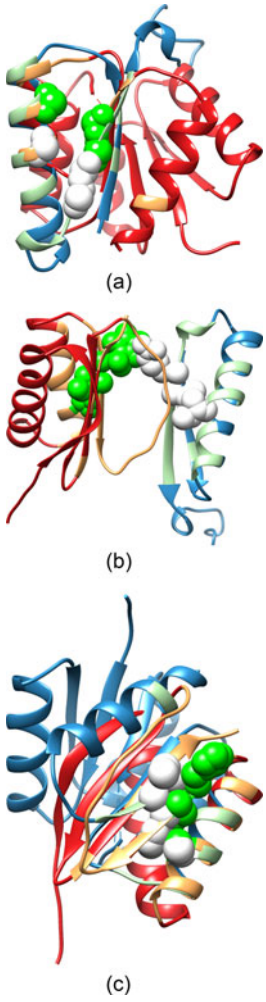


Fig. 10. Alignment results of iAlign for (a) 1otf_ EA versus 1d5w_CB, (b) 1f95_BA versus 1otf_EA, and (c) 1f95_BA versus 1d5w_CB.

to Dynein Light Chain 8 (DLC8) and BIM Peptide Complex which translate molecular cargoes along microtubules [37]; 1otf refers to 4-Oxalocrotonate Tautomerase, which catalyzes the isomerization of unsatured ketones [38]; and 1d5w refers to Phosphorylated FIXJ Receiver Domain, which alters a response regulator's conformation in a variety of adaptive processes [39]. Coincidentally, each of these complexes has three evolutionary conserved residues [29]. Specifically, residues Ala39, His55, and Cys56 are conserved on 1f95_BA; Ile21, Arg40, and Val41 are conserved on 1otf_EA; and Ala90, Glu100, and Phe101 are conserved on 1d5w_CB. Three alignments are performed to compare PBSalign and iAlign's ability to closely superimpose similar binding sites: 1f95_BA versus 1otf_EA, 1f95_BA versus 1d5w_CB, 1otf_EA versus 1d5w_CB. The alignment results of PBSalign are illustrated in Fig. 9, where the interface chains of these complexes are shown in red and blue and their binding site residues are highlighted in orange and light green. The evolutionary conserved residues are shown in spheres. From Fig. 9, we can see that PBSalign can align the conserved local residues closely for the pairwised alignments of the three binding sites. For example, in Fig. 9c, binding site residues Ile21, Arg40, and Val41, of 1otf_EA are aligned to binding site residues Ala90, Glu100, and Phe101 of 1d5w_CB, respectively. During the experiments, iAlign generates the same matching binding site pairs as the dataset D_2 only for the pair 1otf_EA versus 1d5w_CB, which is shown in Fig. 10a. For the other comparisons, namely 1f95_BA versus 1otf_EA (Fig. 10b) and 1f95_BA versus 1d5w_CB (Fig. 10c), iAlign provides conformations that result in larger SI, SAS, and MI than PBSalign.

5 Discussion

We have presented PBSalign, a new method for explicitly comparing binding sites based on the geometric and physicochemical properties of local surfaces. PBSalign uses features extracted from the binding site surface to generate initial seed alignments, which are further refined to produce accurate alignment. Our experimental results demonstrate that PBSalign can capture similarities of homologous and non-homologous protein binding sites accurately and provide alignments with better geometric match measures as compared to iAlign for a larger part of alignments in the selected datasets. The alignment of PBSalign utilizes both the surface and structure information of binding site surfaces. This method is similar to I2I-SiteEngine, but different from iAlign, which mainly relies on structure.

PBSalign is an alignment-based method for comparing binding site. As we discussed in [40], another type is feature-based binding site comparison which utilize features (e.g. shape descriptors) to provide a fast comparison of binding site without explicit alignment of residues. We have recently developed such a type of algorithm, PBSword, which compares a pair of given binding sites by measuring similarities in their overall shapes [40]. However, it does not output results of residue correspondences. PBSalign overcomes the problem of obtaining residue correspondences, which is essential to judge the quality of alignment.

While PBSalign demonstrates high alignment quality, it requires much longer execution time than iAlign. In addition to providing PBSalign's time complexity in Section 2.4, we also measure the average running time to evaluate the efficiency of PBSalign. The experiments are conducted on a Linux Fedora server with AMD Opteron dual-core 1000 series processors and 8 GB RAM. With PBSalign, each alignment takes 28.3 s. For iAlign, each alignment takes only 0.25 s.

We emphasize that PBSalign is not a replacement for, but rather a complement to the existing methods by taking more properties into account. Data mining on larger datasets woud benefit from preprocessed or heuristic approaches that limit datasets for more computationally expensive steps like PBSalign. Preprocessing, which consumes over half of the CPU utilization of PBSalign, is one of the reasons why PBSalign takes significantly more time than methods such as iAlign. Furthermore, PBSalign finds residue correspondences based on the maximal clique detection algorithm, which is known to be NP-hard [41] theoretically and computationally expensive even with heuristic approaches.

Future works includes the integration of PBSword and PBSalign to achieve fast search of large-scale databases by filtering out geometrically dissimilar binding sites using PBSword [42], and then providing accurate structural alignments for the remaining sites using PBSalign. To accelerate the execution time, we plan to parallelize PBSalign on Graphic Processor Unit (GPU) by dispatching each refinement of seed alignment to a computing core of GPU [43] and utilize distributed computing environments, such as MapReduced [44].

Acknowledgments

The authors would like to thank Mu Gao and Jeffrey Skolnick for providing iAlign tool and their dataset. Chi-Ren Shyu was supported by US NSF DBI-1053024 for shape analysis, Dmitry Korkin was supported by US NSF DBI-084519 and IOS-1126992, Bin Pang was supported by the Shumaker Endowment for Bioinformatics, Nan Zhao was supported by US NSF IOS-1126992, and David Schlessman was supported by the Stamps Foundation.

References

[1] R. P. Bahadur and M. Zacharias, "The interface of protein-protein complexes: Analysis of contacts and prediction of interactions," *Cell. Mol. Life Sci.*, vol. 65, no. 7–8, p. 1059, Apr. 2008.

[2] A. W. Ghoorah, M.-D. Devignes, M. Smail-Tabbone, and D. W. Ritchie, "Spatial clustering of protein binding sites for template based protein docking," *Bioinformatics*, vol. 27, no. 20, pp. 2820–2827, Oct. 2011.

[3] S. Mukherjee and Y. Zhang, "Protein-protein complex structure predictions by multimeric threading and template recombination," *Structure*, vol. 19, no. 7, pp. 955–966, 2011.

[4] X. Kuang, J. G. Han, N. Zhao, B. Pang, C. R. Shyu, and D. Korkin, "DOMMINO: A database of macromolecular interactions," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D501–D506, Jan. 2012.

[5] G. Kuzu, A. Gursoy, R. Nussinov, and O. Keskin, "Exploiting conformational ensembles in modeling protein–protein interactions on the proteome scale," *J. Proteome Res.*, vol. 12, no. 6, pp. 2641–2653, 2013.

[6] R. M. Bhaskara, G. A. de Brevern, and N. Srinivasan, "Understanding the role of domain–domain linkers in the spatial orientation of domains in multi-domain proteins," *J. Biomol. Struct. Dyn.*, pp. 1467–1480, Dec. 2013.

[7] M. Gao and J. Skolnick, "APoc: Large-scale identification of similar protein pockets," *Bioinform. Oxf. Engl.*, vol. 29, no. 5, pp. 597–604, Mar. 2013.

[8] S. Wang and W.-M. Zheng, "CLePAPS: Fast pair alignment of protein structures based on conformational letters," *J. Bioinform. Comput. Biol.*, vol. 6, no. 2, pp. 347–366, 2008.

[9] Y. Y. Tseng and W.-H. Li, "Classification of protein functional surfaces using structural characteristics," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 4, pp. 1170–1175, Jan. 2012.

[10] H. Hasegawa and L. Holm, "Advances and pitfalls of protein structural alignment," *Curr. Opin. Struct. Biol.*, vol. 19, no. 3, pp. 341–348, 2009.

[11] M. Gao and J. Skolnick, "iAlign: A method for the structural comparison of protein–protein interfaces," *Bioinformatics*, vol. 26, no. 18, pp. 2259–2265, 2010.

[12] A. Shulman-Peleg, S. Mintz, R. Nussinov, and H. J. Wolfson, "Protein-protein interfaces: Recognition of similar spatial and chemical organizations," in *Algorithms in Bioinformatics*. New York, NY, USA: Springer, 2004, pp. 194–205.

[13] H. Zhu, I. Sommer, T. Lengauer, and F. S. Domingues, "Alignment of non-covalent interactions at protein-protein interfaces," *PLoS ONE*, vol. 3, no. 4, p. e1926, Apr. 2008.

[14] T. Fober, G. Glinca, G. Klebe, and E. Hullermeier, "Superposition and alignment of labeled point clouds," *IEEE ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 6, pp. 1653–1666, Nov./Dec. 2011.

[15] J. Konc and D. Janezic, "ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment," *Bioinformatics*, vol. 26, no. 9, pp. 1160–1168, 2010.

[16] V. Pulim, B. Berger, and J. Bienkowska, "Optimal contact map alignment of protein–protein interfaces," *Bioinformatics*, vol. 24, no. 20, pp. 2324–2328, 2008.

[17] P. Bertolazzi, C. Guerra, and G. Liuzzi, "A global optimization algorithm for protein surface alignment," *BMC Bioinformatics*, vol. 11, no. 1, p. 488, Sep. 2010.

[18] T. A. Binkowski and A. Joachimiak, "Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites," *BMC Struct. Biol.*, vol. 8, no. 1, p. 45, Oct. 2008.

[19] M. F. Sanner, A. J. Olson, and J.-C. Spehner, "Reduced surface: An efficient way to compute molecular surfaces," *Biopolymers*, vol. 38, no. 3, pp. 305–320, 1996.

[20] C. Dorai and A. K. Jain, "COSMOS-A representation scheme for 3D free-form objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 10, pp. 1115–1130, Oct. 1997.

[21] G. Steinkellner, R. Rader, G. G. Thallinger, C. Kratky, and K. Gruber, "VASCo: Computation and visualization of annotated protein surface contacts," *BMC Bioinformatics*, vol. 10, no. 1, p. 32, 2009.

[22] C. Bron and J. Kerbosch, "Algorithm 457: Finding all cliques of an undirected graph," *Commun. ACM*, vol. 16, no. 9, pp. 575–577, 1973.

[23] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[24] Y. Zhang and J. Skolnick, "TM-align: A protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res.*, vol. 33, no. 7, pp. 2302–2309, 2005.

[25] J. Konc and D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," *Proteins*, vol. 4, p. 5, 2007.

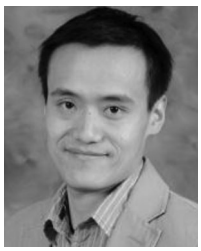
[26] E. Krissinel and K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 60, no. 12, pp. 2256–2268, 2004.

[27] E. Tomita, A. Tanaka, and H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," *Theor. Comput. Sci.*, vol. 363, no. 1, pp. 28–42, 2006.
[28] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.
[29] O. Keskin and R. Nussinov, "Similar binding sites and different partners: Implications to shared proteins in cellular pathways," *Structure*, vol. 15, no. 3, pp. 341–354, 2007.
[30] M. Gerstein and M. Levitt, "Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins," *Protein Sci.*, vol. 7, no. 2, pp. 445–456, 1998.
[31] M. Menke, B. Berger, and L. Cowen, "Matt: Local flexibility aids protein multiple structure alignment," *PLoS Comput. Biol.*, vol. 4, no. 1, p. e10, 2008.
[32] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, 1995.
[33] C. Winter, A. Henschel, W. K. Kim, and M. Schroeder, "SCOPPI: A structural classification of protein–protein interfaces," *Nucleic Acids Res.*, vol. 34, pp. D310–D314, Jan. 2006.
[34] J. Teyra, S. A. Samsonov, S. Schreiber, and M. T. Pisabarro, "SCOWLP update: 3D classification of protein-protein, -peptide, -saccharide and -nucleic acid interactions, and structure-based binding inferences across folds," *BMC Bioinformatics*, vol. 12, p. 398, 2011.
[35] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan, "Fast screening of protein surfaces using geometric invariant fingerprints," *Proc. Nat. Acad. Sci.*, vol. 106, no. 39, pp. 16622–16626, 2009.
[36] R. Kolodny, P. Koehl, and M. Levitt, "Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures," *J. Mol. Biol.*, vol. 346, no. 4, pp. 1173–1188, Mar. 2005.
[37] J. Fan, Q. Zhang, H. Tochio, M. Li, and M. Zhang, "Structural basis of diverse sequence-dependent target recognition by the 8 kDa dynein light chain," *J.Mol.Biol.*, vol. 306, pp. 97–108, Jul. 2000.
[38] H. S. Subramanya, D. I. Roper, Z. Dauter, E. J. Dodson, G. J. Davies, K. S. Wilson, and D. B. Wigley, "Enzymatic ketonization of 2-hydroxymuconate: Specificity and mechanism investigated by the crystal structures of two isomerases," *Biochemistry (Mosc.)*, vol. 35, pp. 792–802, Nov. 1995.
[39] C. Birck, L. Mourey, P. Gouet, B. Fabry, J. Schumacher, P. Rousseau, D. Kahn, and J. P. Samama, "Conformational changes induced by phosphorylation of the FixJ receiver domain," *Struct. FoldDes*, vol. 7, pp. 1505–1515, Oct. 1999.
[40] B. Pang, N. Zhao, D. Korkin, and C.-R. Shyu, "Fast protein binding site comparisons using visual words representation," *Bioinformatics*, vol. 28, no. 10, pp. 1345–1352, May 2012.
[41] M. R. Garey and D. S. Johnson, *Comput. Intractability*, vol. 174. New York, NY, USA: Freeman, 1979.
[42] B. Pang, X. Kuang, N. Zhao, D. Korkin, and C.-R. Shyu, "PBSword: A web server for searching similar protein-protein binding sites," *Nucleic Acids Res.*, vol. 40, no. W1, pp. W428–W434, Jun. 2012.
[43] B. Pang, N. Zhao, M. Becchi, D. Korkin, and C.-R. Shyu, "Accelerating large-scale protein structure alignments with graphics processing units," *BMC Res. Notes*, vol. 5, no. 1, p. 116, Feb. 2012.
[44] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, pp. 107–113, 2008.

Bin Pang received the PhD degree in bioinformatics from the Informatics Institute, University of Missouri, Columbia, MO, and another PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests include data mining, machine learning, computer network, and high-throughput parallel computing with GPU. After graduation from Chinese Academy of Sciences, he was a research assistant and a member of technical staff at NEC Labs China and Lucent Technologies China. He is currently with Microsoft, Redmond, WA.

David Schlessman received the dual BS degrees in computer science and biology from the University of Missouri, Columbia, MO, in 2014. His current research interests include survey-based studies in public health, wet-lab work in cell biology, and software engineering in bioinformatics and mobile application development. He was an undergraduate research assistant at the Center for Data Analytics and Discovery and Informatics Institute to conduct structural bioinformatics research.

Xingyan Kuang received the MS degrees in software engineering from Sichuan University, Chengdu, China. She is currently working toward the PhD degree in Informatics Institute, University of Missouri, Columbia, MO. Her current research interests include macromolecular (protein-protein, protein-nucleic acids and nucleic acids-nucleic acids) interaction database and binding sites prediction, macromolecular interaction network, and system biology. She received the Second Place Student Poster Award-Computational Merit, MidSouth Computational Biology and Bioinformatics Society (MCBIOS) in 2012.



Nan Zhao (S'13-M'14) received the PhD degree in informatics from University of Missouri, Columbia, MO. After one year of postdoctoral training at Missouri, he became a faculty member at the Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, in October 2013. He is currently an assistant research professor of computational biology. His research interests have been focused on structural bioinformatics, computational system biology, deep machine learning and big data mining for infectious diseases. He has received various academic awards, including Outstanding Graduate Student of School of Engineering, Shumaker Fellowship of Bioinformatics, and Recognized Research Award at Missouri Life Sciences Week, during his PhD training. He is a member of the IEEE, the International Society for Computational Biology (ISCB), and the Midsouth Computational Biology and Bioinformatics Society (MCBIOS).

Daniel Shyu is currently an undergraduate student in the Weldon School of Biomedical Engineering at Purdue University, West Lafayette, IN. He has worked in a variety of research labs, ranging from bioinformatics to veterinary medicine at the University of Missouri-Columbia, Mississippi State University, Mississippi State, MS, and Purdue University. Recently, he conducted several experiments related to the molecular study of influenza A virus proteins and their interactions. He is the recipient of the Purdue Trustees Scholarship, Stamps Leadership Scholarship and has utilized this opportunity to travel and serve abroad in Ecuador. His current research interests include a variety of disciplines, all connecting to medicine and the advancement of the field.

Dmitry Korkin received the PhD degrees in computer science from the University of New Brunswick, NB, Canada. Upon completing his postdoctoral training at the University of California, San Francisco, he joined the Department of Computer Science, University of Missouri (MU), Columbia, MO, in September 2007, where he became an associate professor and Core Faculty of the interdisciplinary doctoral program at the MU Informatics Institute, and served as its first associate director of Graduate Studies in Bioinformatics from 2008 to 2010. In September 2014, he joined Worcester Polytechnic Institute, Worcester, MA, as an associate professor in the Department of Computer Science, affiliated with the Departments of Applied Math and Biology and Biotechnology. Dr. Korkin is the recipient of the US National Science Foundation Faculty Early Career Development (NFS CAREER) Award and MU College of Engineering Junior Faculty Research Award. He is a member of the IEEE, the American Association for the Advancement of Science (AAAS) and the International Society for Computational Biology (ISCB).

Chi-Ren Shyu (S'89-M'99-SM'07) received the PhD degrees in electrical and computer engineering from Purdue University, West Lafayette, IN. Upon completing one year of postdoctoral training with Purdue University, he joined the Department Computer Engineering and Computer Science, University of Missouri (MU), Columbia, MO, in October 2000. He is currently the Shumaker Endowed professor and the chairman of the Department of Electrical and Computer Engineering. He also heads the MU Informatics Institute, where 43 core faculty members from 17 departments from MU support an interdisciplinary training and research program in bioinformatics, health informatics, and geoinformatics. His current research interests include biomedical informatics, big data analytics, and visual knowledge reasoning. He received the US National Science Foundation Faculty Early Career Development (NFS CAREER) Award, MU College of Engineering Faculty Research Award, and various teaching awards. He is a senior member of the IEEE, and a member of the American Association for the Advancement of Science (AAAS) and the American Medical Informatics Association (AMIA).