

Protein-Protein Interfaces: Recognition of Similar Spatial and Chemical Organizations

Alexandra Shulman-Peleg^{1,*}, Shira Mintz², Ruth Nussinov^{2,3,†}, Haim J. Wolfson¹

¹ School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel,
Telefax : +972-3-640 9373, e-mail : {shulmana,wolfson}@tau.ac.il;

² Sackler Inst. of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University.

³ Basic Research Program, SAIC, NCI-Frederick, Inc. Lab. of Experimental and Computational Biology
Bldg 469, Rm 151, Frederick, MD 21702, USA

Keywords: Protein-protein interfaces, classification, binding, 3-D database searches

1 Abstract

Protein-protein interfaces, which are regions of interaction between two protein molecules, contain the information regarding patterns of interacting functional groups. Recognition of such patterns is useful both for prediction of binding partners and for the development of drugs that can interfere with the formation of the protein-protein complex. Here we present a novel method, Interface-to-Interface (I2I)-SiteEngine, for structural alignment between two protein-protein interfaces. The method simultaneously aligns between two pairs of binding sites that constitute an interface. The method is based on recognition of similarity of physico-chemical properties and shapes. It assumes no similarity of sequences or folds of the proteins that comprise the interfaces. The similarities between interfaces recognized by I2I-SiteEngine provide an insight into the interactions that are essential for the formation of the complex and can be related to its function. Its high efficiency makes it suitable for large scale database searches and classifications. Here, first we utilize the method to create a classification of a *pilot* dataset of interfaces. Then we apply it to efficiently search the obtained clusters for recognition of similarities and for the prediction of binding.

2 Introduction

Most of the cellular processes are governed by association and dissociation of protein molecules. The understanding of such processes can throw light on the mechanism of molecular recognition. A protein-protein interface is defined by a pair of regions of two interacting protein molecules that are linked by non-covalent bonds. Analysis and classification of protein-protein interfaces [1, 2] is the first step in deciphering the driving forces stabilizing the molecular interactions. Recognition of certain interface binding organizations shared by different protein families, may suggest their important contribution to the formation and stability of the protein-protein complex. This may constitute targets for drug discovery and assist in predicting side effects. Furthermore, similar interfaces may suggest not only similar binding organizations, but also similarity in binding partners and function. This may provide hints for potential drug leads that will mimic these partners.

Previous classification of interfaces made by Keskin et al. [3] used a backbone representation of the interacting proteins to cluster all known structures of protein complexes. The

*To whom correspondence should be addressed, email: shulmana@tau.ac.il.

†The publisher or recipient acknowledges right of the U.S. Government to retain a nonexclusive, royalty-free license in and to any copyright covering the article. Funded in part by the NCI under contract NO1-CO-12400.

geometric hashing procedure used in their method considered only the geometric constraints between the interface C_α atoms. However, side chains play an important role in interaction between two molecules. Current methods that do consider side chain atoms, align between single binding sites and do not consider their interacting partner [4–8].

In this paper, we present a novel method, Interface-to-Interface (I2I)-SiteEngine, to recognize similarities between protein-protein interfaces independent of the sequence or the fold of the proteins that comprise them. In addition to geometric considerations used in previous alignment methods, this method takes into account biological considerations in the form of physico-chemical properties of the interacting atoms (both backbone and side-chain). The novelty of I2I-SiteEngine is in recognition of patterns of interacting functional groups shared by a pair of interfaces. Extending the algorithmic approach of our previous method for comparison of small molecules binding sites, SiteEngine [8], the current method performs a simultaneous alignment of **two** binding sites that constitute an interface. Such simultaneous alignment not only improves the performance of the algorithm, but also provides more significant biological results. The method introduces a hierarchical scoring scheme which is similar to SiteEngine, but is applied simultaneously to both sides of the interface. First, using a low-resolution representation by chemically important surface points, it performs efficient scoring and filtering of all possible solutions, while retaining the correct ones. Then, as the number of potential solutions is reduced to a smaller subset, the resolution of the molecular representation is increased, leading to more precise calculations. These compare the similarity of the surfaces as well as of local shapes of the chemically similar regions.

We apply the method on a *pilot* dataset to define clusters that contain similar interfaces. Some of the clusters include similar interfaces comprised by proteins with different structural folds. We analyze these clusters and show biological applications which emphasize the importance of such classifications and of procedures to search them.

3 Method

We define an interface as an unordered pair of interacting binding sites, that belong to two different, non covalently linked, protein molecules. Two interfaces are considered to be similar, if the binding sites that comprise them share similar physico-chemical properties and shapes. Given two interfaces $I=(A, B)$ and $I'=(A', B')$ the goal is to find an alignment that will maximize the similarity between them. This implies solving two problems: First, the correspondence between the binding sites of the two molecules is *a-priori* unknown and there are two possible ways of alignment. One is to align A to A' and B to B' and the other is to align A to B' and B to A' . Both of these possibilities must be considered. Assume that A is aligned to A' and B is aligned to B' . Then, the second and most complicated task is to simultaneously align A to A' and B to B' . When aligning between each pair of these binding sites, our goal is to maximize the superimposition of their corresponding shapes and physico-chemical properties, according to the representation below.

Efficient, biologically significant, representation of each binding site is crucial for the recognition of functional similarities between unrelated proteins. As depicted in Figure 1, in our method, each binding site is represented by the surface of its binding region and by the set of its important functional groups. The interacting surface is defined by a set of its solvent accessible surface points [9] that are located less than 4\AA from the surface of the other protein. Following the definition of Schmitt et al [5], each amino acid of a protein is represented as a set of its important functional groups, localized by pseudocenters, according to the interactions in which it may participate. The extracted pseudocenters may have one of the following properties: *hydrogen-bond donor*, *hydrogen-bond acceptor*, *mixed donor/acceptor*, *hydrophobic aliphatic and aromatic(pi) contacts*. We retain only those pseudocenters that represent at least

one atom that is exposed to the surface of the interface. Each surface point is then assigned a physico-chemical property according to the functional group to which it belongs. Some of these points may belong to several atoms that are represented by different pseudocenters. In such a case they are assigned the property of the nearest atom center. Surface points that are represented by the same pseudocenter constitute a physico-chemical surface patch. A *patch center* is the surface point nearest to the center of gravity of the patch. The average curvature of the surface patch is estimated by the solid angle shape function [10, 11] computed at the *patch center*. This parameter of shape is assigned to the corresponding original pseudocenter and is used throughout the algorithm for fast shape comparisons. In addition, a set of all the *patch centers* of the interface is used later as a low resolution representation of the surfaces.

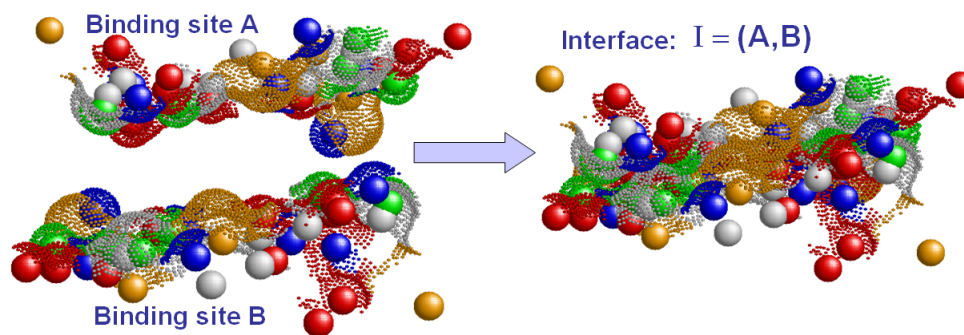


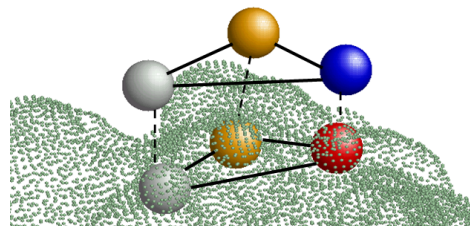
Figure 1: Interface Representation Representation of an interface as a pair of interacting binding sites. Pseudocenters are represented as balls and are colored as following: Hydrogen bond donors - blue, acceptors - red, donors/acceptors - green, hydrophobic aliphatic - orange and aromatic - white. The surface patches are represented as dots and are colored the same as the functional groups that they belong to.

The Matching Algorithm At this stage we compute all of the candidate transformations that can superimpose one interface onto the other. The matching is based on hashing of almost congruent triangles defined by triplets of pseudocenters. When considering protein-protein interfaces we are supplied with valuable information regarding the functional groups of two binding sites that interact with each other. Utilizing this information increases the speed and the quality of the alignment. Therefore, at the matching stage, for each binding site we consider only those pseudocenters that have a complementary physico-chemical property (with which it can interact) at the other binding site. Assuming that at least three such pseudocenters must be present in each interface, we define each triplet of such pseudocenters as an I-triangle (see Box 1).

The flow of the matching algorithm is presented in Figure 2. Given two interfaces $I=(A, B)$ and $I'=(A', B')$ the first stage is to recognize the interacting triangles of each interface. This is achieved by a supplementary hashing procedure, that stores all triplets of pseudocenters from the binding sites B and B' . These hash tables are used to check each triplet of pseudocenters from the binding sites A and A' whether it can form three interaction thus creating an I-triangle. Specifically, each triplet of pseudocenters of A (A') is used to access the hash table of B (B') to check whether there are three centers in B (B') that have complementary properties at suitable spatial locations. If a triangle of A is recognized as an I-triangle it is stored in the main matching hash table, which we denote I2I-Hash. If a triangle of A' is recognized as an I-triangle it is used to access the I2I-Hash of A to look for a similar I-triangle (similar physico-chemical properties and similar side lengths¹). If found, the two triangles are used to define a candidate transformation (rotation and translation) that can superimpose one interface onto the other.

¹The side lengths of the triangles are similar up to a user defined threshold (3.0Å in this paper).

Box 1: Interacting triangle (I-triangle) is a triplet of functional groups (pseudocenters) of one molecule that is recognized to form three interactions with the other molecule. Specifically, let $I=(A, B)$ be an interface, a triplet of pseudocenters of A is called an interacting triangle if and only if B has three complementary pseudocenters at spatial locations that can allow the formation of interactions. The figure depicts a pair of interacting triangles. Pseudocenters are represented as balls and the surfaces of the two molecules as green dots. Functional groups that are above the surfaces belong to one molecule and those that are below the surface to the other. Dotted lines represent the interactions between the pseudocenters. Hydrogen bond donors (blue) are complementary to hydrogen bond acceptors (red). Hydrophobic aliphatic (orange) and aromatic (white) pseudocenters can interact only with similar features in the other molecule. We assume that triangles with such a restricted definition are more significant than others for the complex formation.



The keys to all the hash tables are the three parameters of the side lengths of a triangle and an additional index, that encodes the properties of the nodes. Each two bits of this index encode the physico-chemical property of one node. Thus 6 bits are sufficient to encode a triplet. Nodes that can function both as hydrogen donors and acceptors are encoded twice, once as a donor and once as an acceptor. When retrieving data from the hash tables of B and B' we impose additional distance constraints that will ensure that the matched nodes are located close enough to interact with each other. For the convenience of data retrieval, the query index of physico-chemical properties is converted to an index with complementary properties. In addition to having similar physico-chemical properties and side lengths, triangles of pseudocenters retrieved from the I2I-Hash are required to also have similar values of the solid angle shape function [10].

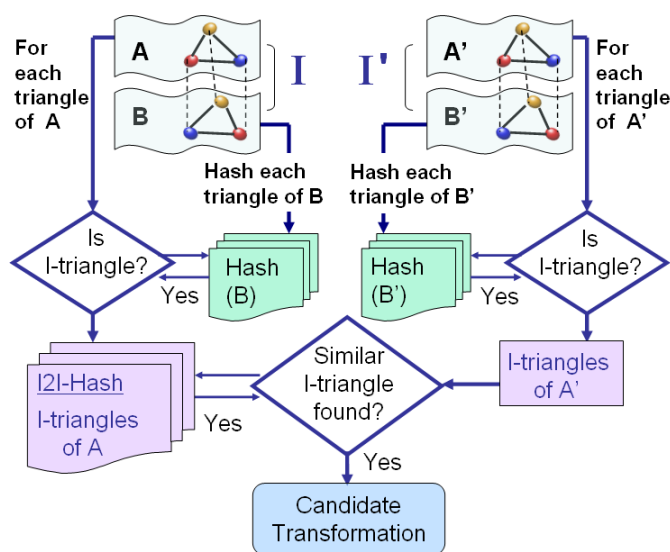


Figure 2: Overview of the matching algorithm

Scoring At this stage we evaluate the candidate transformations that can superimpose one interface ($I'=(A', B')$) onto another ($I=(A, B)$) and select the *one* that provides the best alignment of physico-chemical properties and of shape of the interfaces. A set of scoring functions is applied to simultaneously measure the similarity of a binding site A to A' , and of a binding site B to B' . Here we extend the scoring procedures presented in the SiteEngine method [8] to a simultaneous comparison between two pairs of binding sites.

Surface Scoring We use two scoring functions to compare shapes and physico-chemical properties of the surfaces superimposed by each candidate transformation. Both scoring schemes apply each candidate transformation to a certain set of surface points and compare the physico-chemical properties as well as the shapes of the corresponding regions of the two interfaces. All of the properties are efficiently stored in a chemically labeled distance transform grid [11], which allows their immediate retrieval. The first score, *Fast Low-Resolution Score*, applies this concept to a set of *patch centers*, which provide a low-resolution representation of the surfaces of the interface. Specifically, for each *patch center* of one interface we apply the candidate transformation, access the grid and consider the physico-chemical property and the shape of the region to which it is transformed in the other interface. The more similar these environment are, the higher the calculated score. The computation of this score is extremely fast and it efficiently estimates the potential similarity of the regions superimposed by a candidate transformation. The high ranking solutions are further clustered according to the RMSD between the pseudocenters under the candidate transformation.

After we significantly reduce the number of candidate solutions, the second score, the *Overall Surface Score*, performs a thorough comparison of the overall surfaces aligned by each transformation using a higher level of resolution of molecular representation. Surface points that are transformed close to the surface of the other protein are scored higher than those that fall at a greater distance. In addition, points that are transformed to regions with the same physico-chemical property as their own are scored higher than those with different properties.

1:1 Correspondence Score For each retained candidate transformation, we determine a one-to-one correspondence between the two sets of pseudocenters of the two interfaces. A set of pseudocenters of an interface $I=(A, B)$ is the union of the sets of pseudocenters of binding sites A and B that constitute it. The correspondence is obtained by calculating the maximum weight match in a weighted *bipartite graph* [12, 13], which represents the largest set of pairs of similar pseudocenters. The construction of a weighted *bipartite graph* for comparison between two protein-protein interfaces $I=(A, B)$ and $I'=(A', B')$ is performed in the following way: (1) Each pseudocenter from binding sites A, B, A' and B' defines a node. (2) Assuming that a candidate transformation aligns a binding site A to A' and a binding site B to B' , edges of a *bipartite graph* can only connect nodes of A to A' and nodes of B to B' . Following these restrictions, an edge is added between each pair of pseudocenters that have similar spatial locations and physico-chemical properties. (3) Each edge is assigned a weight that reflects the differences in distance and shape between the nodes. The maximum weight match [13] in this graph, provides a 1:1 correspondence between subsets of pseudocenters of the two interfaces. Due to restriction on the creation of the edges we obtain two separate 1:1 correspondences: one between subsets of pseudocenters of A and A' , and another between subsets of B and B' . The obtained 1:1 correspondence is used for two purposes: One is to improve each candidate transformation by the Least-Squares Fitting method [14]. The other is to score this transformation by summing the similarity scores between the matched pseudocenters and their corresponding surface regions.

Final Ranking Each candidate transformation is assigned a score which is the weighted sum of all the score functions described above. A candidate transformation with the highest score is selected. Since there are two ways of correspondence between the binding sites of the two interfaces, the correspondence that receives the highest score is selected as the correct one for that interface. Only *one solution* with the highest score is selected to represent a comparison between two interfaces. The rest of the solutions are ignored to allow efficient ranking of the results of different pairwise comparisons. When searching a dataset of interfaces for those that are similar to a specific interface of interest, the score of each comparison is normalized by the score of the query compared to itself. We denote this score as *Match Score* and it represents how much of the binding pattern of interest was found to match during the search. This can

suggest potential binding partners and predict their binding modes.

Complexity and Running Times The complexity of the algorithm is $O(N \cdot K \cdot m)$, where N is the maximal number of I-triangles of an interface, K is the maximal number of I-triangles retrieved in a hash table query, and m is the number of pseudocenters in the largest binding site among those that constitute the interfaces (selected within A, B, A', B'). Theoretically, when matching triplets of arbitrary nodes $N \cdot K$ is $O(m^6)$. However, in practice, since we consider only triplets of pseudocenters that define I-triangles, $N \cdot K$ is proportional to $O(m^2)$ (results not shown here). Therefore, because we score each candidate solution, the practical overall running times of the algorithm are proportional to $O(m^3)$. For a typical interface in our *pilot* dataset $m \sim 70$. According to statistics of all the hash tables gathered on all of the 4096 pairwise comparisons performed between 64 interfaces used in this study, the mean number of triangles retrieved in each access to any hash table is 30 and the mean maximal number is 160. Moreover, the mean overall CPU time of program execution for comparison between two interfaces is 28 seconds (3.0 GHz Xeon processors, 4GB memory). Sample running times of specific algorithm executions are provided in Table 1. Additional details regarding the implementation and default parameters can be found on our website: <http://pc-gamba.math.tau.ac.il/I2I-SiteEngine>.

4 Classification of Protein-Protein Interfaces

We applied the I2I-SiteEngine to classify a *pilot* dataset of some of the common protein-protein interfaces. As a result, the 64 interfaces were clustered into 22 different groups, which are detailed in the Appendix. Protein complexes with interfaces that belong to the same cluster are considered to share similar spatial and chemical organizations of their interacting regions.

Similarity Ranking and Clustering The first stage was to rank the dataset proteins according to their similarity to each other. To achieve this, I2I-SiteEngine was applied to all pairs of dataset interfaces (except for where surface area differs by more than 50%). Specifically, each time a certain interface is used as a query. The query is compared to all other dataset interfaces, ranking them in a decreasing order according to their similarity to this query. Table 1 provides three examples of such comparisons. The query interfaces in these examples were: (a) Clip bound to class II Mhc Hla-Dr3 (1a6a), (b) Trypsin(ogen) complexed with Pancreatic Trypsin inhibitor (1bzx) and (c) Glycine N-Methyltransferase (1d2h). The top ranking solutions are the interfaces that are similar to the query. As can be seen, many of them are comprised of proteins that share very low sequence similarity with the query. In addition, most of these proteins have totally different overall protein folds. Therefore, the similarity between them can not be recognized by sequence alignment methods or by structural alignment methods that align the overall backbones of the proteins. Yet, the similarity of these interfaces is successfully recognized by I2I-SiteEngine. At the next stage, a clustering algorithm uses the rank lists of all pairwise comparisons to cluster the interfaces to different groups. In this preliminary implementation we applied a greedy clustering procedure that performs several iterations with decreasing cut off values defined by the *Match Score*.

The Obtained Clusters Following the classification procedures, two types of interface clusters were obtained. The full clusters list is detailed in the Appendix. We will focus our attention on few examples of relatively large clusters that are divided to two types: (1) similar interfaces comprised by proteins with similar overall folds, (2) similar interfaces comprised by proteins with different overall folds. While type 1 is more straightforward, type 2 can suggest either similar binding organizations or similar functions shared by unrelated proteins.

Consider the rank lists of three clusters presented in bold font in Table 1. Table 1(a) and cluster 10 in the Appendix present an example of a type 1 cluster composed of MHC-antigen

| Rank | (a) Rank List of 1a6a | | | | (b) Rank List of 1bzx | | | | (c) Rank List of 1d2h | | | |
|------|-----------------------|-------------|--------------|-------------|-----------------------|-------------|--------------|-------------|-----------------------|-------------|--------------|-------------|
| | PDB | Match Score | Seq. Sim.(%) | Time (sec.) | PDB | Match Score | Seq. Sim.(%) | Time (sec.) | PDB | Match Score | Seq. Sim.(%) | Time (sec.) |
| 1 | 1a6aBC | 100 | 100/100 | 4 | 1bzxEI | 100 | 100/100 | 7 | 1d2hAB | 100 | 100/100 | 2 |
| 2 | 1aqdBC | 52 | 88/20 | 4 | 1tgsZI | 56 | 64/14 | 11 | 1axcCA | 38 | 14/14 | 4 |
| 3 | 1dlhBC | 44 | 87/20 | 4 | 1gl1AI | 51 | 41/17 | 8 | 1kbaBA | 34 | 5/5 | 2 |
| 4 | 1d9kDP | 38 | 60/6 | 5 | 1acbEI | 48 | 40/13 | 8 | 1cdtBA | 34 | 8/8 | 4 |
| 5 | 1jk8BC | 38 | 61/6 | 5 | 3tecEI | 37 | 13/13 | 8 | 1c1yBA | 34 | 6/11 | 3 |
| 6 | 1f3jBP | 35 | 60/6 | 5 | 1sbnEI | 36 | 15/13 | 10 | 1kkIAH | 33 | 12/9 | 3 |
| 7 | 1ydtEI | 27 | 13/5 | 6 | 1h28BE | 28 | 15/8 | 6 | 1czvAB | 32 | 12/12 | 1 |
| 8 | 1axcAC | 27 | 13/1 | 6 | 1cxzAB | 27 | 17/12 | 7 | 1b77AB | 32 | 15/15 | 2 |
| 9 | 1gl1AI | 26 | 15/5 | 8 | 1c1yAB | 27 | 18/10 | 7 | 1ao7AD | 30 | 15/11 | 1 |
| 10 | 1c1yBA | 25 | 12/1 | 4 | 1d9kDP | 27 | 15/3 | 6 | 3tecIE | 30 | 7/16 | 5 |

Table 1: Ten top ranking solutions obtained by I2I-SiteEngine when searching the *pilot* dataset against three interface queries (ranked 1): (a) Clip bound to class II Mhc Hla-Dr3 (1a6a), (b) Trypsin with inhibitor (1bzx), and (c) Glycine N-Methyltransferase (1d2h). For each comparison, column one presents the pdb codes followed by the corresponding chains of the top ranking interfaces. The order of the PDB chains is according to the correspondence defined by the method. Column two presents the score that indicates the extent of similarity. Column three presents the overall sequence similarity to the query chains. The order is the same order as the presentation of the chains, e.g. 1sbnEI - 15/13, means that sequence similarity of chains E of 1sbn and of 1bzx is 15%, and of chains I is 13%. Column four presents the algorithm running times for the selected correspondence (not including the preprocessing). In bold are the interfaces that later on were classified to belong the same cluster as the query.

interfaces. Although in this cluster the MHC molecules are bound to different peptides the similarity between their interfaces is successfully recognized and thus they are clustered together. A more interesting type 2 cluster is presented in Table 1(b) and in cluster 8 in the Appendix. This well-studied [6, 7] functional class of 'serine protease with inhibitor' is comprised of two different folds: the Trypsin-like serine proteases with inhibitor (ranked 1-4 in the Table) and the Subtilisin-like with inhibitor (ranked 5-6). In spite of the different overall folds of the chains, these proteins exhibit a similar function that is related to the interface properties. As expected, our method grouped together all interfaces of serine proteases with inhibitors bound to their catalytic region.

Another interesting example of a type 2 cluster can be seen in Table 1(c) and in cluster 5 in the Appendix. The cluster contains complexes that are comprised of a total of 7 different folds: (1) Methyltransferases (1d2h), (2) Snake toxin-like (1cdt, 1kba), (3) DNA clamp (1axc, 1b77), (4) PEP carboxykinase-like bound to Hpr-like (1kk1), (5) Galactose-binding domain-like (1czv), (6) P-loop containing nucleoside triphosphate hydrolases bound to Beta-Grasp (1c1y), (7) Defensin-like (1dfn, which was added to this cluster at the last iteration of the clustering algorithm and therefore is not presented in the Table). Figure 3 presents a superimposition of members of this cluster according to the transformation obtained by the alignment of corresponding interfaces. Figure 3(a) shows the alignment Coagulation factor V and Cardiotoxin V4II. Figure 3(b) shows the alignment between Glycine N-Methyltransferase and Proliferating Cell Nuclear Antigen. All of these proteins perform different functions, but are recognized to have similar interfaces. This may suggest that similar interface binding organizations may be shared by different protein families.

5 Additional Applications

Besides its usefulness for classification of interfaces, our method can be also applied to search databases of protein-protein interfaces. Searches of this type have two major biological applications: (1) Fast classification of newly determined complexes. (2) Prediction of binding partners

and binding modes by recognition of similarity to known complexes. In the first application, an interface of a 'newly' determined complex is extracted, compared to known interfaces and classified. This may provide valuable knowledge regarding the interface and its biological function. In the second, a complete protein structure is compared to a database of known interfaces to predict its potential binding sites.

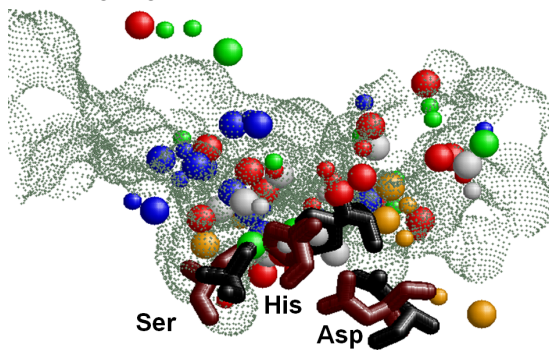
These database searches can be performed more efficiently with the help of a dataset of classified interfaces. Instead of comparing to each interface, we compare only to cluster *representatives*, thus gaining in efficiency and speed. The *representative* of each cluster is selected as the largest interface of that cluster. Below we provide examples of the two search procedures applied on the cluster representatives of our *pilot* dataset.

Interface Type Recognition In this section we describe classification of 'new' interfaces that were not used in the *pilot* dataset. A new interface (query) is compared to all the cluster representatives, except for those whose surface area differs by more than 50% from the query. These examples show the correctness of the obtained clusters and of their representation by representative interfaces.

Subtilisin Carlsberg with Eglin C When the interface of Subtilisin Carlsberg in complex with Eglin C (1cse) was compared to all of the cluster representatives, the interface formed by Trypsin complexed with Animal Kazal-type inhibitor (1tgs) received the highest rank. In spite of the fact that this representative interface belongs to a fold different from that of the query, the 'new' interface was correctly classified to belong to the cluster of serine proteases in complex with an inhibitor.

Box 2: Different Subtilisin and Trypsin folds: similar interfaces and similar functions

The functional class of "serine proteases with inhibitor" is one of the well studied [5–7] examples of proteins with different folds that share similar spatial and chemical binding organizations. Here, we illustrate the ability of our method to recognize these already known similarities using the example of Subtilisin Carlsberg with Eglin C (1cse, Subtilisin-like fold) and Trypsin with Animal Kazal-type inhibitor (1tgs, Trypsin-like fold). Functional groups of the interfaces are represented as balls (1cse smaller, 1tgs larger) and are colored as following: Hydrogen bond donors - blue, acceptors - red, donors/acceptors - green, hydrophobic aliphatic - orange and aromatic - white. The surfaces of the binding sites are depicted as green dots. The catalytic residues of 1tgs (Ser-195, His-57 and Asp-102) are colored black and those of 1cse (Ser-221, His-64 and Asp-32) are brown. The catalytic Asp residues are not surface exposed and therefore were not considered by I2I-SiteEngine, however, the transformation of the solutions provides a good alignment between them.



Beta-Defensin BD In the following example, an interface of Beta-Defensin BD (1fd4) was compared to all cluster representatives. The top ranking interface was that of G-protein Rap1A in complex with the Ras-Binding-Domain (RBD) of C-Raf1 Kinase (1c1y). Examination of the cluster members represented by this complex revealed an interface comprised by Defensin HNP-3 protein (1dfn), which belongs to the same Defensin family as the query. Although the comparison was only between Beta-Defensin BD and the cluster representatives (none of which is from Defensin family), the resulting classification was correct.

SH3 Domain with a Peptide Another example is classification of an interface of Abl tyrosine kinase SH3 domain complexed with a peptide. Once again, this interface was compared to all the representatives. The top ranking interface was the catalytic domain of Prommp-2 E404Q mutant complexed with inhibitor (1eak). The cluster represented by this interface contains

another interface created by c-Src tyrosine kinase SH3 domain complexed with a peptide. As previously, this classification shows the consistency of the method. It is interesting to note that all of the peptides that create the interfaces of this cluster contain at least three Proline residues. Spatial similarity of two of these was recognized by the method. The matching of these relatively rigid residues may be an explanation for the obtained classification.

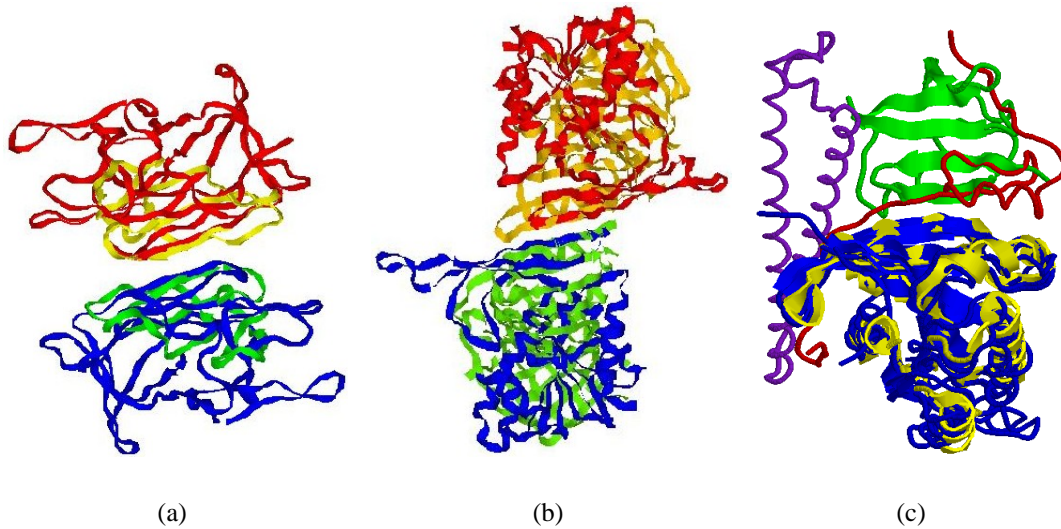


Figure 3: (a) Coagulation factor V (pdb:1czv, chain A - blue, chain B - red) and Cardiotoxin V4II(pdb:1cdt, chain A - green, chain B - yellow) (b) Alignment between Glycine N-Methyltransferase (pdb:1d2h, chain A - blue, chain B - red) and Proliferating Cell Nuclear Antigen (pdb:1axc, chain A - green, chain C - yellow). (c) Superimposition of the three top ranking results obtained in searching with a complete structure of G-protein cH-p21 Ras (yellow). In blue are the G-proteins of the recognized solutions. In green (1c1y), purple (1cxz) and red (1cee) are the binding partners of these interfaces that can potentially bind to the query.

Prediction of Binding Partners and Binding Modes Here we show how the presented method and classification can be used to predict the potential binding partners and modes of a cH-p21 Ras (G-protein). In this application we search the complete surface of this protein for the presence of a binding site, similar to those that constitute known interfaces. Such recognition can provide information regarding its potential binding partners and their binding modes.

From the algorithmic standpoint, comparison between a protein and an interface is only slightly different from comparison between two interfaces. The matching stage compares all of the triangles of the complete protein to I-triangles of each of the binding sites that comprise the interface. At the scoring stage each of these binding sites is scored separately. A region of a protein that achieves the highest score in a comparison to one of the binding sites is predicted to interact in a similar way. Therefore, the binding partners of that site can potentially bind to the recognized region. The superimposition of the complex of the interface on the protein provides prediction of its binding mode. When ranking the results of searching for different interfaces on the surface of a protein it is important to prevent the automatic selection of large interfaces, which have more features, thus receiving a higher score. Therefore, we divide the score of each pairwise comparison by the normalized score of the same binding site when it is searched in its native protein [8].

In order to predict potential binding sites of the G-protein cH-p21 Ras (1he8), its complete protein structure was compared to all of the representative interfaces. Three top ranking solutions were: (1) G-protein RhoA in complex with effector domain of the protein kinase pkn/prk1 (1cxz), (2) G-protein CDC42 in complex with the Gtpase binding domain of Wasp (1cee) and

(3) G-protein Rap1A in complex with c-Raf1 RBD (1c1y). Therefore, the binding partners of these three proteins may potentially bind to G-protein cH-p21 Ras at the recognized regions. Whereas in the first two cases we do not have any specific information regarding the correctness of the prediction, the third interface is created by an RBD domain similar to the one that binds to the query protein [15]. Figure 3(c) presents the superimposition of all these three top scoring interfaces on the query by the transformation recognized by the comparison of the corresponding interfaces. Here we have shown the correctness of the method by examples that can be verified by methods of sequence and structural alignment. However, the presented method considers only the surfaces and the functional groups of the interfaces and thus can recognize similarities shared by proteins with different overall sequences or folds. Thus it can recognize similarities that can not be detected by other methods.

6 Conclusions and Future Work

Recognition of similar patterns of interactions between evolutionary unrelated proteins is important for various biological applications. Here we presented a method that can recognize such similarities without any assumption regarding the similarity of the sequences or of the folds. Our method considers physico-chemical and geometrical considerations of the side-chains as well as of the backbone of the interfaces. It is efficient and can be applied to large scale database searches and classifications. However, it has several weaknesses. First, it addresses protein molecules as rigid bodies and considers flexibility only through a set of thresholds that allow a certain variability in the locations. Second, there no implicit treatment of electrostatic potentials that have a strong impact on the interaction. Such issues will be addressed in future research.

We have applied our method to classify a *pilot* dataset of protein interfaces and have shown its usefulness for searching applications. Although the constructed dataset is limited in size, it is sufficient to already show a clear representation of the interfaces clusters. Motivated by this experience we intend to apply I2I-SiteEngine to all PDB complexes for the complete classification of known interfaces. We hope that this will suggest preferred chemical organizations shared by similar interfaces. The insight we have gained from the pairwise interface alignment will facilitate the development of a tool for multiple interfaces alignment based on functional groups that we intend to develop next.

Acknowledgments We would like to thank Maxim Shatsky and Dina Schneidman for useful discussions and for contribution of software to this project. We would like to thank Dr. Shuo Liang Lin for valuable suggestions. This research has been supported in part by the “Center of Excellence in Geometric Computing and its Applications” funded by the Israel Science Foundation (administered by the *Israel Academy of Sciences*). The research of H.J.W. and A.S-P. is partially supported by the H. Minkowski-Minerva Center for Geometry at Tel Aviv University. The research of R.N. has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

References

- [1] Valdar, W. S. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.
- [2] Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
- [3] Keskin, A., Tsai, C. H., Wolfson, H. J. & Nussinov, R. (2004). A new, structurally non-redundant, diverse dataset of protein-protein interfaces and its implications. *Protein Sci.* . in press.

- [4] Kinoshita, K. & Nakamura, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Sci.* **12**, 1589–1595.
- [5] Schmitt, S., Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence or fold homology. *J. Mol. Biol.* **323**, 387–406.
- [6] Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001–1013.
- [7] Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
- [8] Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. (2004). Recognition of functional sites in protein structures. *J. Mol. Biol.* **in press**.
- [9] Connolly, M. (1983). Analytical molecular surface calculation. *J. Appl. Cryst.* **16**, 548–558.
- [10] Connolly, M. L. (1986). Measurement of protein surfaces shape by solid angles. *J. Mol. Graph.* **4**, 3–6.
- [11] Duhovny, D., Nussinov, R. & Wolfson, H. J. (2002). Efficient unbound docking of rigid molecules. In *Workshop on Algorithms in Bioinformatics*, (Guigo, R. & Gusfield, D., eds), vol. 2452, pp. 185–200. LNCS, Springer Verlag.
- [12] Cormen, T. H., Leiserson, C. E. & Rivest, R. L. (1990). *Introduction to Algorithms*. The MIT Press.
- [13] Mehlhorn, K. (1999). *The LEDA platform of combinatorial and geometric computing*. Cambridge University Press.
- [14] Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **34**, 827–828.
- [15] Paduch, M., Jelen, F. & Otlewski, J. (2001). Structure of small G proteins and their regulators. *Acta Biochim Pol.* **48**, 829–50.
- [16] Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.

7 Appendix

Below we present preliminary results of the classification of our pilot dataset of interfaces. Using I2I-SiteEngine, 64 interfaces were clustered to 22 different groups. Due to the relatively low number of interfaces under comparison, some clusters are single membered. For each interface the table specifies the PDB code of the complex, followed by the chains that constitute the interface as well as the SCOP classification of these chains.

| Cluster num. | PDB codes | SCOP [16] family of the first chain | SCOP [16] family of the second chain |
|--------------|--|---|---|
| 1 | 1ao7AE | MHC antigen-recognition domain | V set domains (antibody variable domain-like) |
| 2 | 1h27BE 1okuBF 1okvBE | Cyclin Cyclin Cyclin | peptide * peptide * peptide * |
| 3 | 1eakAP 1rlqCR | Fibronectin type II module SH3-domain | peptide * peptide * |
| 4 | 1g6rBH 1fo0BH | V set domains V set domains | MHC antigen-recognition domain MHC antigen-recognition domain |
| 5 | 1czvAB 1cdtAB 1dfnAB 1d2hAB 1c1yAB 1kbaAB 1ef7AB 1axcAC 1b77AB 1kklAH | Discoidin domain Snake venom toxins Defensin Glycine N-methyltransferase G proteins Snake venom toxins Papain-like DNA polymerase processivity factor DNA polymerase processivity factor HPr kinase HprK C-terminal domain | Discoidin domain Snake venom toxins Defensin Glycine N-methyltransferase Ras-binding domain, RBD Snake venom toxins Papain-like DNA polymerase processivity factor DNA polymerase processivity factor HPr-like |

| | | | |
|----|--|--|--|
| 6 | 1h26BE 1h28BE 1jsuBC | Cyclin Cyclin Cyclin | peptide * peptide * P27(KIP1) fragment 22-106 |
| 7 | 1ijdAC 1lghGJ | Light-harvesting complex subunits Light-harvesting complex subunits | Light-harvesting complex subunits Light-harvesting complex subunits |
| 8 | 1cseEI 1tgsZI 1acbEI 1bzxEI 1gl1AI 3tecEI 1sbnEI | Subtilases Eukaryotic proteases Eukaryotic proteases Eukaryotic proteases Eukaryotic proteases Subtilases Subtilases | CI-2 family of serine protease inhibitors Animal Kazal-type inhibitors CI-2 family of serine protease inhibitors Small Kunitz-type inhibitors and BPTI-like toxins PMP inhibitors CI-2 family of serine protease inhibitors CI-2 family of serine protease inhibitors |
| 9 | 1ao7AD | MHC antigen-recognition domain | V set domains (antibody variable domain-like) |
| 10 | 1a6aBC 1f3jBP 1d9kDP 1aqdBC 1jk8BC 1dlhBC | MHC antigen-recognition domain MHC antigen-recognition domain MHC antigen-recognition domain MHC antigen-recognition domain MHC antigen-recognition domain MHC antigen-recognition domain | peptide * peptide * peptide * peptide * peptide * peptide * |
| 11 | 1d09AB | Aspartate/ornithine carbamoyltransferase | Aspartate carbamoyltransferase, Regulatory-chain, C-terminal domain |
| 12 | 1gagAB | Protein kinases, catalytic subunit | peptide * |
| 13 | 1cxzAB 1hcoAB | G proteins Globins | Effector domain of the protein kinase pkn/prk1 Globins |
| 14 | 1ydtEI | Protein kinases, catalytic subunit | peptide * |
| 15 | 1hfoAB 1otgAB | MIF-related 5-carboxymethyl-2-hydroxymuconate isomerase (CHMI) | MIF-related 5-carboxymethyl-2-hydroxymuconate isomerase (CHMI) |
| 16 | 1bt6AB 1irjAB 1dt7AB 1e8aAB | Calcyclin (S100) S100 proteins S100 proteins S100 proteins | Calcyclin (S100) S100 proteins S100 proteins S100 proteins |
| 17 | 1b48AB 10gsAB 1pd212 1axdAB | Glutathione S-transferase (GST), N-terminal and C-terminal domains ** Glutathione S-transferase (GST), N-terminal and C-terminal domains ** Glutathione S-transferase (GST), N-terminal and C-terminal domains ** Glutathione S-transferase (GST), N-terminal and C-terminal domains ** | Glutathione S-transferase (GST), N-terminal and C-terminal domains ** Glutathione S-transferase (GST), N-terminal and C-terminal domains ** Glutathione S-transferase (GST), N-terminal and C-terminal domains ** Glutathione S-transferase (GST), N-terminal and C-terminal domains ** |
| 18 | 1iruOP 1iruFG 1g0uOP 1pmaAC | Proteasome subunits Proteasome subunits Proteasome subunits Proteasome subunits | Proteasome subunits Proteasome subunits Proteasome subunits Proteasome subunits |
| 19 | 1ic2CD 1gl2AB 1gl2BC 1gk4AB 1if3AB | Tropomyosin SNARE fusion complex SNARE fusion complex Vimentin coil theoretical model * | Tropomyosin SNARE fusion complex SNARE fusion complex Vimentin coil theoretical model * |
| 20 | 1ceeAB | G proteins | peptide * |
| 21 | 1eboAB | Virus ectodomain | Virus ectodomain |
| 22 | 1fzaAB | Fibrinogen coiled-coil and central regions | Fibrinogen C-terminal domain-like and Fibrinogen coiled-coil and central regions ** |

* - Proteins that are not classified by SCOP.

** - The interface region of this chain is composed of two different domains.