Generation and Analysis of a Protein–Protein Interface Data Set With Similar Chemical and Spatial Patterns of Interactions

Shira Mintz,[1†] Alexandra Shulman-Peleg,[2] Haim J. Wolfson,[2] and Ruth Nussinov[1,3*]
[1]*Sackler Institute of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*
[2]*School of Computer Science, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[3]*Basic Research Program, SAIC-Frederick, Inc., Laboratory of Experimental and Computational Biology, Frederick, Maryland*

ABSTRACT **Protein–protein interfaces are regions between 2 polypeptide chains that are not covalently connected. Here, we have created a nonredundant interface data set generated from all 2-chain interfaces in the Protein Data Bank. This data set is unique, since it contains clusters of interfaces with similar shapes and spatial organization of chemical functional groups. The data set allows statistical investigation of similar interfaces, as well as the identification and analysis of the chemical forces that account for the protein–protein associations. Toward this goal, we have developed I2I-SiteEngine (Interface-to-Interface SiteEngine) [Data set available at http://bioinfo3d.cs.tau.ac.il/Interfaces; Web server: http://bioinfo3d.cs.tau.ac.il/I2I-SiteEngine]. The algorithm recognizes similarities between protein–protein binding surfaces. I2I-SiteEngine is independent of the sequence or the fold of the proteins that comprise the interfaces. In addition to geometry, the method takes into account both the backbone and the side-chain physicochemical properties of the interacting atom groups. Its high efficiency makes it suitable for large-scale database searches and classifications. Below, we briefly describe the I2I-SiteEngine method. We focus on the classification process and the obtained nonredundant protein–protein interface data set. In particular, we analyze the biological significance of the clusters and present examples which illustrate that given constellations of chemical groups in protein–protein binding sites may be preferred, and are observed in proteins with different structures and different functions. We expect that these would yield further information regarding the forces stabilizing protein–protein interactions. Proteins 2005;61:6–20.**

Key words: protein–protein interactions; protein binding sites; data set of protein–protein interfaces; structural comparison of binding sites; alignment of interfaces, physicochemical properties

INTRODUCTION

Most cellular processes involve association and dissociation of protein molecules. Protein–protein interactions are highly diverged: They include homo- and heterocomplexes, and permanent and transient interactions. They may be large or small, hydrophobic or polar. The Protein Data Bank (PDB)[1] contains numerous examples of various types of complexes such as antigen–antibody, hormone–receptor, and enzyme–substrate/inhibitor. Given their central role in most biological processes, protein interactions have been addressed by many research groups.[2–9] They are crucial in practically all *in vivo* functions: cellular regulation, biosynthethic and degradation pathways, signal transduction, initiation of DNA replication, transcription and translation, multimolecular assembly, viral packaging, the immune response, and oligomer formation. Protein–protein interactions are involved in allosteric mechanisms, in turning genes on and off, and are important for drug design. Yet, deciphering the complex nature of protein interactions has proven to be a very difficult task. *In vivo*, a large portion of the protein surface may be involved in binding to various molecules. Moreover, some sites bind multiple proteins of different sizes, shapes, and composition. We may not know the identity and roles of these proteins, and the pathways that might be involved. Further, some binding reactions are cooperative events,

Grant sponsor: Center of Excellence in Geometric Computing and its Applications, funded by the Israel Science Foundation (administered by the Israel Academy of Sciences). Grant sponsor: Hermann Minkowski-Minerva Center for Geometry at Tel-Aviv University (in support of H. J. Wolfson). Grant sponsor: National Cancer Institute of the National Institutes of Health; Contract number: NO1-CO-12400 (to R. Nussinov). The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organization imply endorsement by the U.S. Government. The publisher or recipient acknowledges the right of the U.S. Government to retain a nonexclusive, royalty-free license in and to any copyright covering the article.

†This work was performed in partial fulfillment of the requirements for an M. Sc. degree.

*Correspondence to: Ruth Nussinov, Bldg. 469, Rm 151, NCI-Frederick, Frederick, MD 21702. E-mail: ruthn@ncifcrf.gov

Received 14 October 2004; Revised 13 February 2005; Accepted 4 March 2005

Published online 3 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20580

often with involvement of small molecules or of nucleic acid molecules. From the chemical standpoint, there are variable relative contributions of the hydrophobic effect *versus* electrostatic interactions, and a wide range of motifs that construct the interface region.

The extensive structural data stored in the PDB[1] allows a broad analysis of protein complexes. In this work we have used all 2-chain interfaces in the PDB to create a nonredundant data set that contains clusters of similar protein–protein interfaces. However, unlike in previously constructed data sets of protein–protein interfaces where the residues are represented by their C_α coordinates,[10,11] the current data set is uniquely created using functional (chemical) group and molecular surface shape representation of all protein–protein binding sites. Such a data set allows statistical investigation of interface families that are stored in these clusters, as well as identification and analysis of the chemical forces that account for the protein–protein associations and their local organization. Since the data set accounts for chemical–functional groups rather than for residues, similar groups at similar spatial locations are considered equivalent even if they originate from different residues with noncongruent backbone positions. We expect that such a data set would be useful in protein design, in targeting drugs to specific loci at the protein–protein interface, and in assisting in identification of specific functional groups that contribute significantly to the recognition and stability of the complex.

While a chemistry-based protein interface data set is very useful, its preparation is not straightforward. An interface consists of residues from 2 chains. In addition, the interface region in each of the chains may include several noncontiguous pieces. The functional group representation further makes it a difficult task, since we only use a single protein shell lining the binding site, and the functional groups are distributed on this shell as labeled points in 3-dimensional (3D) space, with the labels referring to the chemical group types. Under such circumstances, the protein sequence–order cannot be taken into account in the data set construction. Figure 1(A) presents an illustration of a protein–protein interface, where the 2 sides of the interface are slightly pulled apart for a clearer view. Figure 1(B) provides an example of a representation of 1 side of the interface. Figure 2 is an illustration of 2 binding sites that share a similar functional group constellation in space.

Several tools exist for the alignment, identification and classification of single binding sites based on their 3D structure. Artymiuk et al.[13] presented a subgraph–isomorphism-based[14] method, ASSAM, to identify the spatially conserved patterns of catalytic residues. This algorithm was recently updated to include additional properties such as solvent accessibility, disulfide bridges, and secondary structure elements.[15] Rosen et al.[16] used geometric hashing to search for regions on the surface of 1 protein that resemble a specific binding site of another. Russell[17] used a recursive depth–first search algorithm to find all possible groups of amino acids common to 2 protein structures. Kinoshita et al.[18,19] developed a method, based on clique detection,[20] that compares the molecular surface geometries and electrostatic potentials. Using this method, functional sites similar to those of phosphoenolpyruvate carboxy kinase were detected in proteins with different folds. Schmitt et al.[12] developed a model for a representation of amino acid functional groups by pseudocenters that denote centers of potential interaction. Using this representation together with a clique detection algorithm,[20] a method for comparison between binding sites was presented by Bron and Kerbosch. They constructed a database of binding sites, Cavebase, and searched it with cavity queries. Wallace et al. presented TESS,[21] an algorithm that derives 3D templates from protein structures. The algorithm was applied to classify the catalytic triads of enzymes such as serine proteases, triacylglycerol lipases, ribonucleases, and lysozymes.[22] The successor of TESS, the JESS algorithm,[23] is a fast and flexible algorithm for searching protein structures for small groups of atoms, with some constraints on geometry and chemistry. Brakoulias and Jackson[24] presented a method for the comparison of protein-binding sites using geometric matching to detect common atomic features. After performing an all-against-all comparison of phosphate binding sites in a number of different nucleotide phosphate-binding proteins, 10 main clusters have been obtained that represent a limited number of unique structural binding motifs for phosphate. Shulman-Peleg et al. presented a geometric hashing–based method, SiteEngine,[25] which can search large protein structures for regions similar to the binding site of interest. Several other methods exist for the identification of local structural similarities in proteins.[26] However, most of these methods are not suitable for our purpose due to the following reasons: (1) The problem of clique detection and subgraph isomorphism is known to be NP-complete[27,28] and its application to the classification of the entire PDB will therefore be time-consuming; (2) all of the methods described above compare binding sites and do not consider their interacting partner. These methods can be applied to separately classify the binding sites that constitute the interfaces. However, it is unclear how the clusters of binding sites can be combined to create clusters of 2-sides interfaces, especially since the transformations that align the binding sites of the same interface may differ and may have a different rank.

Sequence patterns have become widely used for binding classifications.[29–31] A sequence of a newly determined protein may be scanned against a database of sequence patterns in order to identify the protein family it belongs to. However, many functionally similar binding sites are comprised of noncontiguous pieces that are sequence–order independent.[17,22] In such cases, methods that assume sequence order may not be applicable. A method that aligns noncontiguous pieces from both sides of the interface and is suited for the alignment of protein–protein interfaces was used by Tsai et al.[10] In their method, they used the computer vision-based geometric hashing structural comparison technique.[32] The algorithm, which is sequence and directionality independent, uses the C_α atom coordinates and does not require connectivity among the

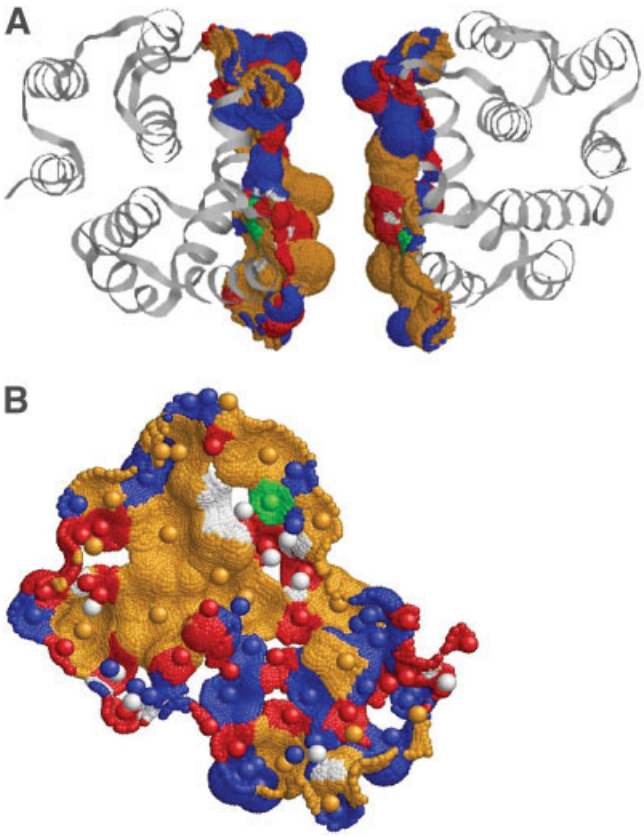

Fig. 1. Physicochemical and surface representation of an interface (PDB code: 1aa7). (**A**) Illustration of a protein–protein interface with its surface representation. Colors of the surface patches are determined by the physicochemical properties of the functional groups to which it belongs. The 2 sides of the interface are pulled apart for a clearer representation. (**B**) An example of a surface representation of one side of the interface. Following Schmitt et al.,[12] each pseudocenter (represented as a colored ball) represents an interaction center of one of the following physicochemical properties: hydrogen bond donor (blue), hydrogen bond acceptor (red), mixed donor–acceptor (green), hydrophobic aliphatic (orange), and hydrophobic aromatic (white). Each surface point is assigned a physicochemical property according to the functional group to which it belongs.

C_α points in the matching. The similarity between 2 protein–protein interfaces was measured based on the extent of the geometrical superposition between their corresponding C_α atoms, the percent residue identity in the match, and the differences in their sizes. This method was used by Tsai et al. to create a nonredundant data set of protein–protein interfaces.[10] As the number of the complexes in PDB grew, Keskin et al.[11] reapplied the method to build an updated data set. In the construction of the data set, the residues considered were those that interact across the 2-chain interface and residues in their spatial vicinity. The C_α atom geometrical representation led to clusters that contain similar interface architectures formed by the 2 chains. Only geometrical constraints and backbone coordinates were used in this procedure, although side-chains are well known to play important roles in the interaction between protein molecules. Another 2-sides interface classification was carried out by Mintseris and Weng.[33] In their work, they used atomic contact vectors (ACVs) as a way to represent the physicochemical properties of interfaces and used the clique detection algorithm[20] to compare them. They applied the ACVs to 2 classification problems—distinguishing homodimers from crystal contacts and obligate oligomers from transient recognition complexes. Using the ACVs, a nonredundant data set of transient recognition complexes was created.

Here, our goal is to investigate the organization of the functional groups derived from both the backbone and side-chains of amino acids. We expect that clusters of interfaces with common chemical groups would be useful in figuring out the organization of the stabilizing intermolecular interactions. Recognition of similar binding organizations shared by different protein–protein complexes is performed by pairwise alignment between the corresponding protein–protein interfaces. Alignments that are based on sequence similarity may not be sufficient for this task, since interacting residues may have similar constellations in space, but their order on the polypeptide chain may be

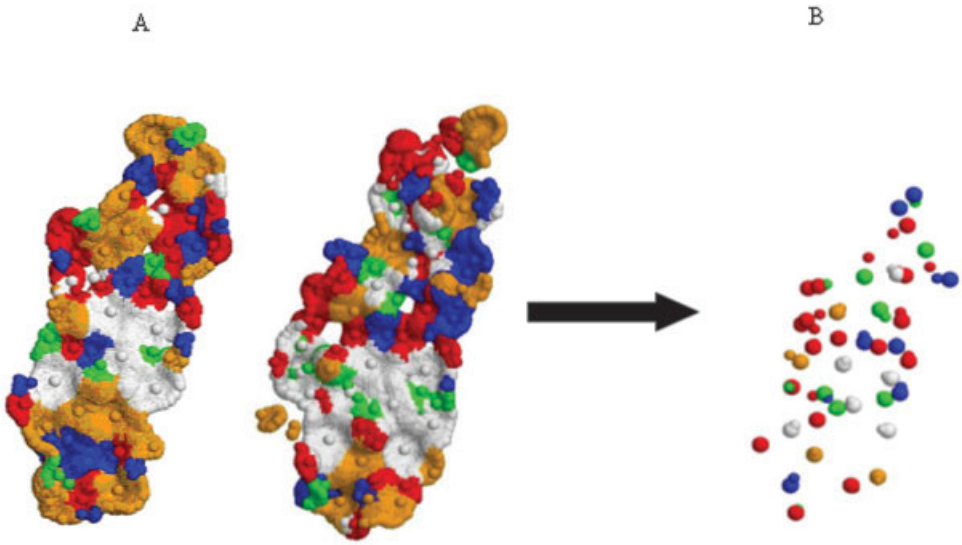

Figure 2.

different. Each region may be derived from residues in different parts of the sequence, or even from single sequentially isolated residues. The problem gets even harder when attempting to align regions consisting of functional chemical groups that may derive from different residues.

To address these requirements, we have developed I2I-SiteEngine (Interface-to-Interface SiteEngine). Unlike SiteEngine,[25] whose goal is recognition of binding sites of small molecules in protein structures, I2I-SiteEngine[34] has been designed to recognize similarities between protein–protein interfaces. I2I-SiteEngine is independent of the sequence or the fold of the proteins that comprise the interfaces. In addition to geometry, the method takes into account the physicochemical properties of the interacting atom groups, both backbone and side-chain. The method performs a simultaneous alignment of 2 binding sites that constitute an interface. Here, we briefly describe the method and focus on its application to the classification of all complexes available in the PDB. We present the classification procedure, as well as the results obtained from analysis of the interface clusters. These illustrate that similar constellations of chemical groups in the protein–protein binding sites may be preferred, and are observed in proteins with different structures and different functions. Recognition of common favored binding organizations shared by different protein families may suggest their contribution to the formation and stability of the complex.

The "hot spot" residues occur in a broad range of interfaces.[35–37] A mutation of a hot spot residue to alanine leads to a significant ($\Delta\Delta G > 2$ kcal/mol) drop in the free energy of the protein–protein association. The hot spots have been defined at the residue level. It has further been shown that the experimental hot spots are highly correlated with the conservation of interface residues in families of 3D structures.[38,39] Our data set will allow the examination of hot spots at the functional group level, hopefully elucidating the chemical origin of the stabilizing interactions, assisting in the identification of their type and organization. Additionally, they may be useful in the generation and design of drug targets that prevent the formation of protein–protein complexes. It may also assist in the prediction of side effects caused by similar binding partners.

METHODS

Data Set Generation

The construction of the protein–protein interface data set includes 5 stages: (1) initial extraction of all 2-chain interfaces from the PDB; (2) filtering of the interfaces to reduce redundancy and the number of pairwise comparisons; (3) generation of interfaces; (4) structural comparisons; (5) ranking; and (6) clustering.

Stage 1: Initial extraction of interfaces from the PDB

Our definition of interface residues is similar to that of Tsai et al.[10] and Keskin et al.[11] Two residues from different chains are defined as interacting residues if the distance between any 2 atoms of the 2 residues is less than the sum of their corresponding van der Waals radii plus 0.5 Å. We extracted all 2-chain complexes from the PDB (excluding modeled structures, structures with a resolution lower than 3.5 Å, and structures with only C_α coordinates) that have at least 10 interacting residues in each chain (the September 2003 release). In this scheme, an interface will include only 2 chains, and a certain polypeptide chain might appear in more than 1 interface, according to the number of polypeptide chains that interact with it. Through this procedure [Fig. 3(A)], 23,912 two-chain interfaces were collected. These interfaces include different types of complexes: homodimers, heterodimers, and portions of higher complexes. Following the nomenclature of Tsai et al.,[10] an interface is named according to its PDB code and the chains it contains (e.g., 1aa7AB is an interface between chain A and chain B, derived from PDB code 1aa7).

Stage 2: Filtering procedures

The PDB is redundant. Many structures share a high percentage of sequence identity, as well as a high structural similarity. In an attempt to minimize the number of interfaces to be compared to each other at the structural comparison stage, 2 filters were applied (Fig. 3). The first filter rules out interfaces that have no biological significance. Protein complexes may yield a set of coordinates that are not independent of the crystallographic symmetry (space group and unit cell). As a result, in a PDB entry, the deposited coordinates may include multiple copies of the complex. PQS[40] (Protein Quaternary Structure) is an automated procedure that has been devised to recognize where multiple copies exist. We used the PQS server in order to rule out interfaces with 2 chains that belong to 2 separate copies of the same complex. Such interfaces most likely have no biological significance. The first filtering procedure removed 2109 interfaces from the initial list [Fig. 3(B)].

The second filtering procedure removes structurally redundant complexes as follows: All 2-chain complexes were grouped by their family structures obtained from Structural Classification of Proteins (SCOP).[41] Each of these SCOP groups contains 2-chain complexes that share the same family structure (e.g., 1aa7AB and 1smtCD are in the same SCOP group if chain A from 1aa7 and chain C from 1smt belong to the same SCOP structural family, and chain B from 1aa7 and chain D from 1smt belong to the same SCOP structural family). For each SCOP group a representative was chosen. Using the MASS method,[42,43]

Fig. 2. An example of 2 similar binding sites that share a similar functional group constellation in space (PDB codes: 1jh5, 1d0g). (**A**) Physicochemical and surface representation of 2 TNF proteins. The interfaces consist mostly of aromatic and aliphatic patches (colored in white and orange, respectively). Donor (blue), acceptor (red) and donor–acceptor (green) patches are also seen on the interface. (**B**) A view of the matched functional groups between the 2 interfaces. The larger balls represent the functional groups of 1jh5 and the smaller represent the functional groups of 1d0g. From a total of 39 matched functional groups, 6 aromatic centers are matched (colored in orange).

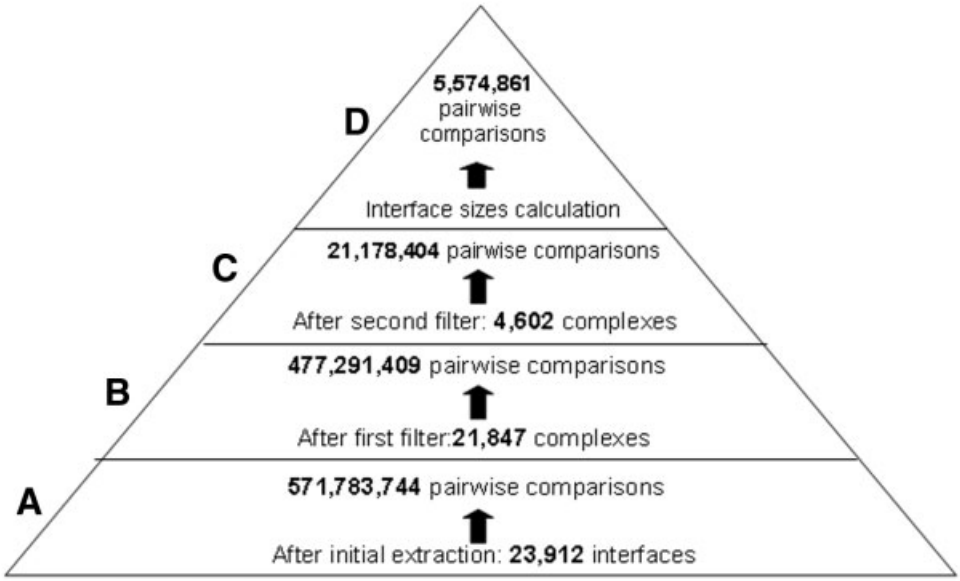

Fig. 3. Filtering procedures to reduce redundancy and minimize the number of interface comparisons. (**A**) The initial step of the data set generation is the extraction of all interfaces from the PDB. An all-to-all pairwise comparison of the 23,912 interfaces will require ~570 million comparisons. (**B**) The first filter rules out 2065 interfaces that have no biological significance (using PQS). (**C**) The second filter removes structurally redundant complexes, resulting in 4602. (**D**) Only 2 interfaces that are compatible in size are being compared, resulting in less than 6 million comparisons.

which performs structural alignment by secondary structures, all complexes in each SCOP group were structurally aligned to the representative of the corresponding group. If the pairwise alignment yields a score above a given (highly strict) threshold,* the complex is retained in the group. If the score is lower than the threshold, the complex is now the new representative of a new group. The groups obtained by this process contain interfaces that share very high structural and sequential similarity. At the final step, we extract only the representatives of the groups, resulting in 4602 interfaces [Fig. 3(C)]. The process described here is performed in order to eliminate structural redundancy. Since a very strict threshold is used in the structural comparison process, interface representatives that share a rather high sequence similarity might still exist. The elimination of sequence similarity redundancy is described below.

Stage 3: Generation of interfaces

Prior to the structural alignment procedure, we preprocess the interfaces of the retained complexes. An interface is defined as an unordered pair of interacting binding sites, which belong to 2 noncovalently linked protein molecules. Each interface is represented by the functional groups of its 2 interacting binding sites, as well as their corresponding physicochemically labeled surfaces. The main stages of the interface generation are summarized in Figure 4. Given a protein–protein complex, the first step is the construction of a smooth molecular surface[44,45] for each of the interacting protein chains. Since we are interested only in the regions of interaction, for each protein chain, we retain only surface points that are within a distance of 4 Å from the surface of its binding partner. Given the atomic coordinates of the interacting protein structure, we extract the functional groups that are exposed to the interface surface. Those are determined by the physicochemical properties of the residues (backbone and side-chains). We follow Schmitt et al.[12] in the definition and representation of each functional group of an amino acid by a 3D point, denoted as pseudocenter. Each pseudocenter represents a center of potential interaction and has one of the following properties: hydrogen bond donor, hydrogen bond acceptor, mixed donor–acceptor, hydrophobic aliphatic and aromatic (π) contacts. For example, the model represents the side-chain of arginine by 3 pseudocenters of type donor (located on nitrogen atoms) and a pseudocenter of type aliphatic (located at the center of mass of the 3 carbon atoms). The side-chain of proline is represented by an aromatic pseudocenter located at the center of its aromatic ring.

Finally, we perform a physicochemical labeling of the interface surfaces according to the properties of the pseudocenters. The obtained surfaces and the pseudocenters of the 2 binding sites are used to represent an interface.

Stage 4: Structural alignment between interfaces

Toward the goal of classification of protein–protein interfaces, we have developed a method, I2I-SiteEngine, which simultaneously aligns between 2 pairs of binding sites and is independent of the sequences and the folds of

*The root-mean-square deviation (RMSD) of the alignment is required to be less than 1.5 Å and at least 85% of the protein should be included in the alignment.

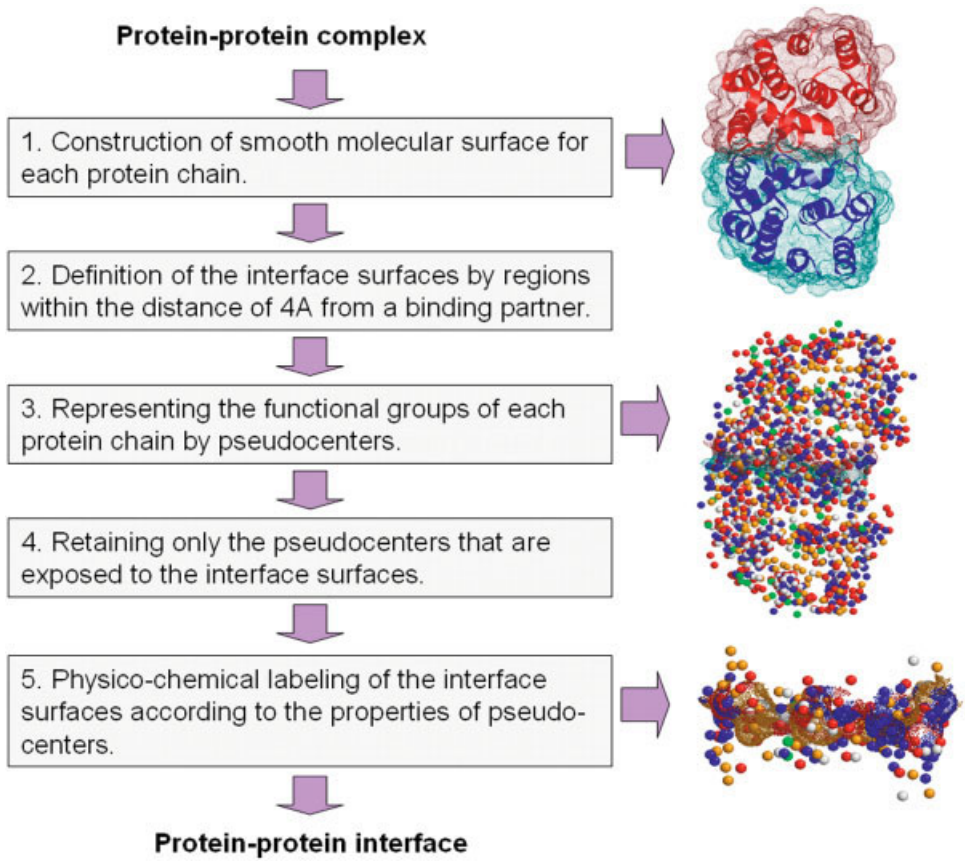

Fig. 4. The main stages of the interface generation process.

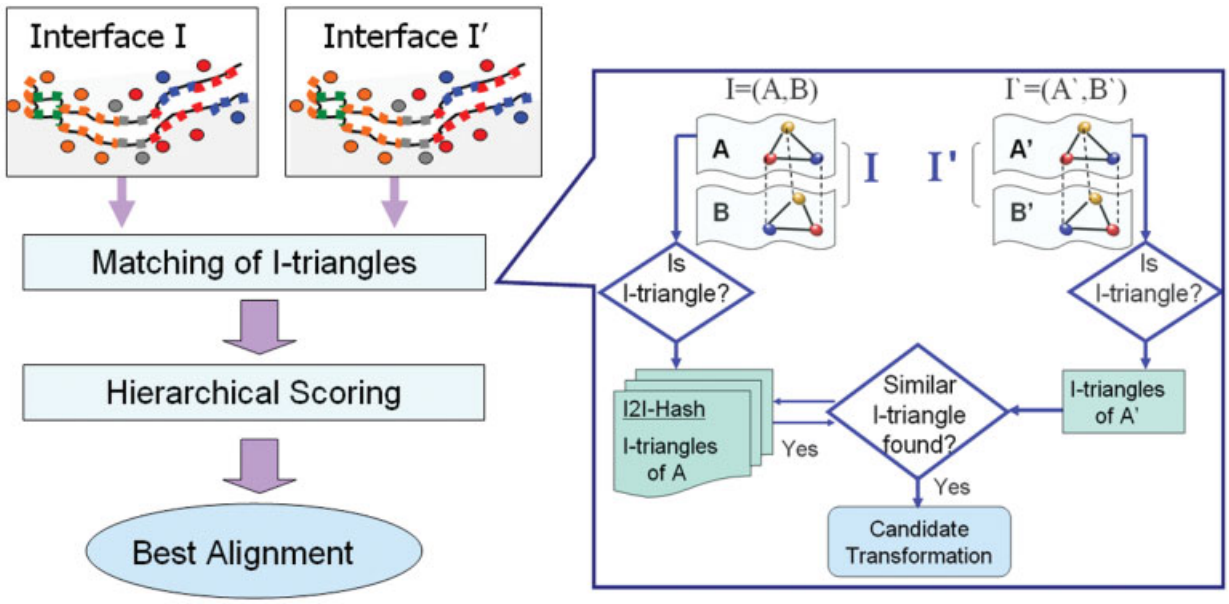

Fig. 5. The main stages of the I2I-SiteEngine algorithm. Given 2 protein–protein complexes, the first stage is the representation of the surfaces and the physicochemical properties of the 2 interfaces. Each triplet of pseudocenters of the binding sites A and A′ is considered. Triplets that have complementary properties in the binding sites B and B′, respectively, are considered to form I-triangles. I-triangles of the 2 interfaces are matched using a geometric hash table with keys defined by the triangle side lengths and physicochemical properties. Each pair of matched I-triangles defines a candidate transformation, which is scored by a set of hierarchically applied scoring functions. These measure the similarity of the physicochemical surfaces, shapes, and pseudocenters of the 2 interfaces aligned by a candidate transformation.

the corresponding proteins. The I2I-SiteEngine software is freely available through the Web server: http://bioinfo3d.cs.tau.ac.il/I2I-SiteEngine/. Below we briefly outline the main stages of the I2I-SiteEngine algorithm. Additional details can be found in Shulman-Peleg et al.[34] Given 2 interfaces, I = (A,B) and I' = (A′,B′), represented by the physicochemically labeled surfaces and pseudocenters as described above, the goal is to find a transformation (rotation and translation) that will maximize the similarity score of the superimposed surfaces and physicochemical properties. The main stages of the algorithm are summarized in Figure 5. Since the correspondence between the binding sites of the 2 complexes is unknown, binding site A can be aligned either to A′ or to B′. To consider these 2 possible options, we repeat stages (2) and (3) twice and select the correspondence that achieves the highest score. The algorithmic stages are the same for both cases and thus are described below only for the alignment of A to A′.

Matching of I-triangles. The goal of the matching stage is to calculate all candidate transformations that can superimpose one interface onto the other. We introduce a definition of an "Interacting triangle" (*I-triangle*), which is a triplet of functional groups (pseudocenters) in one chain that is recognized to form 3 interactions with the other chain. An interaction is determined by the presence of 2 pseudocenters, one from each of the interacting binding sites, which have complementary properties at suitable spatial locations. Specifically, hydrogen bond donors are complementary to hydrogen bond acceptors, whereas aromatic and aliphatic properties can interact with similar ones respectively.

A hashing procedure, detailed in the flowchart of Figure 5, is applied to recognize all almost-congruent I-triangles of the 2 interfaces.[34] Each pair of matched I-triangles defines a rigid transformation (rotation and translation), which is evaluated by the subsequent scoring functions.

Hierarchical Scoring. The scoring is based on a set of hierarchically applied scoring functions that measures the similarity of the physicochemical properties and shapes aligned by a candidate transformation.[34] First, using a low-resolution representation of the surface of each interface, it performs a fast estimation of the obtained similarity. As the number of potential solutions is reduced, the resolution of the molecular representation is increased, leading to more accurate calculations. For 100 top-ranking candidate solutions, we determine a match list of the aligned pseudocenters of the 2 interfaces. This match list is used to calculate a score that estimates the similarity between the matched pseudocenters. The match list is calculated by a maximum weight bipartite matching algorithm[46] on the graphs constructed from the pseudocenters of the 2 interfaces.[34] Finally, each candidate transformation is assigned a score that is a weighted sum of all the scores calculated by the method.

After selecting the top-ranking solution, the output of the program consists of (1) a similarity score defined by the scoring functions; (2) a 3D transformation of the selected solution; and (3) a list of the matched pseudocenters. The method is robust and suitable for our goal of large-scale classification. Its mean running time, measured for the 5,574,861 comparisons performed in this classification, is 26 s [Intel(R) Xeon(TM) CPU 2.40 GHz]. The calculated running time includes the 2 comparisons performed for each possible correspondence. It does not include the time required for the preprocessing of the surfaces and grids, which is performed once for each molecule.

Stage 5: The similarity ranking stage

We applied I2I-SiteEngine to align each pair of interfaces that were retained after the filtering procedure (4602 interfaces) and had a compatible surface area [Fig. 3(D)]. We consider the surface area of 2 interfaces to be compatible if the size of their reentrant surface area[47] differs by less than 50% of the larger interface. Specifically, the structural comparison is performed for each pair of interfaces that fulfill:

$$(\text{small interface}*2) \geq \text{large interface size}$$

In order to rank the results of the pairwise comparisons, we normalize the final score calculated by I2I-SiteEngine. For each comparison between 2 interfaces I and I', the normalization is performed in the following manner:

$$S(I,I') = \frac{\text{Score of } I \text{ aligned to I}'}{\text{Score of } I \text{ aligned to itself}} \times 100. \qquad (1)$$

Interface I in this example is assumed to be larger than I'. The normalized score represents the portion of the aligned region out of the total binding region of the larger interface. It must be noted that the calculated score measures not only the similarity of the pseudocenters of the 2 interfaces but also the similarity in the location and the physicochemical property of every surface point. Therefore, a 100% match is expected to be obtained only between an interface and itself. Even the alignment between 2 interfaces that are derived from homologous proteins and share high sequence similarity is expected to yield a much lower score due to the fluctuations of the protein surface. The calculated normalized scores are used to build a distance matrix that is used in the final clustering stage and represents the similarity relationships between the compared interfaces.

Stage 6: The Clustering Stage

A clustering algorithm was applied to the final scores in order to group the interfaces in clusters according to their similarity. A common clustering algorithm that is used for pairwise comparisons is the Hierarchical Clustering Algorithm.[48] This algorithm transforms a distance matrix, which is the result of pairwise similarity measurements between elements of a group (e.g., interfaces), into a hierarchy of nested clusters.

Hierarchical algorithms are classified into 2 kinds: *Agglomerative* and *Divisive*. The Agglomerative procedure starts with each object forming a cluster containing only itself. Then, the number of clusters is iteratively reduced by merging the 2 most similar clusters, until only 1 cluster

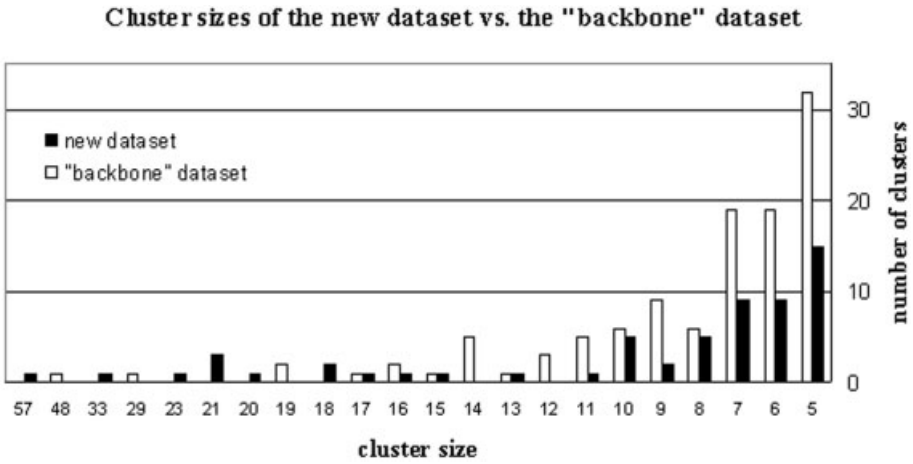

Fig. 6. Cluster sizes of the new data set versus the "backbone" data set. The total number of clusters with at least 5 members is smaller in our data set (59 clusters vs 103 clusters), as well as the number of interfaces these clusters contain (604 in our data set vs 949 in the "backbone data set").

remains. The Divisive procedure starts with 1 cluster and iteratively splits a cluster until all objects are separated. The Agglomerative procedure, used in this work, requires fewer computations to proceed from one level to another, and therefore is used more frequently. The crucial step in the Agglomerative procedure is the merging of the 2 most similar clusters. Different distance functions are used to measure the similarity between different clusters and can lead to different results. In this work we used the *average linkage distance method*[48] to define the distance between 2 clusters C_i, C_j:

$$D_{\text{ave}}(C_i,C_j) = \frac{\sum\sum S(k,l)}{k\epsilon C_i, l\epsilon C_j} \Big/ m*n, \qquad (2)$$

where $S(k,l)$ is the pairwise similarity score between interface k and interface l [defined in Eq. (1)], and m and n are the number of interfaces in clusters C_i and C_j, respectively. In this method, the distance between 2 clusters is defined as the average distance between all possible pairs of interfaces within the 2 clusters. The 2 clusters with the smallest average distance between their data entries are merged at each step. In our implementation of the algorithm, the merging process stops when the average distance between 2 clusters is larger than a predetermined threshold value. The threshold for the similarity score [Eq. (1)] was determined by a preliminary evaluation on a pilot data set[34] and was set to 25. Other distance functions include the single linkage that measures the distance between clusters by the distance between the 2 closest points within the clusters, and the complete linkage that measures the distance by the distance between the farthest pair of data points within 2 clusters. However, both the single linkage and the complete linkage methods are very sensitive to noise and to outliers. The average linkage method is a compromise and is more stable to unknown distributions of objects. A drawback of the average linkage method is that given a threshold value of similarity for the merging stage, the final results of the clustering algorithm might include 2 data points in 1 cluster that share a lower similarity than the threshold. In order to overcome this problem, for each such pair of interfaces, the interface that is less similar to all the rest of the interfaces in the cluster is removed from the cluster and generates a new cluster.

RESULTS AND DISCUSSION

Generation of a Nonredundant Data Set

After applying the clustering algorithm to the pairwise alignment scores, the interfaces that were removed by the *structural* alignment procedure at the initial filtering stage were added to the clusters according to their representatives. This procedure resulted in 2582 clusters. Of these, 740 contain a single interface. In order to be able to carry out a statistical analysis on the data set, a nonredundant data set was generated from the clusters. To eliminate redundancy, the sequences of the chains comprising the interfaces in each cluster were compared by BlastClust,[49] using a threshold identity of 50%. The sequence threshold that is used in SCOP[41] for the classification of the families is 30%. Our method for the alignment between the interfaces considers a very high-resolution representation of the physicochemical surfaces and is sensitive to local changes. In order not to lose relevant data, we have selected a slightly more permissive threshold of 50%. If 2 interface members in the same cluster share a sequence identity of more than 50% in both their chains, one member was removed from the cluster. Finally, a representative was selected for each cluster. The representative is the interface that is most similar to the rest of the cluster members.

Comparison to the "Backbone Data Set"

Figure 6 presents a comparison between the sizes of the clusters in the nonredundant data set obtained by Keskin et al.[11] (the "backbone data set") and the sizes of the clusters obtained by our method here. The backbone data

set is generated at the amino acid level, and the structural comparisons are carried out using the C_α coordinates. The nonredundant backbone data set includes only clusters with 5 or more interfaces. In order to make an appropriate comparison, Figure 6 shows clusters with at least 5 members for both data sets. Both data sets were obtained from a comparable number of interfaces (21,663 interfaces here vs 20,937 in Keskin et al.). However, different representation of the interfaces, different clustering methods, thresholds and parameters, were used that affect the final results. As can be seen in the chart, the total number of clusters with at least 5 members is smaller in our data set (59 clusters vs 103 clusters), as well as the number of interfaces that these clusters contain (604 in our data set vs 949 in the backbone data set). The stronger tendency of interfaces to cluster in the backbone data set is expected and emphasizes the major difference between the 2 methods. While the current I2I-SiteEngine alignment method focuses on the level of the chemical groups and demands similarity between the physicochemical properties of the aligned surfaces, as well as similarity of the geometrical shape, the alignment method used in the backbone data set is based on the geometry of the protein backbone alone, thus focusing on the motifs and the residue-based constellations that are observed in the interfaces. When aligning 2 interfaces solely by their C_α atom coordinates, interfaces with recurring secondary structures motifs will receive high alignment scores even though the side-chains atoms of the interacting residues may display dissimilar orientations and chemical properties.

The Clusters

Similar interfaces, similar overall fold

The function of a protein is most often determined by its ability to interact with other proteins (e.g., enzyme–inhibitor, antibody–antigen). Most often, proteins with similar function will also share similar overall folds. Therefore, the fact that interfaces with similar function and fold are observed to be clustered is not surprising. In our nonredundant data set (that contain only clusters of interfaces with no more than 50% sequence identity), 15 such clusters with 5 or more interfaces that share the same overall superfamily structure (following SCOP) are observed. An example for this type of cluster can be seen in Figure 2. Cluster 673 contains 5 members of the TNF (tumor necrosis factor) family proteins that play a role in mammalian cell host defense processes, inflammation, apoptosis, autoimmunity, and organogenesis.[50] The TNF ligand trimer structure consists entirely of β strands and loops. The interface between 2 monomers of the trimer consists of aromatic residues (which are conserved for all TNF family members). Hydrophobic interactions appear to be the main force driving trimer formation.[50] In Figure 2(A), which gives the surface representation and the functional groups of one side of 2 TNF interfaces, the aromatic patches are clearly seen. From a total of 39 matched functional groups between the 2 interfaces, 6 aromatic centers are matched [Fig. 2(B)].

Similar interfaces, different overall folds

A more surprising type of cluster is when the interfaces are clustered in the same group (i.e., they share similar spatial binding organizations), yet they belong to a different SCOP superfamily (i.e., the chains that compose them have different global folds). Forty-six such clusters with 5 interfaces or more are observed in our nonredundant data set. Moreover, although stable and transient complexes may exhibit different features,[4] the discrimination between their physicochemical properties is not straightforward,[51] and several clusters of this type contain both types of complexes. While the previous type of clusters is to be expected, this type can suggest either preferred binding organizations for proteins with different functions, or similar functions shared by unrelated proteins (possibly selected by convergent evolution). The subtilisin-like and trypsin-like folds are a well studied example of such a case, where the proteins have different overall folds, yet they share a similar function.[22,52] In our data set, interfaces comprised of these proteins complexed with their inhibitors are clustered. Figure 7 shows the multiple alignments between 1 side of the 8 interfaces from this cluster: Five interfaces (PDB codes: 4sgb, 1ppf, 1acb, 1an1, and 1gl0) belong to the trypsinlike serine proteases superfamily, and 3 interfaces (1cse, 2sic, and 1oyv) belong to the subtilisin-like superfamily. The proteins from both superfamilies are known to have a Ser–His–Asp catalytic triad in their active sites (presented as purple sticks). The alignment shown in Figure 7 was done with the MultiBind algorithm.[63] The MultiBind program performs multiple alignments of binding sites in order to detect binding patterns that are common to a family of proteins. Nine structurally conserved functional groups were detected when applying MultiBind to the cluster. Functional groups marked with (*) are derived from the same residue in each of the 9 proteins. As can be seen, 2 structurally conserved functional groups (aromatic, white; donor–acceptor, green) are derived from the catalytic histidine residue, 2 conserved functional groups (donor, blue; donor–acceptor, green) are derived from the catalytic serine residue, and another structurally conserved functional group (acceptor, red) is derived from another serine residue in all 9 proteins. The other 4 conserved functional groups (aliphatic, orange; 2 acceptors, red; and donor, blue) are derived from various residues. As can be seen in Figure 7, only 2 of the catalytic residues (His and Ser) are aligned. The third catalytic residue (Asp) is not exposed on the protein surfaces and therefore is not considered in the alignment.

Two similar interfaces with different overall folds can also be seen in Figure 8. Both proteins are kinases: The first (1b99) belongs to the nucleoside diphosphate kinase superfamily, and the second (1l0o) to the histidine kinase superfamily. We used this case in order to examine whether sequence alignment methods can detect the similarity between the interfaces. Here, the binding region is constructed of contiguous sequence residues, which makes it suitable for sequence alignment. When we used BLASTP[49] with the overall chains of the 1b99 complex as a query against the full PDB data set, no proteins belonging

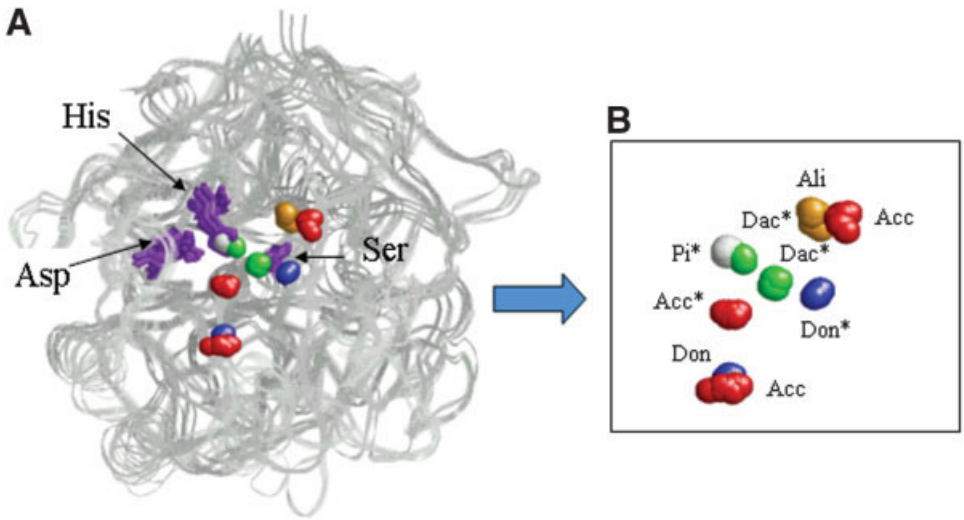

Fig. 7. Eight clustered proteins with 2 different folds. (**A**) Multiple alignment between 1 side of 8 interfaces that belong to 2 different superfamilies: trypsin-like serine proteases (PDB codes: 4sgb, 1ppf, 1acb, 1an1, 1gl0) and subtilisin-like (1cse, 2sic, 1oyv). The Ser–His–Asp catalytic triad is shown as purple sticks. (**B**) Nine structurally conserved functional groups were detected. Five of them are derived from the same residue in each of the 8 proteins (marked with *): aromatic (white) and donor–acceptor (green) functional groups are derived from the histidine catalytic residue; donor (blue) and donor–acceptor (green) functional groups are derived from the serine catalytic residue, and a conserved acceptor (red) is derived from another serine residue. Four conserved functional groups are derived from variant residues: donor (blue), 2 acceptors (red), and an aliphatic (orange) functional group. Only 2 of the catalytic residues (His and Ser) are aligned. The third catalytic residue (Asp) is not exposed on the proteins surfaces and therefore is not considered in the alignment.

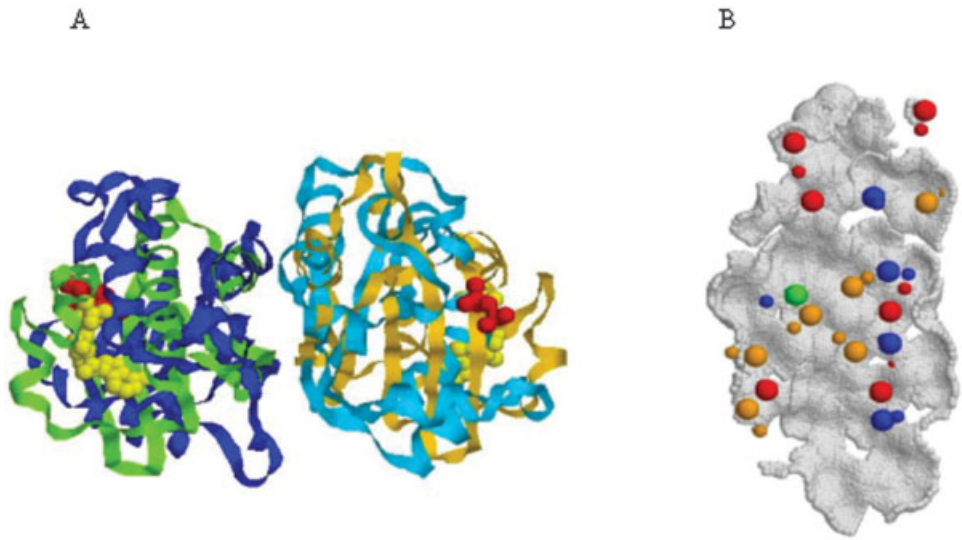

Fig. 8. Two similar interfaces with different overall folds. (**A**) Alignment between nucleoside diphosphate kinase (PDB code: 1b99, colored blue and cyan) and a histidine kinase (PDB code: 1l0o, colored green and orange), using the transformation obtained by the interface alignment. The red and yellow balls represent the ligands of the proteins: FUP and POP, red balls; ADP, yellow balls, for 1b99 and 1l0o, respectively. (**B**) The matched functional groups between the 2 interfaces are all aliphatic (orange balls) or polar (donor, blue; acceptor, red; and donor–acceptor, green). The larger balls represent the functional groups of 1b99 and the smaller represent the functional groups of 1l0o.

to the histidine kinase superfamily were detected. Next, we used bl2seq[49] to make a pairwise alignment between the binding region sequence of 1b99 chain and 1l0o chain. Again, no significant similarity was found. Finally, we searched for sequence patterns using PROSITE.[30] No common sequence patterns were found in the binding region of the 2 proteins. In our data set, both complexes are clustered together. Figure 8(A) presents the alignment between the 2 complexes based on the transformation obtained from the structural interface alignment. Both interfaces are comprised of α-helices and β-strands. As can be seen, the interfaces are located at opposite sides from

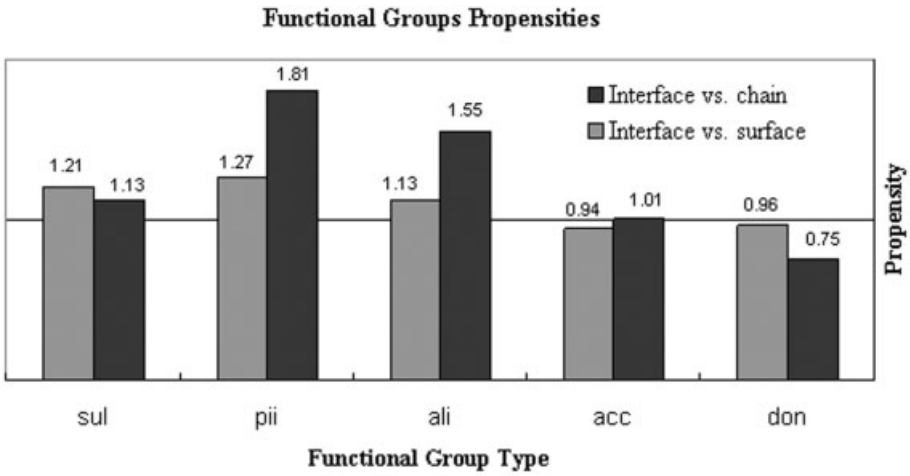

Fig. 9. Propensities of donor, acceptor, aliphatic, aromatic, and sulfur functional groups in the interfaces compared to overall chains (black) and to surfaces (gray). A propensity of greater than 1 indicates that a functional group occurs more frequently in the interface than on the protein chain or surface.

the ligand binding sites (yellow and red space-filled spheres). In Figure 8(B), the matched functional groups of 1 side of each interface are shown. Out of the 17 matched functional groups, 6 represent an aliphatic property, while the other 11 are polar (donor or acceptor).

In the absence of sequence or functional similarity between 2 proteins, a common ancestry is hard to discern. Sandhya et al.[53] detected remote evolutionary relationships by searching in a specialized database of sequences belonging to the same fold. They have assessed their method by exploring the relationships they detected among sequences known to belong to the triosephosphate isomerase (TIM) β/α-barrel fold. An interesting connection was detected between fructose–bisphosphate aldolase II (PDB code: 1b57), which belongs to the aldolase superfamily, and the histidine biosynthesis enzyme (PDB code: 1qo2), which belongs to the ribulose–phosphate-binding barrel superfamily. Although in our data set the 2 interfaces were not compared (given their large size difference), the histidine biosynthesis enzyme (1qo2) is clustered with tagatose-1,6-bisphosphate aldolase interface (PDB code: 1gvf), which is composed of chains from the same superfamily (aldolase) and even from the same family (Class II FBP aldolase) as 1b57. The existence of the histidine biosynthesis enzyme and the aldolase protein in the same cluster may support the previous assumption that these 2 proteins share a common evolutionary origin.[53]

Propensities of functional groups in the interfaces

Several works calculated the propensities of all residues in protein interfaces. Jones and Thornton[3] showed that the hydrophobic residues have a greater preference for the interfaces of homodimers than for heterocomplexes. Glaser et al.[54] found hydrophobic residues to be more abundant in large interfaces and polar in small interfaces. These results are in agreement, as homodimeric interfaces are usually larger than heterocomplexes. The results of Zhou and Shan[55] showed that nonpolar residues are favored in the interface, whereas charged and polar residues are disfavored. These results are in agreement with previous results obtained by Argos[56] and with Tsai et al.[7] Neuvirth et al.[57] found that while the hydrophobic content of interfaces in their data set is similar to that of noninterface regions, hydrophobic residues tend to form larger clusters.

Our method, which represents each interface by a set of its functional groups, allows us the detailed examination of the propensities of the functional groups rather than the residue propensities. We compare the frequencies of functional groups in the interfaces to those in the overall structures and on the surface alone. While the first comparison can yield clues to similar forces that participate in the folding and binding processes, the second comparison may be useful for prediction of binding sites. The propensity of each functional group (P_i) was calculated as the fraction of the number of functional groups of type i in the interface, compared with its fraction in the whole chain or on the surface:

$$P_i = \frac{n_i/n_{\text{total}}}{N_i/N_{\text{total}}}, \tag{3}$$

where n_i is the number of functional groups of type i in the interface, n_{total} is the total number of functional groups in the interface, N_i is the number of functional groups of type i in the whole chain or on the surface, and N_{total} is the total number of functional groups in the whole chain or on the surface. The propensities of the following functional groups were calculated for the interfaces in the nonredundant data set clusters: acceptor, donor, aliphatic, aromatic, and sulfur. Although the sulfur property is not represented separately in the original model of Schmitt et al.,[12] here it was added for the propensity calculations. The results are shown in Figure 9. A propensity greater than 1 indicates that the functional group occurs more frequently in the interface than on the protein surface or on the whole chain.

The propensities in Figure 9 show that, compared to the surface alone, the donor, acceptor, and aliphatic functional groups do not show a significant preference to be either in the interface or on the surface (0.96, 0.94, and 1.13, respectively). The aromatic and sulfur functional groups show a very slight preference for the interface (1.27 and 1.21, respectively). However, when comparing the functional groups in the interfaces to the whole chain, the aromatic and aliphatic functional groups show a relatively high preference for the interfaces (1.81 and 1.55, respectively). The donor functional groups show a slight preference to be in the overall chain (0.75), and the acceptor and sulfur functional group show no significant preference for either one (1.01 and 1.13, respectively). When restricting the propensity calculation to large interfaces alone (greater than 2000 Å), very similar results were obtained with minor insignificant changes. The results shown in the figure support the Lo Conte et al. results.[58] In their work, they compared the relative contribution of the nonpolar groups (all groups containing aliphatic and aromatic carbons), neutral polar groups (all groups containing noncarbon atoms, except those carrying a net electric charge), and charged groups to the interface area, as compared to the total solvent-accessible surface area of small proteins. Our propensity results support their finding that the fraction contributed by nonpolar groups (represented as the aliphatic and aromatic functional group in our model) on the interface and in the surface is similar (the average surface of the 75 protein–protein interfaces they examined is 53% nonpolar, while the average interface is 56% nonpolar[58]). Both results suggest that hydrophobic functional groups do not show a greater tendency to be in an interface regions than on the accessible surface of the proteins (although only small globular proteins were examined[58]). Lo Conte et al. found the fraction contributed by neutral polar groups to be somewhat higher in interfaces versus on surfaces, whereas charged groups are somewhat less abundant in interfaces. A comparison of these results with our results is difficult because of the different functional group models used in these two works. Note that, similar to the results of Lo Conte et al., we also detect a large variance in the propensity values (interfaces that are significantly more/less hydrophobic/polar than their overall surfaces are observed). Our results suggest that the composition of the physicochemical groups in the interface is not sufficient for the prediction of binding sites.

Contribution of Backbone Versus Side-Chains

Polar interactions include backbone–backbone, backbone–side-chain, or side-chain–side-chain. In order to compare the involvement of backbone and side-chain interactions in hydrogen bonds, for each polar functional group (hydrogen donor or acceptor) we calculated the percentage of backbone versus side-chain functional groups in the interfaces. The results, presented in Table I, show that on average, the backbone of the polypeptide chain is highly involved in polar interactions: 38% of the donor functional groups and 55% of the acceptor functional groups are derived from the backbone, making the carbonyl oxygen to be the most frequently used acceptor in the interfaces. Although using different functional group models, these results are similar to those of Le Conte et al.[58] There, the fraction of the number of polar groups involved in backbone–backbone interaction was found to be 0.24; backbone–side-chain is 0.4, and side-chain–side-chain is 0.36. As can be seen in Table I, in total, similar percentages of polar backbone functional groups are scattered in the interface and on the surface of the proteins (36% and 39%, respectively).

TABLE I. Percentage of Backbone Versus Side-Chain Polar Functional Groups

Backbone donors in interface	38%
Backbone acceptors in interface	55%
Backbone total polar functional groups in interface	36%
Backbone total polar functional groups in surface	39%

Distribution of Crystal Interfaces in the Data Set

Several works have attempted to develop tools to distinguish between complexes that are derived from monomeric proteins making crystal-packing contact (so-called *crystal interfaces*) from real biological interfaces.[33,40,59,60] A conclusive solution for this problem has not yet been found. We have created a list of crystal interfaces by using experimental observations from the literature and the list of interfaces published as crystal interfaces in Chakrabarti and Janin.[5] The combined list contains a total of 84 interfaces. We used this list in order to examine the distribution of the crystal interfaces in the clusters. The results are shown in Table II. Forty-one interfaces belong to single-member clusters, 18 to clusters with 2 to 4 members, and 25 to larger clusters. The results show that crystal interfaces can be either unique, and therefore be in single-member clusters, or share similar patterns with biological interfaces. However, we note that for the majority of the interfaces in the data set, no literature examination was performed to elucidate their real tertiary structure. Therefore, it is not clear whether some of the interfaces that share the same cluster with crystal interfaces are indeed real interfaces, or perhaps they too are crystal interfaces. For example, a crystal interface, 1jfmBC is in cluster 179, together with 6 other interfaces. Following the classification of a crystal interface in this cluster, we checked the published experimental data of the cluster members and found that at least 3 of the other members (1gzeAD, 1hnnAB, and 1ql3AC) are known to function as monomers. Therefore, the existence of a crystal interface in a cluster could suggest the existence of other unidentified crystal interfaces.

Functional Prediction of New Interfaces

The clusters of interfaces may be useful for predicting the biological function of a new interface. Given a novel complex with an unknown function, we can search the data set for similar interfaces with similar binding organizations. Even if the recognized cluster is one with similar

TABLE II. Distribution of "Crystal Interfaces" in the Data Set

"Crystal interfaces" in a single-member cluster	13pkAD, 1i4gAB, 1aw7AB, 1ml1AC, 2g3pAB, 1mblAB, 830cAB, 1a3yAB, 1c02AB, 1aq0AB, 2hckAB, 1qciAB, 1xgmAB, 1qdmAB, 1ag9AB, 1hkbAB, 1ckiAB, 1hfvAB, 1g4eAB, 1fgkAB, 4atjAB, 1cqxAB, 1qpaAB, 1toaAB, 1dsuAB, 1a12AB, 1ac1AB, 1qs8AB, 1omeAB, 1bilAB, 1sw6AB, 1f05AB, 1d8hAC, 1fvfAB, 1b6bAB, 1brwAB, 1fqjBE, 1flzAB, 1hz6BC, 1k05BC, 1jnrAD
"Crystal interfaces" in clusters with 2–4 members	1bs4BC, 1ehyAC, 1cpjAB, 2ugiAB, 1binAB, 1c47AB, 1xcaAB, 1dxmAB, 1pvaAB, 1a2lAB, 1csgAB, 1b5zAB, 1cbiAB, 1dkdAC, 1fguAB, 1gcqBC, 2rslAB, 1ifqAB
"Crystal interfaces" in clusters with at least 5 members	1rb3AB, 1ilr12, 1thtAB, 1kptAB, 1bc2AB, 1bkzAB, 1vlzAB, 3ng1AB, 1ae9AB, 1aohAB, 1k6jAB, 1e69AB, 1a7vAB, 1dxxAD, 1czvAB, 1iepAB, 1dg1GH, 1lvfAB, 1fdqAB, 1gggAB, 1fvrAB, 1i0cAB, 1iazAB, 1jfmBC, 1rhgBC

interfaces that are constructed of chains of different folds, it can still be used to obtain hints for the function of the new interface. We used the Gene Ontology (GO) database[61] to examine whether such clusters contain recurring biological features. The GO database contains 3 hierarchies that represent (1) the molecular function of a protein, (2) the biological process that the protein participates in, and (3) the cellular component it functions in. The different levels of the hierarchies allow the examination of not only the specific type of function or process but also a more general view of the role of the protein (e.g., a hierarchal annotation of a nucleic acid binding protein may include RNA binding annotation, DNA binding annotation or translation factor activity, and each of these goes further down in the hierarchy of annotations). Clusters of the first type (similar interface–similar global folds) include interfaces that belong to the same family and thus will have similar annotations even at high levels of hierarchy. On the other hand, clusters from the second type (similar interfaces–different global folds) may contain interfaces with different high-level annotations but similar low-level annotations. These clusters may be useful in the global functional prediction of newly determined interfaces that are clustered within them. Several clusters have been found where half or more of the members have different specific functions but they share similar low-level annotation of function or pathway. Three of these clusters are presented in Table III. For the third cluster in Table III, 2 of its members (1nlxAB and 1hv2AB) do not have any GO annotations. However, Batuyan et al.[62] published the fact that the 1hv2 protein is a transcription elongation factor, which binds the DNA as the rest of its cluster members. Obviously, functional annotation becomes more complex when all members of the cluster exhibit different functions. However, the above examples illustrate that even if the new interface is assigned to a cluster that contains complexes with different functions, the assignment to this cluster may still contribute to the functional prediction.

CONCLUSIONS

Here, we have presented a nonredundant data set of protein–protein interfaces. The clustering is based on the similarity between both the spatial and the physicochemical properties of the backbone and the side-chain atoms, regardless of the sequence or overall structural similarity of the chains. As expected, compared to the C_α residue–based data set of interfaces, the different representation, thresholds, parameters, and clustering methods affect the final results. The total number of clusters with at least 5 elements is smaller in the current data set (59 clusters vs 103 clusters). The number of interfaces that these clusters contain is also smaller (604 in our data set vs 949 in the residue-based data set). This highlights the differences in the representations of the interfaces and the data set generation. The I2I-SiteEngine alignment focuses on the chemical groups and the geometries of the aligned surfaces, whereas the C_α-based data set focuses on backbone interface motifs. There, the side-chain atoms of the interacting residues may display dissimilar orientations and chemical properties.

The choice of which data set to use depends on the goal. Using a functional group representation leads to higher accuracy. Thus, inevitably, the functional group–based data set has fewer clusters with at least 5 members, reflecting its sensitivity to side-chain and backbone functional group motions. Given its representation and sensitivity, it is not expected to behave well in large-scale applications of lower resolution structures and modeled structures. However, it is expected to be the data set of choice when elucidating the stabilizing interactions on the microscopic scale and the preferred constellation of chemical groups, when searching for protein–protein binding sites with similar chemistry, when designing mutations or new binding sites, in functional assignment and in drug design.

The SCOP database classifies proteins into structural families (when there is sequence similarity), structural superfamilies (when the proteins lack sequence similarity but have functional and structural similarity), and structural folds (when there is only coarse structural similarity). When cross-comparing the data set clusters with the SCOP classification of the overall fold, 2 types of clusters are revealed. The first, more trivial, are clusters that contain interfaces composed of chains classified in the same superfamily. The second cluster type contains interfaces composed of chains from different SCOP superfamilies. We have shown several examples of such clustered proteins, most of which perform different functions. Such clusters probably illustrate favorable chemical organization of interfaces.

TABLE III. Three Clusters With Members That Share Similar Low-Level Annotation of Function or Process

Cluster members	Specific function (obtained by SCOP)	Shared low-level function/process (obtained by GO)
113bAD	Precorrin-6Y methyltransferase	Transferase activity
110oAB	Anti-sigma factor spoIIab	Transferase activity
1b99AD	Nucleoside diphosphate kinases	Transferase activity
1e7pAD	Fumarate reductase	—
1gttBC	4-hydroxyphenylacetate degradation bifunctional isomerase	—
1iunAB	Product hydrolase CumD	—
1a3gAB	Branched-chain amino acid aminotransferase	Transferase activity
1e4yAB	Adenylate kinase	Transferase activity
1f80AE	Acyl carrier protein	Transferase activity
1ir2BG	Ribulose 1,5-bisphosphate carboxylase-oxygenase	—
1d5wBC	Transcriptional regulatory protein FixJ	DNA binding
1nlxAB	Pollen allergen	(No annotations)
1ihwAB	Retroviral integrase	DNA binding
1petAD	p53 tetramerization domain	DNA binding
1saeAB	p53 tetramerization domain	DNA binding
1hv2AB	Elongin C	(No annotations)
1je8AB	Nitrate/nitrite response regulator	DNA binding
1ng7AB	Poliovirus core protein 3a	RNA binding
1q0wAB	(NO SCOP ENTRY)	RNA binding
1otrAB	Protein Cue2 and Ubiquitin	Damaged DNA binding
1pmxAB	(NO SCOP ENTRY)	—
1h59AB	Insulinlike growth factor and IGFBP-5	—
1jyoEF	Virulence effector SptP domain	—
1pmpAB	P2 myelin protein	—

The mark (—) implies that the recurring annotation does not appear in this complex.

When the structure of a new complex is determined, it can serve as a query and can be compared by methods like I2I-SiteEngine to the representative of each cluster. In case a similarity has been found, the annotations of the representative cluster members can be examined. Although an accurate functional prediction is not guaranteed by such a method, one can still obtain initial leads that can be examined further. The current classification uses the I2I-SiteEngine method, which was designed to specifically suit the requirements of interface matching. However, it has several weaknesses. First, it does not implicitly address protein flexibility. Second, electrostatic potentials are not taken into account; therefore, electrostatic interactions are not properly considered. In addition, the method uses very high resolution of molecular representation, which in many cases may be too sensitive. In such cases, a lower resolution representation may be sufficient and may lead to larger clusters of similar interfaces.

Currently, a method for multiple alignment of functional groups in binding sites is being developed. Such an algorithm should be able to detect physico-chemical patterns common to a set of binding sites. Moreover, we intend to develop a method that will perform a simultaneous multiple alignment of protein–protein interfaces. Such a method is expected to allow us an examination of the structurally conserved interacting functional groups in each cluster. It may assist in addressing questions: What are the interactions that stabilize different types of complexes? Are the conserved functional groups scattered on the interface surface or clustered to form a network of interactions? Do conserved functional groups arise only from similar residues? Hopefully, the combination of this data set and the multiple alignments of the chemical features within the clusters would prove useful toward these applications.

ACKNOWLEDGMENTS

We thank Dina Schneidman-Duhovny, Maxim Shatsky, and Oranit Dror for many useful suggestions and for contribution of software to this project.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Chothia C, Janin J. Principles of protein–protein recognition. Nature 1975;256:705–708.
3. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
4. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53:708–719.
5. Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins 2002;47:334–343.
6. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Protein–protein interfaces: architectures and interactions in protein–protein interfaces and in protein cores: their similarities and differences. Crit Rev Biochem Mol Biol 1996;31:127–152.
7. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.
8. Xu D, Tsai CJ, Nussinov R. Mechanism and evolution of protein dimerization. Protein Sci 1998;7:533–544.
9. Valdar WS, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
10. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A data set of protein–protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 1996;260:604–620.
11. Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. Protein Sci 2004;13:1043–1055.
12. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
13. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.

14. Ullmann JR. An algorithm for subgraph isomorphism. J Assoc Comput Mach 1976;23:31–42.
15. Spriggs RV, Artymiuk PJ, Willett P. Searching for patterns of amino acids in 3D protein structures. J Chem Inf Comput Sci 2003;43:412–421.
16. Rosen M, Lin SL, Wolfson H, Nussinov R. Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng 1998;11:263–277.
17. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
18. Kinoshita K, Furui J, Nakamura H. Identification of protein functions from a molecular surface database, eF-site. J Struct Funct Genomics 2002;2:9–22.
19. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12:1589–1595.
20. Bron C, Kerbosch J. Finding all cliques of an undirected graph. Commun ACM 1973;16:575–577.
21. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. Protein Sci 1997;6:2308–2323.
22. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 1996;5:1001–1013.
23. Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 2003;19:1644–1649.
24. Brakoulias A, Jackson RM. Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins 2004;56:250–260.
25. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
26. Jones S, Thornton JM. Searching for functional sites in protein structures. Curr Opin Chem Biol 2004;8:3–7.
27. Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-Completeness. New York: W. H. Freeman; 1979.
28. Cormen TH, Leiserson CE, Rivest RL. Introduction to algorithms. Cambridge, MA: MIT Press; 1990.
29. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003;31:400–402.
30. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3:265–274.
31. Li H, Li J. Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction data sets. Bioinformatics 2005;21:314–324.
32. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88:10495–10499.
33. Mintseris J, Weng Z. Atomic contact vectors in protein–protein recognition. Proteins 2003;53:629–639.
34. Shulman-Peleg A, Mintz S, Nussinov R, Wolfson HJ. Protein–protein interfaces: Recognition of similar spatial and chemical organizations. In Jonassen I and Kim J, editors. Workshop on Algorithms in Bioinformatics 2004, Bergen, Norway. Lecture Notes in Computer Science, Springer Verlag, Volume 3240. p 194–205.
35. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
36. Clackson T, Wells JA. A hot spot of binding energy in a hormone–receptor interface. Science 1995;267:383–386.
37. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 2002;12:14–20.
38. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
39. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
40. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
41. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
42. Dror O, Benyamini H, Nussinov R, Wolfson H. MASS: multiple structural alignment by secondary structures. Bioinformatics 2003;19(Suppl 1):i95–i104.
43. Dror O, Benyamini H, Nussinov R, Wolfson HJ. Multiple structural alignment by secondary structures: algorithm and applications. Protein Sci 2003;12:2492–2507.
44. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
45. Connolly ML. Molecular surface calculation. J Appl Crystallogr 1983;16:548–558.
46. Mehlhorn K. The LEDA platform of combinatorial and geometric computing. London: Cambridge University Press; 1999.
47. Richards FM. Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 1977;6:151–176.
48. Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley; 2001.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
50. Liu Y, Xu L, Opalka N, Kappler J, Shu HB, Zhang G. Crystal structure of sTALL-1 reveals a virus-like assembly of TNF family ligands. Cell 2002;108:383–394.
51. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein–protein interfaces. J Mol Biol 2004;336:943–955.
52. Fischer D, Wolfson H, Lin SL, Nussinov R. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. Protein Sci 1994;3:769–778.
53. Sandhya S, Kishore S, Sowdhamini R, Srinivasan N. Effective detection of remote homologues by searching in sequence data set of a protein domain fold. FEBS Lett 2003;552:225–230.
54. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
55. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
56. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng 1988;2:101–113.
57. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 2004;338:181–199.
58. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
59. Elcock AH, McCammon JA. Identification of protein oligomerization states by analysis of interface conservation. Proc Natl Acad Sci USA 2001;98:2990–2994.
60. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
61. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32[Database issue]:D258–D261.
62. Botuyan MV, Mer G, Yi GS, Koth CM, Case DA, Edwards AM, Chazin WJ, Arrowsmith CH. Solution structure and dynamics of yeast elongin C in complex with a von Hippel–Lindau peptide. J Mol Biol 2001;312:177–186.
63. Shatsky M, Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of binding patterns common to a set of protein structures. In Miyano S, editor. RECOMB 2005, Cambridge, MA, May 14–18, 2005. Lecture Notes in Computer Science, Springer Verlag, Volume 3500. p 440–455.