UNCOVERING THE STRUCTURAL BASIS OF PROTEIN INTERACTIONS WITH EFFICIENT CLUSTERING OF 3-D INTERACTION INTERFACES

Z. Aung*

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 and
School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543
**Email: azeyar@i2r.a-star.edu.sg*

S.-H. Tan

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613,
Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada[†]*, and*
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada[†]
Email: chris.tan@utoronto.ca

S.-K. Ng

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Email: skng@i2r.a-star.edu.sg

K.-L. Tan

School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543
Email: tankl@comp.nus.edu.sg

The biological mechanisms with which proteins interact with one another are best revealed by studying the structural interfaces between interacting proteins. Protein–protein interfaces can be extracted from 3-D structural data of protein complexes and then clustered to derive biological insights. However, conventional protein interface clustering methods lack computational scalability and statistical support. In this work, we present a new method named "PP*i*Clust" to systematically encode, cluster and analyze similar 3-D interface patterns in protein complexes efficiently. Experimental results showed that our method is effective in discovering visually consistent and statistically significant clusters of interfaces, and at the same time sufficiently time-efficient to be performed on a single computer. The interface clusters are also useful for uncovering the structural basis of protein interactions. Analysis of the resulting interface clusters revealed groups of structurally diverse proteins having similar interface patterns. We also found, in some of the interface clusters, the presence of well-known linear binding motifs which were non-contiguous in the primary sequences. These results suggest that PP*i*Clust can discover not only statistically significant but also biologically significant protein interface clusters from protein complex structural data.

1. INTRODUCTION

Proteins and their molecular interactions with one another are essential for many different biological activities in the cell. Unlike the DNA, a protein is composed of a sequence of amino acid (AA) residues folded into a three-dimensional (3-D) form. It is widely-understood that the 3-D structure of a protein, rather than its AA sequence, is the key determinant of its biological function.

A substantial amount of research work on understanding the mechanisms of protein-protein interactions (PPIs) from the primary sequences of proteins has already been reported—for example, see Ref. 1, 2. In comparison, there has been relatively limited amount of work done based on 3-D structures. In this paper, we will study the interactions between proteins in terms of their 3-D protein–protein interfaces. These are regions in 3-D protein complexes that consist of interacting residues belonging to two different chains that are in spatial vicinity. The interface residues have been known to be highly conserved.[2] Identifying and understanding the underlying mechanisms of these interface clusters can lead to important biological insights that can be useful for appli-

[*] Corresponding author.
[†] Present affiliations.

clusters generated by our algorithm are also statistically significant (in addition to the conventional visual and biological verifications).

(2) All the existing methods employed time-consuming comparison techniques to measure the similarity of the interfaces, resulting in unscalable approaches. For instance, I2I-SiteEngine[8] took an average of 26 seconds for each pairwise interface comparison, and required a total of $5,574,861$ such comparisons in its entire clustering process.[8] This means it will require about $1,677$ days (over 4 years) to carry out the clustering process on a single PC. (Actually, I2I-SiteEngine was implemented on a cluster of PC workstations, and it took about 1 month processing time.[12]) Here, by using a novel algorithmic scheme, we are able to perform interaction interface clustering in a much more time-efficient manner while at the same time maintaining the statistical quality of the clusters generated.

3. DATA REPRESENTATION

Many biological processes in the cell involve the formation of *protein complexes* which are molecular aggregations of numerous proteins that are in stable protein–protein interactions. The interacting proteins can be collectively crystallized and their 3-D structure determined as a single group. Such structural information are usually deposited as a single entity into PDB[10] database and given a unique PDB ID. The member proteins of a protein complex are called *protein chains* or simply *chains*. Within a particular complex, each chain is assigned a unique chain ID. A pair of protein chains that are directly interacting with each other form an *interface* region through which they interact spatially. A residue from a protein chain is considered to be a part of an interface if it has at least one counterpart residue from the other chain with the distance between their nearest atoms less than or equal to 5Å.[6]

For example, in Figure 1, the protein complex *gamma delta resolvase* is designated with the PDB ID 2rsl. It has three protein chains which are assigned the chain IDs A, B, and C. In this complex, there are direct interactions between chains A and B, and also between chains B and C respectively. The interface for each interacting protein pair is highlighted in the figure. The interface for chains A and B is denoted as 2rslAB, and that for chains B and C as 2rslBC.[6]

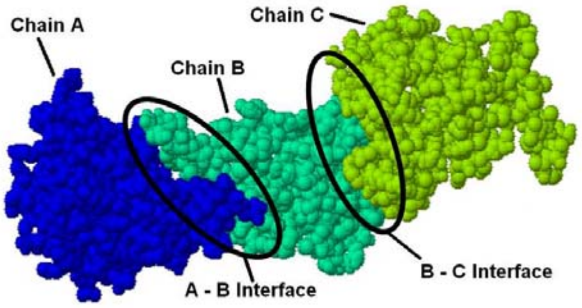


Fig. 1. The protein complex *gamma delta resolvase* (PDB ID 2rsl) with three protein chains A, B and C.

The residues that constitute an interface are not always sequential in nature, according to the observations in Ref. 6. It is therefore inadequate to represent the interfaces by the literal sequences of the constituent residues from the N-terminus to the C-terminus of the chains. Furthermore, to overcome the weakness of those methods[6, 13] that handled the two interface fragments separately, we need to find a better way to encode the interfaces as a single entity such that processing it is equivalent to processing its two constituent interface fragments simultaneously.

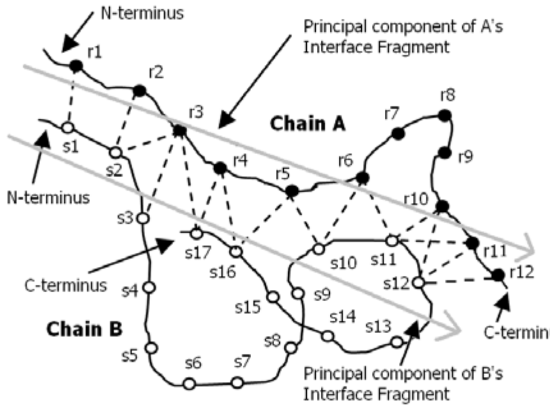


Fig. 2. An example protein complex with chains A and B. The dotted line means that the two residues are in contact (i.e. distance of their nearest atoms ≤ 5Å).

To do so, we encode interfaces as *interface ma-*

trices as follows. For the two interface fragments in an interface, we first derive their respective principal component vectors by means of principal component analysis.[14] We then arrange the residues in each interface fragment by their positions along its principal component vector as shown in Figure 2. The interface fragment for chain A is an ordered set of 9 residues: $\{r_1, r_2, r_3, r_4, r_5, r_6, r_{10}, r_{11}, r_{12}\}$, whereas that for chain B is an ordered set of 8 residues: $\{s_1, s_2, s_3, s_{17}, s_{16}, s_{10}, s_{11}, s_{12}\}$.

An interface matrix encoding of each protein interface can then be obtained by storing the pairwise distances between the centers of residues, each from an interface fragment, in a matrix that effectively captures the "interface pattern" of the interface fragments. The interface matrix for the interacting proteins chains A and B in the above example is a 9×8 matrix:

$$\begin{pmatrix} d(r_1,s_1) & d(r_1,s_2) & d(r_1,s_3) & d(r_1,s_{17}) & \dots & d(r_1,s_{12}) \\ d(r_2,s_1) & d(r_2,s_2) & d(r_2,s_3) & d(r_2,s_{17}) & \dots & d(r_2,s_{12}) \\ d(r_3,s_1) & d(r_3,s_2) & d(r_3,s_3) & d(r_3,s_{17}) & \dots & d(r_3,s_{12}) \\ d(r_4,s_1) & d(r_4,s_2) & d(r_4,s_3) & d(r_4,s_{17}) & \dots & d(r_4,s_{12}) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d(r_{12},s_1) & d(r_{12},s_2) & d(r_{12},s_3) & d(r_{12},s_{17}) & \dots & d(r_{12},s_{12}) \end{pmatrix}$$

where $d(\bullet, \bullet)$ is the Euclidean spatial distance between the centers of two given residues.

4. CLUSTERING METHOD

There are four major steps in our proposed PP*i*Clust method for discovering the significant clusters of similar protein–protein interfaces:

(1) *Extracting representative interfaces.* First, we extract representative interfaces from the 3-D protein complexes in PDB and encode them as interface matrices;
(2) *Generating interface feature vectors.* We then generate feature vectors for the representative interface matrices extracted;
(3) *Clustering.* Clustering is then performed on the interface feature vectors to discover groupings of the protein interfaces; and
(4) *Statistical validation.* Finally, we quantitatively ascertain the statistical quality of the interface clusters generated.

4.1. Extracting Representative Interfaces

First, we extracted a set of representative protein–protein interfaces from the protein complexes. We used the 3-D structural data of protein complexes from PDB. After removing the irrelevant structures, such as single chains, low-resolution models, etc., we obtained a data set of 17,300 protein chains which belonged to 5,503 protein complexes.

We then extracted protein–protein interfaces from the interacting protein pairs of the protein complexes. From 5,503 complexes, we obtained 17,012 interfaces. After pruning away the interfaces with too few (less than 10) or too many (more than 200) interacting residues in each side, 11,558 interfaces were left.

Some of these interfaces may be redundant. Two interfaces are considered redundant if both of their corresponding chains are sequentially homologous (with more than 30% sequence identity using BLASTClust, which is a part of BLAST[15] suite). Using this criterion, we identified groups of redundant interfaces. For each such group, we chose the one with the best resolution and the largest interface size as the *representative interface.* After this process, we ended up with 1,445 representative interfaces for further analysis. The interfaces were then encoded into interface matrices as described in Section 3.

4.2. Feature Vector Generation

Our objective is to group similar interface matrices into their respective clusters. To do so, we need to be able to compare the interface matrices and determine their similarity values quantitatively.

The DALI method[16] was previously used to align 2-D distance matrices derived from individual 3-D protein structures. Unfortunately, we cannot employ DALI here because it is known to be a time consuming pairwise alignment method, and it will take a very long time (several months on a stand-alone PC) to align 1,445 interface matrices all-against-all for our systematic analysis.

As such, we devise a new scheme for encoding the interface matrices so that they can be compared efficiently and effectively. We opt for a scheme where we represent each interface matrix as a multi-dimensional feature vector based on the frequencies of the "local features" exhibited in the interface matrix. Such a frequency-based approach has been extensively used in various histogram methods in image processing.[17] It has also been used in structural bioinformatics, particularly for protein

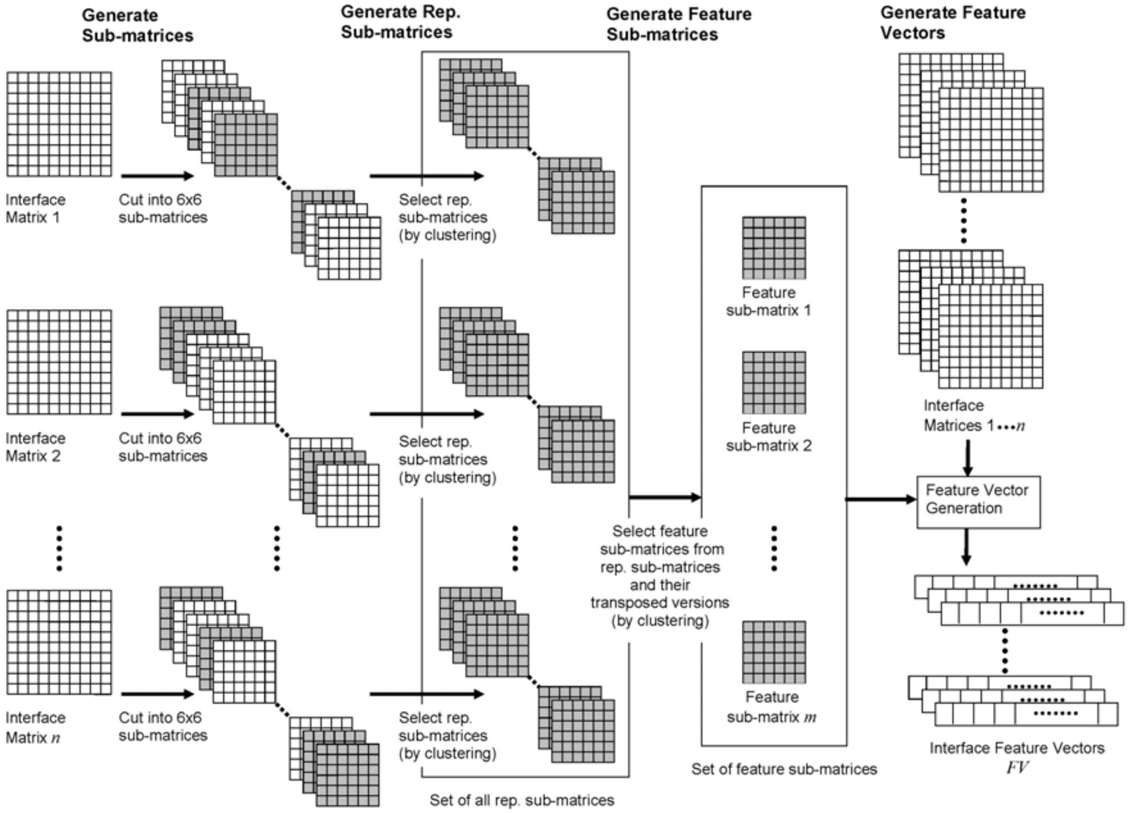

Generate Sub-matrices
Generate Rep. Sub-matrices
Generate Feature Sub-matrices
Generate Feature Vectors
Interface Matrix 1
Cut into 6x6 sub-matrices
Select rep. sub-matrices (by clustering)
Interface Matrix 2
Cut into 6x6 sub-matrices
Select rep. sub-matrices (by clustering)
Interface Matrix *n*
Cut into 6x6 sub-matrices
Select rep. sub-matrices (by clustering)
Set of all rep. sub-matrices
Select feature sub-matrices from rep. sub-matrices and their transposed versions (by clustering)
Feature sub-matrix 1
Feature sub-matrix 2
Feature sub-matrix *m*
Set of feature sub-matrices
Interface Matrices 1•••*n*
Feature Vector Generation
Interface Feature Vectors *FV*

Fig. 3. Generating feature vectors from representative interface matrices. Representative sub-matrices for each representative interface matrix are shown in gray.

fold classification.[18, 19]

We can view an interface matrix as a set of 6×6 overlapping sub-matrices.[16, 20] Our basic idea is to represent an interface matrix as a "bit-vector" where each bit corresponds to the presence or absence of a single type of sub-matrix which constitute the whole interface matrix. However, there are over one million distinct sub-matrices for all $1,445$ interface matrices. If we use them all, our resultant bit-vector will be too long. To reduce the number of sub-matrix types, we group the similar ones together and select a representative sub-matrix from each group. This is done by two rounds of nearest-neighbor clustering[21] and medoid selection processes. The process of generating feature vectors from the representative interface matrices is outlined in Figure 3. Using this method, we finally came up with 409 features sub-matrices.

Using the 409 feature sub-matrices, we can now systematically encode each interface matrix as a *feature vector*. Basically, it is the frequency profile of the sub-matrix features in the interface matrix, with the dimension of the frequency vector being equal to the number of feature sub-matrices.

4.3. Clustering

What we have done in the previous steps is to reduce the 3-D structural information of protein interfaces into 2-D interface matrix and then into 1-D feature vectors. We are now ready to cluster the extracted protein interfaces. For any two feature vectors FV_i and FV_j, we measure their *feature vector distance* with the inverse cosine distance function[17] defined as:

$$df(FV_i, FV_j) = \cos^{-1} \frac{(FV_i \cdot FV_j)}{(\| FV_i \| \cdot \| FV_j \|)} \qquad (1)$$

where $(\bullet \cdot \bullet)$ is the dot product between two vectors, and $(\| \bullet \|)$ is the norm of a vector. While $df(\bullet, \bullet)$ is a non-metric distance (it violates the triangular inequality property), it is well-suited for reflecting the human's perceptions of similarity and non-similarity.[17] Note that we have also tested our system with the metric Euclidean distance function, but it confirmed that the inverse cosine distance function is indeed superior.

In this study, we discover the clusters of interface feature vectors for $1,445$ interfaces by employing the nearest-neighbor clustering algorithm[21] using the distance function $df(\bullet, \bullet)$ and the distance threshold df_t. (We will discuss the effects of different df_t values in the next sub-section.)

In the clustering process, for every object (feature vectors in our case) in the input data set, we allocate it to the cluster in which its nearest neighbor exists and all the other existing cluster members are also near enough to it, with regard to the given distance threshold (df_t in our case). If we cannot detect such a cluster, we create a new cluster with this object as the first member.

4.4. Statistical Validation

During the clustering, we also conducted a statistical test called *silhouette analysis*[22] to quantitatively ascertain the quality of the interface clusters that were discovered. The *silhouette width* $s(i)$ of an object i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(\, a(i), b(i)\,)} \qquad (2)$$

where $a(i)$ is the average distance of i to all other objects in its own cluster, and $b(i)$ is the average distance of i to all objects in its nearest neighbor cluster. The silhouette width of an object is between -1 (the worst case) and $+1$ (the ideal case). The *average silhouette width* $\overline{s}$ of a clustering scheme is the average of the silhouette widths of all the members in all clusters. The larger the value of $\overline{s}$, the better the clustering scheme.

Figure 4 shows the effect of varying distance threshold df_t values on the average silhouette width ($\overline{s}$) of the clustering scheme. The lower df_t values give higher quality clusters (i.e. in terms of high $\overline{s}$ values) than those generated with larger df_t values. However, the lower df_t values also generated many more useless singleton interface clusters. As such, we have to make a decision based on the tradeoff between the $\overline{s}$ value and the coverage of non-singleton clusters. We use here a criterion of having at least half of the total number of interfaces covered by non-singleton clusters. This can be attained by setting $df_t = 0.35$, which covered 50.6% of the interfaces. This corresponds to an $\overline{s}$ value of 0.85 if all the clusters are taken into account, and a value of 0.58 if we consider only the non-singleton clusters. As explained in Ref. 23, a clustering is considered reasonably good if its $\overline{s}$ is between 0.51 and 0.7. In this way, our method ensures the statistical quality of the interface clusters generated.

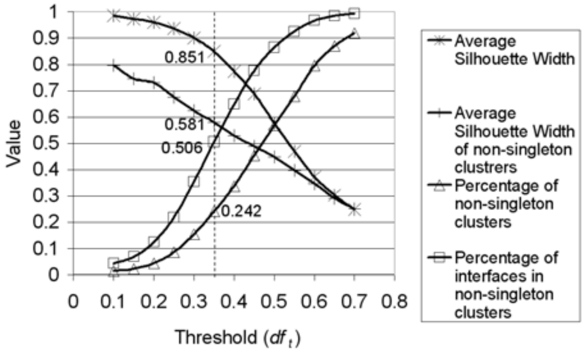


Fig. 4. Effect of feature vector distance threshold (df_t) on clustering.

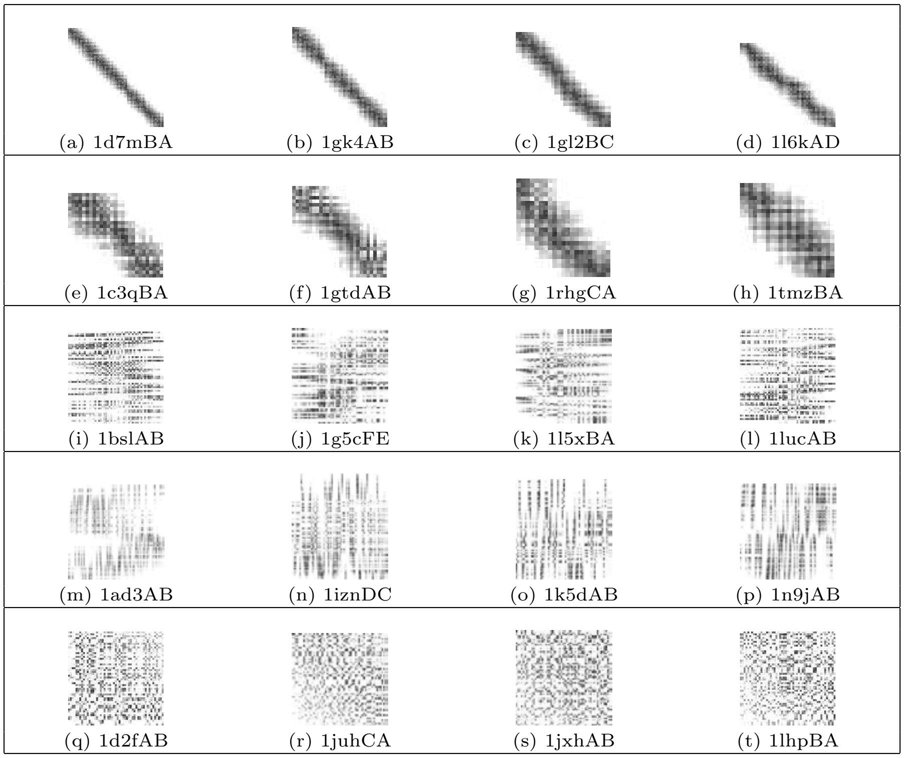
5. RESULTS

We implemented PPiClust on a stand-alone PC with 2 Pentium IV 3.0GHz CPUs and 1GB main memory. The resultant clusters are presented in the webpage: `http://www1.i2r.a-star.edu.sg/~azeyar/genesis/PPiClust/`.

In our current study on $1,445$ representative interfaces from $5,503$ protein complexes, our method was time-efficient with a total processing time of only about 8 hours. This is much faster compared to the other interface clustering methods such as Ref. 7, 8, 6, which are too slow to be practically implemented on a single PC.

5.1. Visual Verification

As a preliminary analysis, we inspected the visual quality of the interface clusters. Figure 5 shows the various sample interfaces observed in some of the clusters. The interfaces were represented as interface matrices, depicted in the figure as gray-scale images. Darker tones indicate closer residue–residue distances in an interface, while the lighter tones depict the larger distances. It is observed that the interfaces belonging to a same cluster generally look similar. For example, interface patterns (a)–(d) belong to a particular cluster with the characteristic



(a) 1d7mBA (b) 1gk4AB (c) 1gl2BC (d) 1l6kAD

(e) 1c3qBA (f) 1gtdAB (g) 1rhgCA (h) 1tmzBA

(i) 1bslAB (j) 1g5cFE (k) 1l5xBA (l) 1lucAB

(m) 1ad3AB (n) 1iznDC (o) 1k5dAB (p) 1n9jAB

(q) 1d2fAB (r) 1juhCA (s) 1jxhAB (t) 1lhpBA

Fig. 5. Examples of similar interface patterns (represented as interface matrices) belonging to various interface clusters: (a)–(d) thin diagonals, (e)–(h) thick diagonals, (i)–(l) horizontal ripples, (m)–(p) vertical ripples, and (q)–(t) sparse patterns.

appearance of a thin diagonal, interface patterns (e)–(h) belong to another cluster with the common appearance of a thick diagonal, etc.

Next, we further analyze our resulting protein interface clusters of protein interfaces to see whether our method can generate not only statistically significant protein interface clusters but also biologically interesting ones. In particular, we investigated whether the clusters contained non-trivial discoveries such as similar interface patterns from structurally diverse proteins, as well as whether the well-known linear binding motifs were also found in the resulting protein interface clusters.

5.2. Structural Diversity of Interfaces' Parent Chains

Despite the built-in statistical assurance in PPiClust, it is still plausible that the resultant clusters contained protein interfaces whose parent protein chains are all structurally similar. Discovering such interface clusters would not be very biologically significant, since the interacting interface from structurally similar parent chains are expected to be clustered together. What would be more interesting would be the discovery of clusters that contained similar interfaces whose parent chains are structurally quite different. We have found a surprisingly large number of such interfaces in the clusters of interaction interfaces using our method.

Let us systematically determine if the interface clusters generated by our method contained mostly interfaces from structurally diverse parental chains. We can measure the diversity of a given interface cluster C with its Fold pair-based *Shannon's entropy value*[24] as follows:

$$Ent(C) = \sum_{i=1}^{k} -p_i \times \log_2 p_i \qquad (3)$$

where k is the total number of distinct parent Fold pairs that the interfaces in cluster C belongs to, and p_i is the proportion of C belonging to a particular Fold pair i.

In the case when a cluster is totally homogeneous (i.e. all interfaces in the cluster belongs to a single Fold pair), its entropy value will be 0. On the other hand, if a cluster is totally diverse (i.e. each member interface belongs to a distinct Fold pair from the others), its entropy value will be $\log_2 n$, where n is the number of members in the cluster.

Figure 6 shows the average entropy values for the different cluster sizes. We also show two reference curves for the ideal (zero) and the maximum entropy ($\log_2 n$) cases in the figure. Observe that the entropy values for the interface clusters found by our method are indeed generally close to the maximum values. Thus, we can infer that our methods have detected mostly biologically interesting clusters of structurally similar interfaces belonging to the structurally diverse parent proteins.

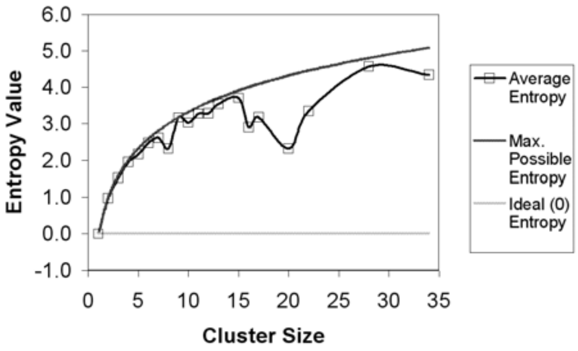


Fig. 6. Average entropies for different cluster sizes.

Overall, the average entropy of each cluster is 1.37. The result indicates that similar interfaces are indeed mediating interactions among diverse structural folds. These interface clusters represent favorable binding structural scaffolds that have been reused in nature for interactions. They are thus useful for understanding the underlying structural basis for proteins to interact with each other (such as identifying putative binding sites on proteins of known structures).[25] The interfaces could also facilitate studies on the critical residues[4] and the motifs[26] important for the stability of protein–protein interactions.

Figure 7 shows the interface 1kacAB of protein complex 1kac (*The λ Repressor C-Terminal Domain Octamer*) and 1mbxCA of protein complex 1mbx (*ClpSN with Transition Metal Ion Bound*). According to SCOP[27] structural classification system, 1kacA belongs to *Fold* b.21, 1kacB to Fold b.1, 1mbxC to Fold d.45, and 1mbxA to Fold a.174. In other words, 1kacAB belongs to the parent *Fold pair* b.21–b.1, and 1mbxCA belongs to the parent Fold pair d.45–a.174. Thus, while the interface structures of 1kacAB and 1mbxCA are quite similar, their parent chain structures are very different. This finding enables us to further investigate the possible functional similarity of 1kacAB and 1mbxCA, even though their global structures bear no significant resemblance to each other. In fact, as we will discuss in the next section, we actually found an important linear motif **KPxx[QK]** (ELM ID: LIG_SH3_4) commonly embedded in both of them.

5.3. Occurrences of Important Biological Motifs

We also observed that the discovered interfaces tend to be compact—each interface fragment contains an average of 30.81 residues. This has biological significance as it implies that the provision of a large complementary surface between two structures is not an essential prerequisite for interactions. In fact, it is likely that the interactions are mediated by short residue fragments or motifs on these compact interfaces.

Biologists have recently discovered that there are small contiguous sequence segments of 3–10 residues that play critical roles in many protein interactions, post-translational modifications and trafficking.[2] In fact, it is estimated that 15%–40% of interactions may be mediated by a short, linear motif (expressed commonly in regular expression) in one of the binding partners.[28] To further assess the biological significance of the clusters derived, we also attempt to identify linear binding motifs[2] from our clusters.

For each cluster generated by our method, we derive two sets of interface residue sequences that are sequential in 3-D space after the principal component analysis transformation. Note that these interface residues may not be contiguous in terms of their primary sequences. To detect whether occurrences of important biological motifs can be found in our interface clusters, we attempt to match a set of linear binding motifs extracted from biomedical literature and ELM database[29] to the interface sequences derived above. The most significant matches are listed in Table 1. For example, the common **AxxxA**[30] helix-helix interaction motif (where **x** denote any AA) were repeatedly detected in our derived sequences. In particular, the popular **PxxP** binding motif [31] in various signaling pathways were also detected in one of our clusters.

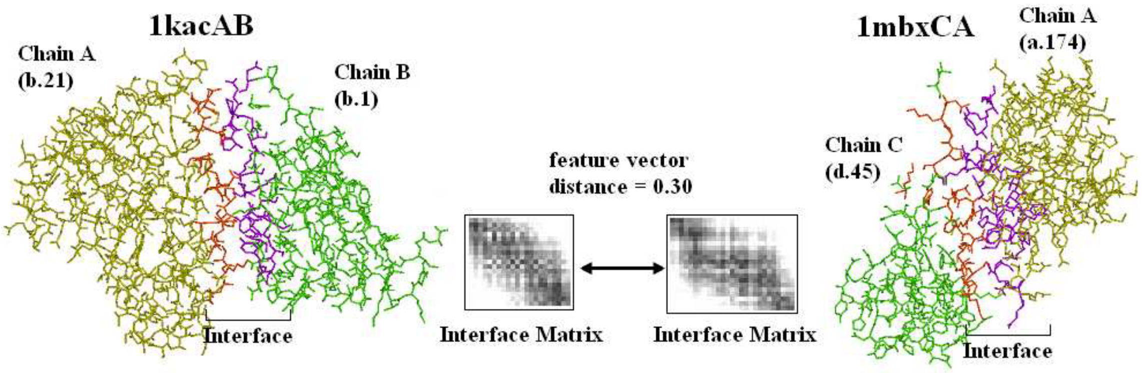


Fig. 7. Similar interfaces in different protein complexes.

Table 1. Significant matches between known linear binding motifs and clusters of interface sequences. Motifs are expressed as regular expression where "x" represent any AA. For matched interface sequences, the chain ID and the corresponding AA numberings are given. The odd-ratio is calculated as O/E where O is observed occurrence of linear motif in the cluster, and E is occurrence of linear motif expected by random in the cluster.

Linear Binding Motif	Matched Interface Sequences	Odd-Ratio	References
KPxx[QK]	1kacA $\mathbf{K}_{429}\mathbf{P}_{418}\mathbf{P}_{417}\mathbf{P}_{416}\mathbf{Q}_{487}$ 1mbxC $\mathbf{K}_{23}\mathbf{P}_{24}\mathbf{P}_{25}\mathbf{S}_{26}\mathbf{K}_{105}$	150.76	LIG_SH3_4 (Ref. 29)
RxLx[EQ]	1n7sA $\mathbf{R}_{56}\mathbf{K}_{59}\mathbf{L}_{60}\mathbf{L}_{63}\mathbf{E}_{62}$ 1hdhA $\mathbf{R}_{390}\mathbf{A}_{38}\mathbf{L}_{394}\mathbf{I}_{37}\mathbf{Q}_{397}$	66.70	(Ref. 32)
RGD	1bslA $\mathbf{R}_{115}\mathbf{G}_{50}\mathbf{D}_{18}$ 1bouA $\mathbf{R}_{127}\mathbf{G}_{126}\mathbf{D}_{19}$	64.94	LIG_RGD (Ref. 29)
L[IVLMF]x[IVLMF][DE]	1lm8V $\mathbf{L}_{178}\mathbf{I}_{180}\mathbf{S}_{183}\mathbf{L}_{184}\mathbf{D}_{187}$ 1b79A $\mathbf{L}_{96}\mathbf{L}_{83}\mathbf{A}_{87}\mathbf{I}_{84}\mathbf{E}_{91}$	28.14	LIG_Clathr_ClatBox_1 (Ref. 29)
PxxP	1nkzE $\mathbf{P}_{12}\mathbf{A}_{13}\mathbf{I}_{16}\mathbf{P}_{17}$ 1ix2A $\mathbf{P}_{52}\mathbf{K}_{38}\mathbf{R}_{86}\mathbf{P}_{94}$	27.56	(Ref. 33)
[VILMAFP]KxE	1hqgC $\mathbf{V}_{203}\mathbf{K}_{205}\mathbf{D}_{204}\mathbf{E}_{256}$ 1rypI $\mathbf{V}_{195}\mathbf{K}_{29}\mathbf{A}_{27}\mathbf{E}_{197}$	15.05	MOD_SUMO (Ref. 29)
[PSAT]x[QE]E	1kacB $\mathbf{A}_{127}\mathbf{P}_{128}\mathbf{Q}_{52}\mathbf{E}_{50}$ 1mbxA $\mathbf{P}_{80}\mathbf{F}_{24}\mathbf{Q}_{79}\mathbf{E}_{23}$	19.10	LIG_TRAF2_1 (Ref. 29)
AxxxA	1svfC $\mathbf{A}_{179}\mathbf{V}_{175}\mathbf{H}_{171}\mathbf{V}_{168}\mathbf{A}_{167}$ 1gl2B $\mathbf{A}_{213}\mathbf{H}_{216}\mathbf{V}_{217}\mathbf{Q}_{219}\mathbf{A}_{220}$ 1bgyE $\mathbf{A}_{48}\mathbf{G}_{46}\mathbf{V}_{45}\mathbf{T}_{44}\mathbf{A}_{41}$ 1gmjC $\mathbf{A}_{21}\mathbf{K}_{24}\mathbf{G}_{23}\mathbf{Q}_{27}\mathbf{A}_{28}$ 1n7sB $\mathbf{A}_{240}\mathbf{Y}_{243}\mathbf{V}_{244}\mathbf{R}_{246}\mathbf{A}_{247}$ 1bkvB $\mathbf{A}_{47}\mathbf{L}_{46}\mathbf{G}_{45}\mathbf{R}_{44}\mathbf{A}_{43}$ 1ek9B $\mathbf{A}_{343}\mathbf{A}_{347}\mathbf{Q}_{346}\mathbf{T}_{378}\mathbf{A}_{351}$	6.81	(Ref. 30)

Interestingly, on visual inspection, we found many cases whereby the interface residue sequences that matched the known linear binding motifs are themselves non-sequential in their primary sequences. This is rather intriguing because linear sequence motifs have traditionally been assumed to occur as contiguous sequence segments; yet, we have found in our interface clusters numerous instances whereby the residues from different parts of a protein chain come together spatially to mimic some known linear binding motifs. For example, Figure 8 shows two interface residue sequences in one cluster that come together spatially to re-assemble the **KPxx[QK]** linear motif (ELM ID: LIG_SH3_4). Figure 9 shows another example of sequentially discontinuous interface residues re-assembling another known linear motif (**RxLx[EQ]**[32]). In this example, both sets of residues corroborate to form a similar interface that interacts with an α-helix.

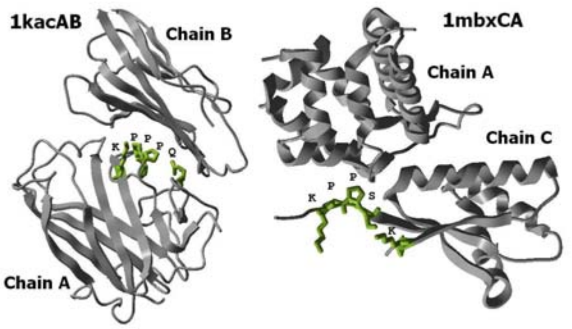


Fig. 8. Conservation of motif **KPxx[QK]** in a particular cluster.

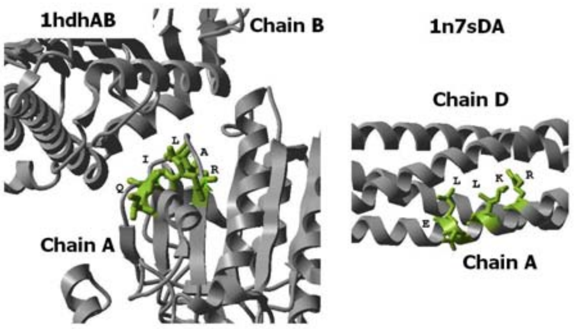


Fig. 9. Conservation of motif **RxLx[EQ]** in a particular cluster.

These intriguing examples discovered in our interface clusters suggest that foldings of protein chains can be combined to yield interaction sequence motifs. This would imply that the many reported biologically important linear motifs could occur more frequently than expected, as we have yet to take into account of the possibility of sequentially discontinuous occurrences. For example, the **RxLx[EQ]** motif which was attributed to the virulency of malarial parasite *P. falciparum* in human was found in 250 to 350 of the parasite proteins by primary sequence match.[32] The actual number of proteins containing this motif could be more based on what we have observed in this work.

Currently, only $\sim$ 200 linear binding motifs out of few thousands speculated to exist are known[2]—there might also be many important biological motifs that are sequentially discontinuous that have yet to be detected. We have shown here that it is possible to relate the protein interface clusters with biologically important motifs by adopting a principle component analysis to transform residues at interaction interfaces for linear binding motif discovery. Our efficient PP*i*Clust method could thus form an alternative framework to facilitate the discovery of more novel linear motifs in the as yet unexplored structural space.

6. CONCLUSIONS

In this paper, we have proposed a novel interaction interface clustering scheme named PP*i*Clust (Protein–Protein *i*nterface Clusterer) to extract statistically significant and biologically interesting clusters of protein interfaces from 3-D protein complex structural data. As we have taken care to encode the 3-D structural patterns of interfaces with compact 1-D feature vectors, the proposed method is also time-efficient—the total time taken for the whole process is about 8 hours on a stand-lone PC. This is important as most other methods cannot scale up to mine the increasingly available structural information.

Our analysis on the resultant interaction interface clusters revealed that the structurally similar interfaces in our clusters can belong to parent proteins that have very diverse structural folds. This suggests the possibility of similar protein functions among proteins with different structural fold types, an observation that was also made in other existing works.[6–8] More interestingly, our analysis also revealed that many highly conserved linear binding motifs of well-known biological functions can also be detected in the interface clusters generated by our method. This included sequentially discontinuous occurrences of the motifs, suggesting that residues from different parts of protein can come together spatially to mimic the functions of linear motifs. In fact, there might still be important biological motifs that are spatially conserved but sequentially discontinuous yet to be detected. Our efficient PP*i*Clust method can thus enable the exploration of the yet unexplored structural space to uncover the structural basis of protein interactions.

References

1. Ng SK, Zhang Z, Tan SH. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* 2003; **19**: 923–929.
2. Neduva V, Russell RB. Linear motifs: evolutionary interaction switches. *FEBS Lett* 2005; **579**: 3342–3345.
3. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. *J Mol Biol* 1999; **285**: 2177–2198.
4. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. *Prot Struct Funct Genet* 2000; **39**: 331–342.
5. Teyra J, Doms A, Schroeder M, Pisabarro MT. SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics* 2006; **7**: 104.
6. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A dataset of protein–protein interfaces generated with sequence-order-independent comparison technique. *J Mol Biol* 1996; **260**: 604–620.
7. 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.
8. Mintz S, Shulman-Peleg A, Wolfson HJ, Nussinov R. Generation and analysis of a protein–protein interface data set with similar chemical and spatial patterns of interactions. *Prot Struct Funct Bioinfo* 2005; **61**: 6–20.
9. Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 2005; **21**: 1901–1907.
10. 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.
11. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 2005; **21**: 3201–3212.
12. Shulman-Peleg A. Personal communications 2005.
13. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. *J Mol Biol* 2004; **339**: 607–633.
14. Murtagh F, Heck A. *Multivariate Data Analysis*. Kluwer Academic 1987.
15. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–410.
16. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993; **233**: 123–138.
17. Zachary J, Iyengar SS, Barhen J. Content based image retrieval and information theroy: a general approach. *J Amer Soci Info Sci Tech* 2001; **52**: 840–852.
18. Carugo O, Pongor S. Protein fold similarity estimated by a probabilistic approach based on c(alpha)–c(alpha) distance comparison. *J Mol Biol* 2002; **315**: 887–898.
19. Choi IG, Kwon J, Kim SH. Local feature frequency profile: a method to measure structural similarity in proteins. *Proc Natl Acad Sci, USA* 2004; **101**: 3797–3802.
20. Aung Z, Fu W, Tan KL. An efficient index-based protein structure database searching method. In: *Proc 8th Intl Conf Database Systems for Advanced Applications (DASFAA'03)* 2003; pp. 311–318.
21. Dunham MH. *Data mining: introductory and advanced topics*. Prentice Hall 2003.
22. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis*. Wiley-Interscience 1990.
23. `http://www.unesco.org/webworld/idams/advguide/Chapt7_1_1.htm`.
24. Shannon CE. A mathematical theory of communication. *Bell Sys Tech J* 1948; **27**: 379–423.
25. Pieper U, Eswar N, Braberg H, *et al.* MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* 2004; **32**: D217–222.
26. Tsai CJ, Xu D, Nussinov R. Structural motifs at protein–protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci* 1997; **6**: 1793–1805.
27. 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.
28. Ceol A, Chatr-aryamontri A, Santonico E, Sacco R, Castagnoli L, Cesareni G. DOMINO: a database of domain–peptide interactions. *Nucleic Acids Res* 2006; **35**: D557–560.
29. Puntervoll P, Linding R, Gemund C, *et al.* ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 2003; **31**: 3625–3630.
30. Kleiger G, Grothe R, Mallick P, Eisenberg D. GXXXG and AXXXA: common alpha-helical interaction motifs in proteins, particularly in extremophiles. *Biochemistry* 2002; **41**: 5990–5997.
31. Dombrosky-Ferlan P, Grishin A, Botelho RJ, Sampson M, Wang L, Rudert WA, Grinstein S, Corey SJ. Felic (CIP4b), a novel binding partner with the Src kinase Lyn and Cdc42, localizes to the phagocytic cup. *Blood* 2003; **101**: 2804–2809.
32. Przyborski J, Lanzer M. Parasitology: the malarial secretome. *Science* 2004; **306**: 1897–1898.
33. Ravi Chandra B, Gowthaman R, Raj Akhouri R, Gupta D, Sharma A. Distribution of proline-rich (PxxP) motifs in distinct proteomes: functional and therapeutic implications for malaria and tuberculosis. *Protein Eng Des Sel* 2004; **17**: 175–182.