Discovery of Binding Motif Pairs from Protein Complex Structural Data and Protein Interaction Sequence Data

H. Li, J. Li, S.H. Tan, S.-K. Ng

Pacific Symposium on Biocomputing 9:312-323(2004)

DISCOVERY OF BINDING MOTIF PAIRS FROM PROTEIN COMPLEX STRUCTURAL DATA AND PROTEIN INTERACTION SEQUENCE DATA

H. LI [1,2] J. LI [1, a] S. H. TAN [1] S.-K. NG [1]
[1] *Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613*
[2] *School of Computing, National University of Singapore, Singapore, 119260*
Email: {*haiquan,jinyan,soonheng,skng@i2r.a-star.edu.sg*}

Abstract

Unravelling the underlying mechanisms of protein interactions requires knowledge about the interactions' binding sites. In this paper, we use a novel concept, *binding motif pairs*, to describe binding sites. A binding motif pair consists of two motifs each derived from one side of the binding protein sequences. The discovery is a directed approach that uses a combination of two data sources: 3-D structures of protein complexes and sequences of interacting proteins. We first extract *maximal contact segment pairs* from the protein complexes' structural data. We then use these segment pairs as templates to sub-group the interacting protein sequence dataset, and conduct an iterative refinement to derive significant binding motif pairs. This combination approach is efficient in handling large datasets of protein interactions. From a dataset of 78,390 protein interactions, we have discovered 896 significant binding motif pairs. The discovered motif pairs include many novel motif pairs as well as motifs that agree well with experimentally validated patterns in the literature.

1 Introduction

Protein-protein interactions play a crucial role in the operations of many key biological functions such as inter-cellular communications, signal transduction, and regulation of gene expressions. Unravelling the underlying mechanisms of these interactions will provide invaluable knowledge that could lead to the discovery of new drugs and better treatments for many human diseases.

Physically, protein interactions are mediated by short sequences of residues that form the contact interfaces between two interacting proteins, often referred as their binding sites. Though many experimental methods[1] and computational[2] methods have been developed to detect protein interactions with increasing levels of accuracies, few methods can

[a] To whom correspondence should be addressed.

pinpoint the specific residues in the proteins that are involved in the interactions. Such information are necessary for the interaction data to be directly useful for drug discovery. To determine the binding sites between interacting proteins, usually experimental methods include mutagenesis studies and phage display[3], which are tedious and time-consuming. Computational methods often include docking approaches and domain-domain interaction approaches. The docking approach is based on the analysis of bound protein structures. The use of this approach is very limited. The main reason is that resolved structures of proteins are often not available due to the limitations in scalability and coverage of current protein structural determination technologies. The domain-domain interaction approach assumes that protein interactions are determined by the interactions between domains and is aimed to figure out the interactions only among predefined domains[4,5,6]. However, some domains may not directly determine the interactions, but only function as determinants of protein folding. Even though the domains involve in protein interactions, not all of their residues are contained in the binding sites and contribute to the role of the interactions.

In this work, we study the problem of binding site at residue level rather than at domain level. Our basic idea is that correlated sequence motif pairs determine the interactions. A similar concept, correlated sequence-signature pairs, was first proposed by Sprinzak[4] with the expression of domain pairs. We focus on efficient *in silico* discovery of our motif pairs from multiple data sources about protein interactions. Ideally, such interacting motif pairs should be discovered from protein complex structural data. However, as discussed above, the availability of such data is very limited. Alternatively, interacting motif pairs may be discovered by analyzing their co-occurrence rates in interacting protein pairs' sequences. However, as high-throughput detection technologies such as two-hybrid screenings[7,8] can rapidly generate large datasets of experimentally determined protein interactions, the search space on the associated protein sequences is enormous. The high false positive rates observed in high-throughput protein interaction data could also diminish the biological significance of motif pairs detected solely from protein interaction sequences.

To address these issues in mining motif pairs, we propose a joint approach that makes use of the two available types of interaction data: (1) the limited structural data of protein complexes that provide exact information on inter-protein contact sites, and (2) the abundantly available interacting protein sequence pairs from high-throughput interaction detection experiments. The structural data of protein complexes are carefully mined for contact residues; these are then computationally extended into the so-called *maximal contact segment pairs* which we will define later. The complexes' maximal segment pairs are then de-

ployed to seed the discovery of motif pairs from large sequence datasets of interacting proteins, followed by an iterative refinement procedure to ensure the significance of the derived motif pairs. This combined directed approach reduces the formidable search space of interacting protein sequences while providing some biological support for the motifs discovered. Indeed, many of our motif pairs discovered this way can be confirmed by biological patterns reported in the literature, as we will show later.

We present the overall picture of our method in Section 2. In Sections 3 and 4, we describe new algorithms to discover maximal contact segment pairs from protein complex data, and then to discover binding motif pairs from interacting protein sequence data. Results showing the effectiveness and significance of this joint approach are presented in Section 5. Finally, we conclude and discuss about possible future work in Section 6.

2 Overview of Our Method and Data Used

A key idea in our proposed method for discovering significant binding motif pairs is the detection of *maximal contact segment pairs* between two proteins residing in a complex. First, all possible pairs of spatially contacting residues are determined from the 3-D structure data of a protein complex. These contact points are then extended to capture as many continuous binding residues along the two proteins as possible, deriving the maximal contact segment pairs. Computationally, the derivation of maximal contact segment pairs is a challenging problem. In Section 3, we will describe an algorithm to discover them efficiently.

Our objective is to discover *significant binding motif pairs* from protein-protein interaction sequence datasets. Using the maximal contact segment pairs that we have discovered from the protein complex structural data, we cluster the interacting protein sequence data into sub-groups, each corresponding to one maximal contact segment pair. Then from each sub-group, we use a new motif discovery algorithm and an iterative optimization refinement algorithm to discover a binding motif pair. To assess the significance of binding motif pairs in the refinement procedure, we define a measure called *emerging significance*, which is similar to the concept of emerging patterns[9]. This measure is based on both positive and negative interaction datasets: A pattern or motif pair is said to have a high emerging significance if it has a high frequency in the positive dataset but a relatively low frequency in the negative dataset.The iterative refinement is terminated when the motif pairs reach an optimized level of emerging significance.

The protein complex dataset used in this study is a non-redundant subset from PDB where the maximum pairwise sequence identify is 30% and only structures with resolution 2.0 or better are included. The set

used was generated on 9th June 2003 and contained 1533 entries in which each entry has at least 2 chains. As mentioned, our emerging significance approach requires the use of both positive and negative instances of pairwise protein-protein interactions. For positive protein-protein interaction sequence data, we used the data by von Mering *et al*[1]. This dataset covers almost all those interaction data generated by experimental methods and in-silico methods for yeast proteins. In total, there are 78,390 non-redundant interactions in this dataset. However, there are currently no large datasets of experimentally validated negative interactions. As such, we generated a *putative* negative interaction dataset by assuming that any possible protein pair in yeast that do not occur in the positive dataset as a negative interaction. As our emerging significance measure only requires that the detected patterns have relatively lower frequency in the negative datasets, the effect of potential false negative interactions in this putative negative dataset is minimal.

3 Discovering Maximal Contact Segment Pairs from Protein Complexes

3.1 Preprocessing: Compute Contact Sites

Given a pair of proteins in a complex, a *contact site* is an elemental pair of two residues or atoms, each coming from one of the two proteins, that are close enough in space. A protein complex usually consists of multiple proteins, in this study we then consider all pairs of proteins in a protein complex to obtain all contact sites in this step.

We define a *contact site* mathematically as follows: *Suppose two proteins with 3-D structural coordinates in (x,y,z),* $L_a = \{(a_i, x_{a_i}, y_{a_i}, z_{a_i}), i = 1...m\}$ *and* $L_b = \{(b_j, x_{b_j}, y_{b_j}, z_{b_j}), j = 1...n\}$. *The pair* (a_i, b_j) *is a contact site if* $dist(a_i, b_j) \leq \varepsilon$, *where* a_i *and* b_j *are the atom id in the protein* L_a *and* L_b *respectively, and* ε *is an empirical threshold for the Euclidean distance function* $dist(.,.)$. *Such a pair is denoted* $Contact(a_i, b_j)$, *or equivalently* $Contact(b_j, a_i)$.

Note that a contact site in the atom level directly implies a contact site in residue level because each atom is a part of a unique residue. Hereafter, we will discuss contact sites only at the residue level. Since two residues are said to be in contact if one of the atoms in a residue is in contact with one atom in the other residue, it is possible for a residue to be in contact with multiple residues.

3.2 Extract Contact Segment Pairs

Next, we extend the concept of contact sites to the concept of *contact segment pairs*, aiming to search for large areas of contact sites in a pair of

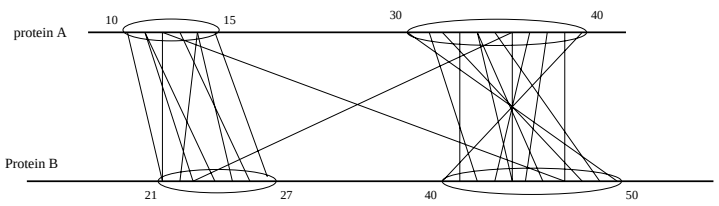

Figure 1: An illustration of contact segment pairs in a pair of interacting proteins A and B. Here, protein A is said to be the *opposite* protein of B, and vise versa.

binding proteins. Figure 1 shows our idea, depicting a typical scenario where segments of residues in one protein are continuously in contact with segments of residues in the other protein. As an illustration, the segment $[a_{10}, a_{15}]$ in protein A of Figure 1 is in contact with the segment $[b_{21}, b_{27}]$ in protein B. That is, they are a contact segment pair. But the segment $[a_{30}, a_{40}]$ in protein A and the segment $[b_{21}, b_{27}]$ in protein B are collectively not a contact segment pair.

Formally, the definition is: *A contact segment pair is a segment pair* $([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}])$ *satisfying, for* $\forall a_i \in [a_{i_1}, a_{i_2}]$, $\exists b_j \in [b_{j_1}, b_{j_2}]$ *such that* (a_i, b_j) *is a contact site, where* $a_{i_1}, a_{i_2}, b_{j_1}, b_{j_2}$ *are residue ids in two proteins* L_a *and* L_b*. Such a pair of segments is sometimes denoted* $Contact([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}])$.

A **maximal contact segment pair** is then defined as a contact segment pair such that no other contact segment pair can contain the both segments of this contact pair. In this paper, we are interested in the following problem:

Definition 1 ***Maximal Contact Segment Pairs Problem:*** *Given a pair of binding proteins* L_a *and* L_b*, suppose* $C = \{(a_i, b_j) |\ Contact(a_i, b_j)$ *with respect to the two proteins* L_a*and* $L_b\}$*, the problem is how to find all possible maximal contact segment pairs from* C *with their segment lengths all longer than a threshold.*

A naive approach to solving this problem would require testing all possible segment pairs. Suppose two proteins L_a and L_b have m and n residues respectively, then the proteins L_a and L_b will have m^2 and n^2 possible segments respectively. For each combination, $O(mn)$ time complexity would be required for the computation. So, the total time complexity for such a naive approach will be $O(m^3 * n^3)$ per pair of proteins in each complex. This is very expensive particularly when the protein complexes are large and there are hundreds or thousands of protein complexes need to be examined. We present a more efficient method to discover maximal contact segment pairs here.

Observe that for each residue, it may be in contact with multiple

residues in the opposite protein (see Figure 1). We introduce a concept named *coverage* to capture this phenomenon; it will be shown later that this is a useful concept for improving the efficiency of our discovery algorithm. The coverage of a residue a_i, denoted $Cov(a_i)$, is the set of all residues in the opposite protein that are in contact with this residue, namely $Cov(a_i) = \{b_j | (a_i, b_j) \in C\}$.

The coverage of a segment $[a_{i_1}, a_{i_2}]$, denoted $Cov([a_{i_1}, a_{i_2}])$, is the union of the coverages of all its residues, namely,
$Cov([a_{i_1}, a_{i_2}]) = \cup_{a_i \in [a_{i_1}, a_{i_2}]} Cov(a_i)$.

The following proposition is useful in our algorithm to discover maximal contact segment pairs efficiently.

Proposition 1 *A segment pair* $([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}])$ *is a contact segment pair iff the coverage of any of the two segments contains the other segment, i.e.* $Contact([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}]) \Longleftrightarrow (Cov([a_{i_1}, a_{i_2}]) \supseteq [b_{j_1}, b_{j_2}]) \wedge (Cov([b_{j_1}, b_{j_2}]) \supseteq [a_{i_1}, a_{i_2}])$.

Proof: $\Rightarrow$: We use contradiction to prove. Suppose $Cov([a_{i_1}, a_{i_2}]) \supseteq [b_{j_1}, b_{j_2}]$ is not true, then there exists a $b_j \in [b_{j_1}, b_{j_2}]$ but this $b_j \notin Cov([a_{i_1}, a_{i_2}])$. This means there is no $a_i \in [a_{i_1}, a_{i_2}]$ in contact with b_j. This contradicts the assumption. Therefore, $Cov([a_{i_1}, a_{i_2}]) \supseteq [b_{j_1}, b_{j_2}]$. We can prove $Cov([b_{j_1}, b_{j_2}]) \supseteq [a_{i_1}, a_{i_2}]$ in a symmetrical manner.

$\Leftarrow$: If $Cov([a_{i_1}, a_{i_2}]) \supseteq [b_{j_1}, b_{j_2}]$, this means that for each $b_j \in [b_{j_1}, b_{j_2}]$, there exist at least one contact site in $[a_{i_1}, a_{i_2}]$. Similarly, the residues in the other segment have the same property. ■

Our algorithm is a top-down recursive algorithm. At the initial step, each entire protein in a pair is treated as a segment. A series of recursive breaking-down are then performed to output maximal contact segment pairs, using the above proposition to determine when to break-down a segment into several smaller segments and when to terminate producing a new candidate segment pair. The details of our algorithm are as follows:

Input: Two proteins $L_a = \{(a_i, x_{a_i}, y_{a_i}, z_{a_i}), i = 1...m\}$ and $L_b = \{(b_j, x_{b_j}, y_{b_j}, z_{b_j}), j = 1...n\}$, two special segments $[a_1, a_m]$, and $[b_1, b_n]$, and $C = \{(a_i, b_j) | \ Contact(a_i, b_j),\ 1 \leq i \leq m,\ 1 \leq j \leq n\}$.

Output: A set of maximal contact segment pairs.

Preparation Step: Compute $Cov(a_i)$ and $Cov(b_j)$ for all $1 \leq i \leq m$, $1 \leq j \leq n$.

Initialization Step: Put the initial segment pair $([a_1, a_m], [b_1, b_n])$ into the candidate list.

repeat

Segment Coverage Step: Remove the first segment pair from the candidate list, denoted $([x_{i_1}, x_{i_2}], [y_{j_1}, y_{j_2}])$; Compute the coverage for $Cov([x_{i_1}, x_{i_2}]) \cap [y_{j_1}, y_{j_2}]$.

Splitting Step:

if $(Cov([x_{i_1}, x_{i_2}]) \cap [y_{j_1}, y_{j_2}]) == [y_{j_1}, y_{j_2}]$ **then**
if $(Cov([y_{j_1}, y_{j_2}]) \cap [x_{i_1}, x_{i_2}]) == [x_{i_1}, x_{i_2}]$ **then**
Output the segment pair.
else
Add $([y_{j_1}, y_{j_2}], [x_{i_1}, x_{i_2}])$ into the candidate list.
end if
else
Split $Cov([x_{i_1}, x_{i_2}]) \cap [y_{j_1}, y_{j_2}]$ into w number of continuous sub-segments, denoted $[y_{k_{2t-1}}, y_{k_{2t}}], t = 1....w$, put each segment pair $([y_{k_{2t-1}}, y_{k_{2t}}], [x_{i_1}, x_{i_2}]), t = 1...w$, into the candidate list.
end if
until The candidate list is empty.

A detailed example can be found in this paper's supplementary information[10].

4 Discovering Binding Motif Pairs from Interacting Protein Sequence Pairs

Next, we describe how to discover binding motif pairs from protein interaction sequence data using the maximal contact segment pairs detected from protein complexes.

4.1 Seeded Sub-grouping and Consensus Motif Discovery

We use each of the discovered maximal contact segment pairs as seed to sub-group the interaction sequence pairs such that all the interaction pairs that *contain* the contact segment pair are grouped together. We then conduct a consensus motif discovery in each of the sub-groups of protein interaction sequences.

First, let us give the following two definitions:

Contain: *Suppose a sequence $S = s_1 s_2 ... s_u$, and a segment $P = p_1.p_2...p_v$. S contains P, denoted Contain(S, P), if Local_Alignment(S, P) $\geq \lambda$, where λ is an empirical threshold.*

Cluster of a Contact Segment Pair: *Given an interaction dataset D consisting of n sequence pairs, denoted $D = \{(S_i^1, S_i^2), 1 \leq i \leq n\}$, and a segment pair $P = (P_1, P_2)$, the cluster of this segment pair with respect to D, denoted $G_D(P)$, is*

$$\begin{array}{ll} & \{(S_1', S_2') |\ (S_1', S_2') \in D, Contain(S_1', P_1), and\ Contain(S_2', P_2)\} \\ \cup & \{(S_1'', S_2'') |\ (S_2'', S_1'') \in D, Contain(S_1'', P_1), and\ Contain(S_2'', P_2)\} \end{array}$$

By this way of sub-grouping the interaction dataset, the resulting clusters of different segment pairs may overlap with one another. Biologically, this is important because one protein may involve interactions with different proteins in different locations.

Given the cluster of a contact segment pair, our subsequent step is to find two consensus motifs, one from all those S_1' plus all those S_1'' (namely the left-side sequences of those protein sequence pairs), and the other from all those S_2' plus all those S_2'' (namely the right-side sequences of those protein sequence pairs). At each side, we align all the sequences according to the best alignments with respect to the corresponding segment (P_1 or P_2 in this case). We used the score matrix developed by Azarya[11] for the local alignment[12], since structure is preserved for any residue pairs that have high scores in the matrix.

To obtain the consensus motif from each side of these alignments, every column in the alignment is examined as follows: If the occurrence of a residue in this column is above the stated threshold, we include it in the the consensus motif. If there are no such residues, we treat this column as a wildcard. It is also possible to use alternative methods such as EMOTIF[13] to find the consensus motifs.

These two consensus motifs form a *binding motif pair*. Note that we derive this binding motif pair starting from one contact segment pair. So, given a set of maximal contact segment pairs discovered from the protein complex dataset, we can obtain a set of binding motif pairs by going through all these maximal contact segment pairs on the interacting protein sequence datasets.

4.2 Iterative Refinement

Next, we perform an iterative refinement on the binding motif pairs discovered in the last subsection. The purpose of doing this is to optimize these binding motif pairs. Given a binding motif pair Q, our refinement algorithm uses Q to sub-group the interacting protein sequences dataset, and generates a new binding motif pair Q' (using *exact match* instead of local alignment here), as discussed in the last subsection but replacing the maximal contact segment pair P with Q. Iteratively, the algorithm repeats the procedure, using Q' as Q, until Q' reaches an optimized state.

The stopping criteria used here is based on a concept of *emerging significance* of consensus motifs. Recall that we have established two protein sequence pair datasets: the interaction dataset (also called the positive dataset) and the negative dataset. So far, we have used only the positive dataset in generating the consensus motifs. To measure the emerging significance of a pair of consensus motifs, we make use of both of the positive and negative datasets. If a motif pair is significant, it is reasonable to expect the pair to occur in the positive dataset much more frequently than in the negative dataset. We give the definitions for emerging significance below:

Frequency of a motif pair with respect to a dataset: *Suppose*

we have a dataset D consisting of sequence pairs $D=\{(S_i^1, S_i^2)|1 \leq i \leq n\}$, *the frequency of a motif pair* $P=(P_1, P_2)$ *with respect to D is defined as:* $Freq(P, D) = \frac{|G_D(P)|}{n}$.

Significant motif pairs: *Suppose we have a positive dataset* D_{Pos} *and a negative dataset* D_{Neg}. *A motif pair P is significant if:* $ratio(P, D_{Pos}, D_{Neg}) = \frac{Freq(P, D_{Pos})}{Freq(P, D_{Neg})} \geq \tau$, *where* τ *is a threshold. We also call* $ratio(P, D_{Pos}, D_{Neg})$ *the* **emerging significance** *of P.*

4.3 Time Complexity of the Method

The time complexity for sub-grouping based on a segment pair is $O((|D_{Pos}|+|D_{Neg}|) * |CP|)$ because of using local alignment. Here CP represents the set of maximal contact segment pairs. The size of binding motif pairs is $O(|CP|)$ in the case of using our column-by-column consensus algorithm. The time used to compute the clusters for motif pairs in each pass is linear if the suffix tree approach[14] is applied to conduct the exact match for regular patterns. The complexity of computing a consensus motif pair from a cluster is also linear. Suppose there is at most K passes for the algorithm to terminate, the number of motif pairs is N_{CP}, the time complexity for the refinement of motif pairs is $O(((|D_{Pos}| + |D_{Neg}|) * N_{CP} + |CP|) * K)$. In total, the time complexity for this step is $O((|D_{Pos}| + |D_{Neg}|) * (|CP| + N_{CP} * K) + |CP| * K)$.

5 Implementation and Results

In the initial step of computing contact sites from the protein complex data, we set the threshold ε to $5\dot{A}$. More than 56% of the complexes were found to contain at least one contact site. We also set the number 4 as the threshold of segment length. We found 1403 maximal segment pairs from this complex dataset.

For sub-grouping the interaction dataset using the maximal segment pairs, a threshold should be set in the *contain* operation. Instead of setting λ to be a constant, it is more reasonable to set the threshold strictly for short segments but loosely for long segments. The actual parameters used in our experiment are provided in our supplementary information [10].

Our refinement procedure was performed for 7 iterative passes. After that all the motif pairs became stable. We found a total of 896 motif pairs to be significant when the emerging significance threshold τ was set to be 2. The detailed distribution of emerging significance values can again be found in our supplementary information [10].

All our source codes of the algorithms were run on a Pentium 4 PC with 2.4 GHZ CPU and 256M RAM. Most of the time (around 12 hours) were spent to sub-group the interaction sequence data using the maximal

contact segment pairs. The mining of all the maximal segment pairs was very fast, spending only 50 seconds. The refinement algorithm was also fast, spending about 1 hour. Note that this time cost is acceptable considering the enormity of the problem space.

Although the objective is to discover novel motif pairs, to evaluate the biological significance of the motif pairs found by our algorithms, it is important to verify that some of the discovered motifs agree well with experimentally validated patterns in the literature. However, most publications on the experimental discovery of binding motifs only report a single motif on one side rather than a pair of binding motifs. As such, we can only confirm the coincidence of *individual* motifs in our motif pairs with the reported binding motifs found by traditional experimental methods. For example, for the mutagenesis method, we used key words 'binding motif OR site AND mutagenesis' to search all biomedical abstracts in PUBMED of NCBI. 202 motifs were found, in which 91 motifs are compatible with at least one in our motifs, 58 motifs are highly similar with ours. We show the first 5 matches in Table 1. Similar comparison with the phage display method is provided in our supplementary information [10].

Table 1: Motif coincidence with the mutagenesis method.

Our Motif	Mutagenesis Motif	PMID of Mutagenesis Motif
ALETS	LETS	11435317
P[IV]DL	PVDLS	11373277
L[DN]LL	LLDLL	11451993
K[DE]K[EK]	KEKE	10748065
PIDLSLKP	P*DLS	11062046

Table 2 illustrates how we can compare motif *pairs* using the individual binding motifs reported in the literature. As an example, we use the binding consensus sequences in the list compiled by Kay et al [15] for various proteins by phage display. First, we identify the individual motifs in our population of discovered motif pairs that match closely with a binding consensus sequence in the compiled list. Then, for each of such matched motifs, we verify whether the motif on the other side of the corresponding motif pair are found in proteins known to bind to the particular consensus sequence. In Table 2, we list six example binding consensus sequences from Kay et al [15] compiled list in the first column. In the second column, we list the individual matched motifs from our population of discovered motif pairs—we arbitrarily assign these motifs as the "left motifs". In the third column, we show the motifs on the other sides (the "right motifs") of the matched motif pairs. Since these right motifs are also found in the proteins (shown in the fourth column) reported to bind to the corresponding consensus sequence, the motif pairs

can be considered to be biologically verified. More examples are detailed in our website [10].

Table 2: Motif pair coincidence between our motif pairs and peptide-protein binding pairs.

Consensus Sequence	Left motif	Right Motif	Binding Protein
P*LP*[KR]	P[EK]*P	GV[FI]S	CRK A
P*LP*[KR]	P[ILV][FIL]PG	P[ILV][FL]PG	CRK A
P*LP*[KR]	P[ILV][FL]PG	P[ILV][FIL]PG	CRK A
[RKH]PP[AILVP]P[AILVP]KP	P[IV][EP][IV]A	AAS[FI]	Cortactin
RLP*LP	P[EK]*P	GV[FI]S	Synaptojanin I
[RKH]PP[AILVP]P[AILVP]KP	P[IV][DP]P[FL]	PL[DP]PL	Shank

6 Conclusion and Further Work

The mining of binding motif pairs from protein interaction data is important for extracting knowledge that can lead to the discovery of new drugs. Most of the work reported in the literature only dealt with finding individual binding motifs rather than pairs of interacting motifs. Since motif pairs—unlike single binding motifs—can provide better information for understanding the interactions between proteins, we studied the problem of finding binding motif pairs from large protein interaction datasets.

Our approach combines the mining of large protein interaction sequence datasets with the use of smaller protein complex structural datasets to direct the search. For mining protein complex structural data, we have formulated the detection of maximal contact segment pairs as a novel computational search and optimization problem, and we have provided an efficient algorithm for that. The maximal contact segment pairs derived can then be deployed as seeds for sub-grouping the vast dataset of interacting protein sequence pairs so that motif discovery algorithms can be directed to find the motif pairs within sub-groups. By iteratively applying this technique, we refine these motif pairs until they reach a satisfactory level of emerging significance.

The results have shown that our combination approach is efficient and effective in finding biologically significant binding motif pairs. Many of the motif pairs that we have discovered coincided well with known motif pairs independently discovered by experimental methods. However, our this directed approach heavily depends on protein complex data source. As the current complex dataset is very limited, our approach may miss many other important motif pairs. On the other hand, it is worthwhile to improve our approach for discovering more significant binding motif pairs. For example, in our current definition of contact segment pairs, each residue in one segment is strictly required to have at least one contact residue in the other segment. Biologically, contact segment pairs are still valid even if a few residues in the segments are not in contact.

Computationally, however, our top-down recursive algorithm for finding maximal contact segment pairs will no longer be valid without this constraint. Therefore, one future research direction will be to explore the relaxation of this constraint while retaining the efficiency of the algorithm.

References

1. Von Mering C et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
2. Valencia A and Pazos F. Computational methods for the prediction of protein interactions. *Curr Opin Struc Biol*, 12(3):368–373, 2002.
3. B.k. et al Kay. *Phage display of peptides and proteins*. Academic Press, New York, 1996.
4. Sprinzak E and Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–92, 2001.
5. Deng M et al. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, 12(10):1540–8, 2002.
6. Ng SK et al. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–9, 2003.
7. Uetz P et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–7, 2000.
8. Ito T et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci*, 98(8):4569–74, 2001.
9. Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM SIGKDD*, pages 43–52, USA, Aug 1999.
10. Supplementary Information. http://sdmc.i2r.a-star.edu.sg/protein-interaction/.
11. Azarya-Sprinzak E et al. Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Eng*, 10(10):1109–22, 1997.
12. Smith TF and Waterman MS. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981.
13. Nevill-Manning CG et al. Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci*, 95:5865–71, 1998.
14. E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14:249–260, 1995.
15. Kay BK et al. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.*, 14(2):231–41, 2000.