Discovery of Stable and Significant Binding Motif Pairs from PDB Complexes and Protein Interaction Datasets

Haiquan Li[1, 2] *and Jinyan Li*[1]

[1]*Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613 and*
[2]*School of Computing, National University of Singapore, Singapore, 119260*

ABSTRACT

Motivation: Discovery of binding sites is important in the study of protein-protein interactions. In this paper, we introduce *stable* and *significant* motif pairs to model protein binding sites. The stability is the pattern's resistance to some transformation. The significance is the unexpected frequency of occurrence of the pattern in a sequence dataset comprising known interacting protein pairs. Discovery of stable motif pairs is an iterative process, undergoing a chain of changing but converging patterns. Determining the starting point for such a chain is an interesting problem. We use a protein complex dataset extracted from PDB to help identifying those starting points, so that the computational complexity of the problem is much released.

Results: We found 913 stable motif pairs, of which 765 are significant. We evaluated these motif pairs using comprehensive comparison results against random patterns. Wet-experimentally discovered motifs reported in literature were also used to confirm the effectiveness of our method.

Supplementary Information: http://sdmc.i2r.a-star.edu.sg /BindingMotifPairs

Contact: {haiquan,jinyan}@i2r.a-star.edu.sg

INTRODUCTION

Protein-protein interactions play important roles in many biological processes such as for inter-cellular communication, for signal transduction, and for regulation of gene expressions. Binding sites are crucial clues to unraveling protein-protein interactions. The discovery of binding sites is also useful for the prediction of unknown protein-protein interactions, for the library design of phage display (Smith, 1985), and for drug design as targets in proteomics.

The discovery of binding sites can be categorized into two different approaches. One is focused on the single sides of binding sites, while the other emphasizes the cooperation of both sides. We are interested in the second approach and call it the binding-pair approach. Both experimental and computational methods can deduct binding pairs. Experimental methods include those for analyzing protein complexes (Josephson et al., 2001), phage display (Smith, 1985; Rodi et al., 2001; Sidhu et al., 2003) and mutagenesis (Botstein and Shortle, 1985; Clemmons, 2001). These methods usually lead to relatively high accuracy, but they are time-consuming and cost-expensive. As complementary methods, computational ones are fast and economical to narrow down the search space.

Current computational methods for discovering binding pairs are mainly concentrated on domain-domain interactions. Sprinzak and Margalit (2001) first termed domain-domain interactions as correlated sequence-signatures. Deng et al. (2002) applied a maximum likelihood method to statistically estimate domain-domain interactions. Ng et al. (2003) used an integrative score system to deduct domain-domain interactions. All these methods stand at the domain level and study only pre-defined domains. Note that domain themselves may not be binding sites. For example, they can be folding determinants instead. Also, most domains are lengthy segments of residues, where only a part of them are contained in binding sites. Therefore, how to pinpoint those specific regions that are really involved in binding behavior becomes our research interests.

In this paper, we examine a simple type of binding pairs where each side of a binding site consists of a short sequence of continuous residues and where the two sides approach spatially with each other closely. We call these short sequences *motifs* and the binding sites *binding motif pairs*. We present a computational method to discover such-specified binding motif pairs from a combination of two protein interaction datasets.

We require our binding motif pairs to be *stable*. The notion of stable motif pairs is rooted in the fact that many biological phenomenon exist in stable status, and these stable status might be evolved from their past unstable status. So, from a certain starting point to the stable point, mathematically it is a chain of changing but converging patterns. Second, we require our binding motif pairs to be *significant*. We propose statistical measurements to evaluate the significance not only for single motifs but more importantly for their co-occurrence as pairs. By significance, we mean that their observations or supports should be much higher than their random expectations.

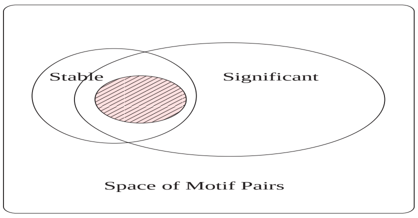

Fig. 1. The relation of stable motif pairs and significant motif pairs. The special subset of stable and significant motif pairs are our aim in this paper.

Combining these two ideas, our binding motif pairs are mathematically stable and statistically significant.

The discovery of all stable and significant motif pairs is a challenging problem as the number of candidates is huge. In this paper, we narrow down our search space by looking for only a special subset of stable and significant motif pairs. The starting points of this subset of stable motif pairs are derived from a protein complex dataset that is known to contain the most biologically reliable data about protein binding sites. By this way, the binding motif pairs aimed to discover would have high confidence because of the biological support from the complex data. This is strongly confirmed by our comprehensive comparison experiments with random patterns and by wet-experimentally discovered binding motifs reported in the literature. Figure 1 relates stable motif pairs and significant motif pairs and locates where are motif pairs that we are most interested in.

In the next section, we describe two datasets and define basic notations. Then we explain stable motif pairs, significant motif pairs, and starting motif pairs in a formal way using three sections of this paper. Finally, we report our discoveries and evaluations.

DATA AND BASIC NOTATIONS

We use two interaction datasets in this paper: a sequence dataset of interacting protein pairs collected by von Mering et al. (2002), and a protein complex dataset derived from PDB (http://www.rcsb.org/pdb/). The sequence dataset consists of 78390 non-redundant interactions, containing almost all the latest interacting protein pairs in yeast genome produced by various experimental and high confident computational methods. The protein complex dataset was generated from PDB on the 9th of June, 2003, containing 1533 such entries that have at least two chains, by using online search tools in PDB-REPRDB (http://mbs.cbrc.jp/pdbreprdb-cgi//reprdb_query.pl). In this complex dataset, the maximum pairwise sequence identity between any two complexes is 30% and each complex has a structure of resolution 2.0 or higher. In this study, the complex dataset is first used to generate starting points for stable motif pairs, then the interacting sequence dataset is used to transform those starting motif pairs so as to output a set of stable and significant binding motif pairs.

The following are basic notations that are frequently used in this paper.

Σ	the alphabet of the 20 amino acids
P	a protein: a sequence of amino acids
M	a motif: a sequence of amino acid sets
PPr	$= \{P_1, P_2\}$, a protein pair
MPr	$= \{M_L, M_R\}$, a motif pair
PrtnDB	the protein database
$\mathcal{D}$	a sequence dataset of interacting protein pairs
π	the absolute support of a motif or a motif pair
π^c	the contributive support of a motif pair
z_s	Z-score of a motif
p_s	P-score of a motif pair

More formally, a protein P is denoted by $a_1 a_2 \cdots a_l$, where $a_i \in \Sigma$ and $l > 0$. A motif M is denoted by $\mathcal{A}_1 \mathcal{A}_2 \cdots \mathcal{A}_k$, where $\mathcal{A}_i \subseteq \Sigma$ and $k > 0$. For example, $M = \{E, K, N\}\{P\}\Sigma\{D, E\}$. (Traditionally, it is also written as $M = [EKN]Px[DE]$.) A protein P *contains* a motif M, denoted $M \subseteq P$, if there exists a continuous segment in P of length k, denoted $a_v a_{(v+1)} \cdots a_{(v+k-1)}$, such that $a_j \in \mathcal{A}_j, v \leq j \leq (v + k - 1)$. PrtnDB is a set of m proteins and is denoted $\{P^i, i = 1, \ldots, m\}$. The sequence dataset $\mathcal{D}$ of n interacting protein pairs is denoted by $\{PPr^i = \{P_1^i, P_2^i\} \mid i = 1, \ldots, n, P_1^i, P_2^i \in \text{PrtnDB}\}$, where P_1^i and P_2^i have interactions.

STABLE MOTIF PAIRS

In this section, we introduce the new concept of *stable motif pairs*. This notion is in light of evolution principles: a stable motif pair is evolved from its neighboring motif pairs, and it should maintain such a status for a long time. We emulate such an evolution using a function f defined in accordance with a widely accepted concept called *consensus discovery*. By this function, only strong residue signals are conserved in heritage but weak ones are filtered out in the transformation of motif pairs.

Given a motif pair, our algorithm for the consensus discovery is to find a *consensus pattern* from the *cluster* of this motif pair.

DEFINITION 1. **[Cluster of a motif pair]** *Let MPr be a motif pair and $\mathcal{D}$ be a sequence dataset of interacting protein pairs. The cluster of MPr in $\mathcal{D}$, denoted $Cluster(MPr, \mathcal{D})$, is a subset of $\mathcal{D}$ such that for every PPr in this subset, PPr contains MPr. That is, $Cluster(MPr, \mathcal{D}) = \{PPr^i \in \mathcal{D} \mid MPr \subseteq PPr^i\}$, or denoted $Cluster(MPr)$ simply when $\mathcal{D}$ is understood.*

where, a protein pair $PPr = \{P_1, P_2\}$ *contains* $MPr = \{M_L, M_R\}$ if $(M_L \subseteq P_1 \wedge M_R \subseteq P_2) \vee (M_L \subseteq P_2 \wedge M_R \subseteq P_1)$. This is also denoted $MPr \subseteq PPr$.

To find the consensus pattern MPr' from the cluster of a given motif pair $MPr = \{M_L, M_R\}$, we use the following method.

1. Split the cluster vertically into two *semi-clusters*: $Cluster_L$ and $Cluster_R$
   $Cluster_L(MPr, \mathcal{D}) = \{P^i \,|P^i \in \{P_1^i, P_2^i\},\ M_L \subseteq P^i, M_R \subseteq (\{P_1^i, P_2^i\} - \{P^i\}), \{P_1^i, P_2^i\} \in \mathcal{D}\}$,
   $Cluster_R(MPr, \mathcal{D}) = \{P^i \,|P^i \in \{P_1^i, P_2^i\},\ M_R \subseteq P^i, M_L \subseteq (\{P_1^i, P_2^i\} - \{P^i\}), \{P_1^i, P_2^i\} \in \mathcal{D}\}$,
2. Align all the occurrences in $Cluster_L$ or in $Cluster_R$ according to the motif M_L or M_R respectively,
3. Find a consensus motif M'_L or M'_R respectively from the two alignments by extracting all those residues in each column of an alignment whose occurrence rate is larger than a threshold (20% in this paper).
4. Combine M'_L and M'_R into a motif pair $MPr' = \{M'_L, M'_R\}$, then it is the transformed motif pair of MPr.

Table 1 gives an example, showing the cluster of a motif pair $\{AGGG[IY], [FV]G[EK][AE][ENS][IL]A\}$ in a sequence dataset $\mathcal{D}$ used in von Mering et al. (2002). Observe that this cluster consists of 7 interacting protein pairs, which is indeed a nonempty subset of $\mathcal{D}$. Columns 2 and 3 of Table 1 list the alignment for the two semi-clusters of the motif pair $\{AGGG[IY], [FV]G[EK][AE][ENS][IL]A\}$. The consensus motifs ($M'_L$ or M'_R) are derived separately from these two semi-clusters. The consensus pattern (the motif pair $\{M'_L, M'_R\}$) is listed at the second last row of the table.

Table 1. The cluster of the motif pair $\{AGGG[IY], [FV]G[EK][AE][ENS][IL]A\}$ and the consensus pattern $\{AGGG[IY], [FV]G[EK]A[ES]IA\}$ derived from this cluster.

Protein 1		Protein 2	
Name	Sequence	Sequence	Name
YKL085W	. . . AGGGI . . .	. . . FGKASIA. . .	YPL004C
YGR204W	. . . AGGGY. . .	. . . FGKASIA. . .	YPL004C
YLL018C	. . . AGGGI. . .	. . . FGKASIA. . .	YPL004C
YGR204W	. . . AGGGY. . .	. . . VGEAEIA. . .	YLR153C
YLL018C	. . . AGGGI. . .	. . . VGEAEIA. . .	YLR153C
YKL085W	. . . AGGGI. . .	. . . VGEAEIA. . .	YLR153C
YKL182W	. . . AGGGY. . .	. . . VGEENLA. . .	YDL052C
	AGGG[IY]	[FV]G[EK]A[ES]IA	
	consensus pattern		

Alternatively, we can use other algorithms such as EMOTIF (Nevill-Manning et al., 1998) to discover consensus patterns.

From the consensus discovery, we can see that MPr' is a transformation of MPr. We use function f to describe this transformation process. Therefore, finding the consensus pattern MPr' from the cluster of a given motif pair MPr can be denoted by $f(MPr) = MPr'$.

DEFINITION 2. **[Stable motif pairs]** *A motif pair* MPr *is stable if*

$$f(MPr) = MPr.$$

Mathematically, this definition follows the Brouwer's Fixed Point Theorem (Mohamed and William, 2001). That is, a thing X, after a transformation f, is still the same thing X, denoted $f(X) = X$. Some basic biological phenomenon can be interpreted as fixed points. For example, the DNA of a cell can be split into two cells with the same DNA after self-replicating where the X is the DNA, and the f is the laws of Physics and Chemistry applied to DNA. As another example, some C2H2 Zinc-Finger genes can be translated into the same type of protein after frameshifts (Meng et al., 2004). Here the X is the protein type, and the f is the frameshifting.

Starting from any motif pair MPr, it is possible to find a stable motif pair. Suppose $f(MPr) = MPr^{(1)}$. Then we apply f to $MPr^{(1)}$ and get $MPr^{(2)}$. Iteratively, it is possible to get $f(MPr^{(i)}) = MPr^{(i)}$. If $MPr^{(i)}$ is nonempty, then it is called a stable motif pair. The whole process is called *refinement*. Table 2 shows an example of such refinement from a starting motif pair. We can prove that the refinement from any motif pair converges to either a stable motif pair or an empty pattern.

Table 2. The refinement from a starting motif pair $\{AG[DGS][GS][IVY], [FV]G[EK][AE][DENS][IL]A\}$. The resulting stable motif pair $\{AGGG[IY], [FV]G[EK]A[ES]IA\}$ is listed at the last row.

Left Motif	Right Motif	Cluster Size
AG[DGS][GS][IVY]	[FV]G[EK][AE][DENS][IL]A	13
AG G [GS][IVY]	[FV]G[EK][AE][ENS][IL]A	9
AG G [GS][I Y]	[FV]G[EK][AE][ENS][IL]A	8
AG G G [I Y]	[FV]G[EK][AE][ENS][IL]A	7
AG G G [I Y]	[FV]G[EK] A [E S] I A	6

SIGNIFICANT MOTIF PAIRS AND THEIR EFFICIENT COMPUTATION

As shown later, not all stable motif pairs are statistically significant. So we introduce *significant motif pairs* in this section to capture more information for binding motif pairs. A significant motif pair requires that the two motifs in the pair must be significant as well. The significance is statistically evaluated against randomness. We begin with definitions for the absolute support and statistical score of single motifs and their efficient computation. Then we explain significant motif pairs and give efficient methods to compute their significance indices.

DEFINITION 3. **[Support for a motif]** *The absolute support of a motif M in* PrtnDB *is the number of proteins in* PrtnDB *that contain M, denoted by $\pi(M, \mathsf{PrtnDB}) = |\{P^i \in \mathsf{PrtnDB} | M \subseteq P^i\}|$, or simply denoted by $\pi(M)$.*

The Z-score measurement is widely used to evaluate the significance of single motifs (Atteson, 1998). The Z-score of a motif M is defined as

$$z_s(M, \mathsf{PrtnDB}) = \frac{\pi(M, \mathsf{PrtnDB}) - exp(M, \mathsf{PrtnDB})}{\sigma(M, \mathsf{PrtnDB})} \quad (1)$$

where $exp(M, \mathsf{PrtnDB})$ is the expectation support for M in PrtnDB, $\sigma(M, \mathsf{PrtnDB})$ is the standard deviation for the random occurrence (support) of M in PrtnDB. With Z-scores, we can distinguish significant motifs from random ones. If the occurrence of a motif is far away from its random expectation, this motif is considered to be statistically significant.

Through the software package provided by Nicodeme et al. (2002), the expectation and deviation for a motif $M = \mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_k$ with respect to PrtnDB can be calculated as follows, where m is the number of proteins in PrtnDB(see *Appendix A* for details):

$$exp(M, \mathsf{PrtnDB}) = \frac{\prod_{i=1}^{k} |\mathcal{A}_i|}{|\Sigma^k|} * (|\mathsf{PrtnDB}| - m * (k-1)) \quad (2)$$

Nicodeme et al. (2002) also showed that for most motifs,

$$\sigma(M, \mathsf{PrtnDB}) \approx \sqrt{exp(M, \mathsf{PrtnDB})} \quad (3)$$

From formula (2) and (3), we can see that after one pass of pre-computation for the number of residues in PrtnDB and the number of proteins m, the expectation and standard deviation of any motif can be calculated in linear time with respect to the number of positions in the motif, *i.e.* in $O(k)$ time.

Next, we introduce the concept of significant motif pairs. Let's first define the support and contributive support of motif pairs with respect to a sequence dataset $\mathcal{D}$ of interacting protein pairs.

DEFINITION 4. **[Support a motif pair]** *The absolute support of a motif pair $MPr = \{M_L, M_R\}$ in $\mathcal{D}$ is defined as the number of interacting protein pairs in $\mathcal{D}$ that contain MPr, denoted by $\pi(MPr, \mathcal{D}) = |\{PPr^i \in \mathcal{D} \mid MPr \subseteq PPr^i\}| = |Cluster(MPr, \mathcal{D})|$.*

Since not all motif pairs contained in an interacting protein pair can play a role for the interaction, we define contributive support of motif pairs to reflect the true contributors for the interaction.

DEFINITION 5. **[Contributive support for a motif pair]** *The contributive support of a motif pair MPr in $\mathcal{D}$ is the number of protein pairs in $\mathcal{D}$ whose interaction is partially contributed by MPr, denoted by $\pi^c(MPr, \mathcal{D}) = |\{PPr^i \in \mathcal{D} | MPr \subseteq PPr^i, MPr\ contributes\ PPr^i\}|$, or simply denoted by $\pi^c(MPr)$.*

Contributive support is only a theoretical concept when structure data for the protein complexes are unavailable. Later on, we will show how to estimate contributive support values based on a sequence dataset $\mathcal{D}$ of interacting protein pairs and a set of motif pairs.

Similarly as Z-scores (Atteson, 1998) used to measure the significance of single motifs with regard to PrtnDB, we define P-scores to measure the significance of motif pairs. Given an $MPr = \{M_L, M_R\}$ and a protein interacting sequence dataset $\mathcal{D}$,

$$p_s(MPr, \mathcal{D}) = \frac{\pi^c(MPr, \mathcal{D})}{exp(MPr, \mathcal{D})} \quad (4)$$

where $exp(MPr, \mathcal{D})$ is expectation support of random co-occurrences of MPr in $\mathcal{D}$.

Based on the Z-scores of single motifs and P-scores of motif pairs, now we define significant motif pairs:

DEFINITION 6. **[Significant motif pairs]** *A motif pair $MPr = \{M_L, M_R\}$ is significant in a protein interacting sequence dataset $\mathcal{D}$ and the corresponding protein set* PrtnDB *if $z_s(M_L, \mathsf{PrtnDB}) \geq \tau_L$, $z_s(M_R, \mathsf{PrtnDB}) \geq \tau_R$, and $p_s(MPr, \mathcal{D}) \geq \tau_B$, where $\tau_L \geq 0, \tau_R \geq 0, \tau_B \geq 1$ are pre-set thresholds.*

This definition emphasizes that the observations should be far away from the expectation values.

Computationally calculating P-scores is not straightforward because the accurate contributive support is almost impossible to be obtained without wet-experimental examination. So, we present an approximate solution. First, assume M_L and M_R are independent, the expectation can be calculated as follows:

$$exp(MPr, \mathcal{D}) = n * \frac{\pi(M_L)}{m} * \frac{\pi(M_R)}{m} \quad (5)$$

where m is the number of unique proteins in $\mathcal{D}$. Therefore, the P-score can be re-written as

$$p_s(MPr, \mathcal{D}) = \frac{m^2 * \pi^c(MPr)}{n * \pi(M_L) * \pi(M_R)} \quad (6)$$

Assume an interaction contains only one binding motif pair, then the contribution of a motif pair to a protein pair is influenced by other motif pairs. Given a sufficiently large set of motif pairs S_{MPr}, we can estimate the contributive support using the following

$$\begin{aligned} \pi^c(MPr) &= \lim_{|S_{MPr}| \to \infty} \sum_{i=1}^{n} \frac{p_s(MPr, \mathcal{D})}{\sum_{MPr' \subseteq PPr^i} p_s(MPr', \mathcal{D})} \delta_i(MPr) \\ \delta_i(MPr) &= \begin{cases} 1 & if\ MPr \subseteq PPr^i \\ 0 & otherwise \end{cases} \end{aligned} \quad (7)$$

It can be seen that for a motif pair, the supports of its two contained motifs are fixed values in a given protein set PrtnDB. So, when handling a large motif pair set S_{MPr}, formula (6) and (7) will consist of a large group of equations with two types of variables: the P-scores and the contributive supports of the motif pairs.

To solve this group of equations, we explore the use of iterative programming. First we set an identical initial value for the P-score of every motif pair. Then we use the current P-scores to calculate the contributive support for all motif pairs by formula (7). We can thereafter get new P-scores using formula (6) for each motif pair and start a new round of calculation, until the changes of most variables are less than a threshold.

DETERMINE STARTING POINTS TO DERIVE STABLE MOTIF PAIRS

Since the number of possible starting motif pairs is huge, it is a computational difficult problem to find all stable motif pairs from a large dataset of interacting protein pairs. In this section, we present a heuristic method to find a subset of the stable motif pairs with the biological guidance from a protein complex dataset. Our motivation is that the protein complex data is the most reliable data about protein interactions, and its 3-D co-ordinate information is an easy platform to find binding sites. Hereby, we first compute binding sites from the complex dataset, and then use them to produce starting motif pairs to search for stable and significant motif pairs from the dataset of interacting protein sequence pairs. By this way, we can get high confidence to the discovered stable and significant motif pairs since they are stemmed from the biologically reliable protein complex data.

A core step to determine starting motif pairs is to discover the so called *maximal contact segment pairs* (Li et al., 2004) from a protein complex dataset. Let's explain a bit more about this concept. Two segments from different proteins are a *contact segment pair* if any residue in one segment can find at least one contact residue in the opposite segment, where the contact of two residues means that at least one of their atom pairs has an Euclidean distance less than a threshold. A contact segment pair is a maximal contact segment pair if no any other contact segment pair in the same protein pair contains both segments of this contact segment pair, capturing contact segment pairs as lengthy as possible. These definitions and the search algorithms can be found in our previous work (Li et al., 2004). To be self-contained for this paper, we also outline these in Appendix B. As an example (see more in *Figure S1* and *Figure S2* of the *supplementary information*), the segment pair $([a_{16}, a_{20}], [d_{41}, d_{47}])$ with sequence (AGSSY, VGRANMA) between chain A and chain D of the complex *pdb1mbm* is a maximal contact segment pair.

However, directly using maximal contact segment pairs as starting motif pairs is not a smart choice. Because these segment pairs are highly specific in corresponding species, they may not occur in yeast interacting protein dataset $\mathcal{D}$. So, we need to *generalize* these contact segment pairs. We achieve this goal by using the principle proposed in Azarya-Sprinzak et al. (1997). The principle says that even some residues in some positions are changed to other residues, their structures are still unchanged. Since the structures maintain the same, the binding behavior is highly likely to maintain as well. Basically, we use local alignment and consensus discovery to implement this generalization and to get satisfactory starting motif pairs.

Given a maximal contact segment pair SPr and a protein interaction dataset $\mathcal{D}$, the generalization of SPr is as follows:

1. Find a subset of $\mathcal{D}$, denoted $ApproxCluster(SPr) = \{PPr^i \in \mathcal{D} \mid Local\_Alignment(SPr, PPr^i) \geq \lambda\}$, where λ is an empirical threshold,
2. Discover the consensus pattern MPr from $ApproxCluster(SPr)$.

Thus, MPr is a generalized pattern for SPr. Then we use MPr as a starting point to discover a stable motif pair. For the maximal segment pair $(AGSSY, VGRANMA)$ mentioned above, we found 34 interactions for its ApproxCluster. From this cluster, we induced a consensus motif pair, $\{AG[DGS][GS][IVY], [FV]G[EK][AE][DENS][IL]A\}$, which was then used as the starting point to derive a stable motif pair $\{AGGG[IY], [FV]G[EK]A[ES]IA\}$.

Summary of the whole flow to discover a subset of stable and significant motif pairs

The whole flow of our method is summarized as follows:

```
Input: A sequence dataset $\mathcal{D}$ of interacting protein pairs,
  a complex dataset $\mathcal{C}$
Output: A set of stable and significant motif pairs $S_{MPr}$
  for all complex $CPL$ in $\mathcal{C}$ do
    for all protein pair $P_a$ and $P_b$ in $CPL$ do
      find the set of maximal contact segment pairs $S_{SPr}$;
    end for
  end for
  for all contact segment pair $SPr$ in $S_{SPr}$ do
    generalize $SPr$ to produce a starting motif pair $MPr$
  end for
  for all  starting motif pair $MPr$ do
    refine $MPr$ to either a stable motif pair $MPr'$ or an
    emptyset.
  end for
  for all stable motif pair $MPr'$ do
    filter those stable motif pairs $MPr'$ if $MPr'$ is not
    significant
  end for
```

IMPLEMENTATIONS AND RESULTS

In the computation of contact residues in a complex, we set the distance threshold as $5\mathring{A}$, that is, any residue/atom pair which have a distance less than $5\mathring{A}$ is regarded to be contacted. In the computation of maximal contact segment pairs, we required that every contact segment should contain at least 4 residues. In the generalization from maximal contact segment pairs to starting motif pairs, we set different λ thresholds for local alignment based on the segment lengths: λ was set strictly for short segments but loosely for long segments. Actual λ values used in this study can be referred to *Figure S3* of the *supplementary information*.

After obtaining starting motif pairs from the complex dataset, we conducted the refinement process to find stable motif pairs from the sequence dataset of interacting protein pairs. For a motif pair MPr, to discover $f(MPr)$—the consensus pattern—and subsequently $f(\cdots f(f(MPr)))$ until a stable state, we computed a latter cluster based on its previous cluster instead of the whole dataset. The efficiency was therefore greatly improved. This is correct because the refinement leads to more and more specific motif pairs.

After obtaining a set of stable motif pairs from the starting motif pairs and the refinement, we filtered the insignificant ones. The thresholds for the significance indices were set as: $\tau_L = 0, \tau_R = 0, \tau_B = 1$. The computation of the supports and Z-scores are straightforward according to our algorithm. However, the computation of the P-scores is an iterative process. The initial P-score for every motif pair was set as 1.0 in this work. We observe that the P-score trends of most motif pairs ($> 90\%$) in the iterative process are convergent, either monotonically increasing or monotonically decreasing. Given a set of motif pairs S_{MPr}, the overall score difference between the j-th and the $(j-1)$-th iteration is calculated by an index $\Delta(j)$ [†]. If $\Delta < 0.01$, we stop the iterative process. For most sets of motif pairs, the process could stop within 4 iterations.

Results overview

In total, we discovered 765 stable and significant motif pairs from the sequence dataset of interacting protein pairs using 1403 maximal contact segment pairs identified from the protein complex dataset. See Table 3 for these results and other related results such as the support information.

Table 3. Our results in overview.

Num of Contact Segment Pairs	Num of Starting Motif Pairs	Num of Stable Motif Pairs	Num of Significant Stable Motif Pairs	Support of Stable Motif Pairs	Support of Significant Motif Pairs
1403	1222	913	765	122193	107028

The P-sore values of the 765 stable and significant motif pairs differ very much from one another. Figure 2 shows the distribution of these P-scores (under $\log_2$ scale). It can be seen that our algorithm can discover motif pairs with both high and low P-scores (larger than a threshold).

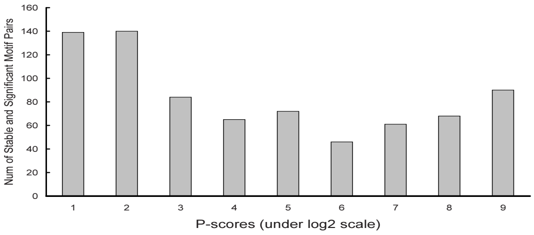

Fig. 2. The distribution of the P-scores (under $\log_2$) for the 765 stable and significant motif pairs.

[†]

$$\Delta(j) = \frac{2 * \sum\limits_{MPr^i \in S_{MPr}} (p_s(MPr^i, \mathcal{D})_j - p_s(MPr^i, \mathcal{D})_{j-1})^2}{\sum\limits_{MPr^i \in S_{MPr}} (p_s(MPr^i, \mathcal{D})_j)^2 + (p_s(MPr^i, \mathcal{D})_{j-1})^2} \quad (8)$$

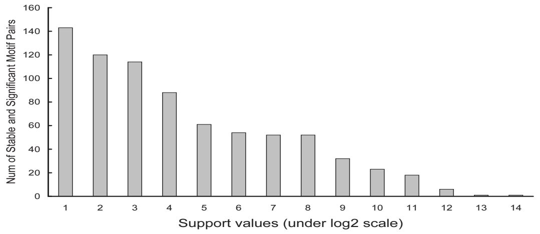

Fig. 3. The distribution of the absolute support values (under $\log_2$ scale) of the 765 stable and significant motif pairs.

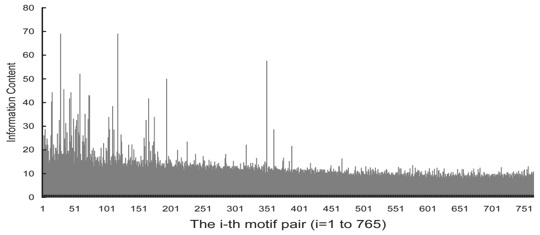

Fig. 4. The distribution of information content of our discovered stable and significant motif pairs.

Besides P-scores, another important information is support. The distribution of the support values (under $\log_2$ scale) of the 765 stable and significant motif pairs is depicted in Figure 3. It can be seen that our algorithm preferred to discovering motif pairs with relatively low supports. This is an advantage of our algorithm as the support of many real binding motif pairs is quite possible to be low in an incomplete dataset. The distribution of the estimated contributive support values for our discovered motif pairs exhibits almost the same shape as the one in Figure 3.

To evaluate the lengths of our discovered motif pairs, we used *information content* (Tompa, 1999) as the index. Assume each residue has equal distribution, the information content of a motif $M = \mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_k$ can be computed by:

$$I(M) = k \log_{10} |\Sigma| - \sum_{i=1}^{k} \log_{10} |\mathcal{A}_i| \qquad (9)$$

For a motif pair $MPr = \{M_L, M_R\}$, we define

$$I(MPr) = I(M_L) + I(M_R) \qquad (10)$$

So, the information content largely reflects the length of a motif. The distribution of the information contents of the 765 motif pairs is presented in Figure 4. It can be seen that most of the motif pairs have an information content between 10 and 20, except for very few cases. So, these motif pairs roughly have residues between 10 to 20.

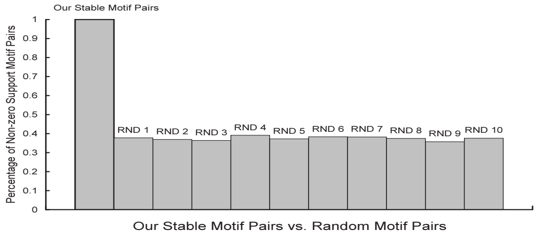

Fig. 5. The percentage of non-zero support motif pairs in our discovered stable motif pairs and those in 10 sets of equal size of random motif pairs.

Effectiveness comparison with random patterns

To demonstrate our discovered stable and significant motif pairs are credible, and also to illustrate that our choice of the starting motif pairs makes benefits to the discovery, we conduct a comprehensive computational comparison between our patterns and random patterns. These experiments include: (1) the comparison between our 913 stable motif pairs versus 10 random sets each consisting of 913 random motif pairs; (2) the comparison between our 1222 starting motif pairs versus 10 random sets each consisting of 1222 random starting motif pairs.

A random motif pair is generated by substituting every residue in our pattern with a random residue. So, the random pattern has the same length as ours. The distribution of the randomly generated residues follows the same distribution of all the residues in the contact sites of our complex dataset. (In fact, it has no significant difference between this distribution and that in the whole yeast genome (Fariselli et al., 2002)).

First, we compare our 913 stable motif pairs with the 10 sets of random motif pairs of equal size to see how much percentage of them are significant. We observed that

- About two thirds of the random motif pairs have a zero-support in the interaction dataset $\mathcal{D}$, namely $\pi(MPr^{random}, \mathcal{D}) = 0$. However, for every MPr of our 913 stable motif pairs, $\pi(MPr, \mathcal{D}) \neq 0$. Figure 5 shows the percentage of random patterns having non-zero support for the 10 rounds of random experiments.

- Only about one ninth of the random motif pairs are significant. However, about 84% of our 913 stable motif pairs are significant. Full results can be seen in Figure 6.

- The total support of our stable and significant motif pairs is much larger than that of significant random motif pairs, which is shown in *Figure S4* of the *supplementary information*.

These results indicate that our discovered stable motif

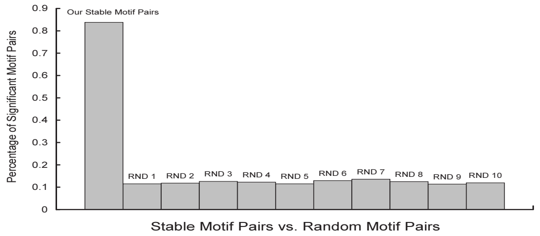

Fig. 6. The percentage of significant motif pairs for our discovered stable motif pairs and those for 10 sets of equal size of random motif pairs.

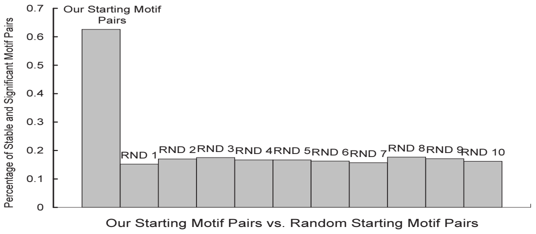

Fig. 8. The percentage of stable and significant motif pairs derived from our starting motif pairs and those derived from 10 sets of equal size of random starting motif pairs.

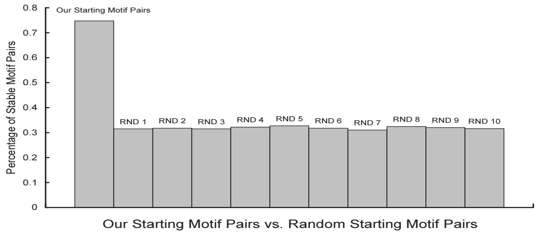

Fig. 7. The percentage of stable motif pairs derived from our starting motif pairs and those derived from 10 sets of equal size of random starting motif pairs.

pairs are much more statistically significant than random patterns. Therefore, they are most likely to be potential binding motif pairs.

Secondly, we substitute our 1222 starting motif pairs with random starting motif pairs to see how much percentage of stable motif pairs can be discovered, and how much percentage of stable and significant can be discovered. Such substitution is repeated for ten times. We observed that

- Our starting motif pairs can lead to 75%(913) of stable points, but those random starting points in each round lead to less than 33% of stable motif pairs. Full results can be seen in Figure 7.
- Our starting motif pairs can lead to about 63% of stable and significant motif pairs, but less than 18% of those random starting points can lead to stable and significant motif pairs. See Figure 8 for full results.

From these comparison, we can conjecture that the generalization from maximal contact segment pairs to our starting motif pairs is a useful method because it contributes much more number of stable and significant motif pairs than the random method does.

From these various random experiments, we can see that the stable and significant motif pairs that we discovered are far way from random expectation, which benefits from the choice of starting points. Therefore, it is reasonable that they provide much information to find real binding motif pairs. This is also confirmed by our literature searching results reported in the next subsection.

Literature validation

To demonstrate the biological significance of our discovered patterns, ideally, they should be validated by wet-experimental methods. Unfortunately, there are few well-known wet-experimental methods which can determine the two sides of the binding sites simultaneously. Current available technique such as phage display (Smith, 1985) can determine only one side of the binding sites and produce protein-motif binding pairs or protein-peptide binding pairs. On the other hand, there is still limited data about binding sites, mostly spanning across various individual literature, without an integrative and comprehensive database available, which makes our validation even harder.

Nevertheless, we still find some evidences to show the biological significance of our discovered patterns. First, we check the coincidence of the *individual* motifs in our motif pairs with the reported binding motifs determined by various wet-experimental methods. For example, using key words ‘binding motif OR site AND mutagenesis’, we extracted 202 binding motifs from the abstracts of NCBI PUBMED; 89 of them have at least 3 positions compatible to ours and 40% overall similarity. Of these 89 binding motif pairs, 42 motif pairs are highly similar with our discovered motifs, having at least 4 positions compatible and 50% overall similarity. We show the top 5 matches in Table 4. Similar examples comparing with the phage display method is provided in *Table S1* of the *supplementary information*.

Secondly, we check our discovered motif pairs with protein-motif binding pairs determined by phage display. First, we identify the individual motifs in our population of discovered motif pairs that match closely with a binding motif/peptide in the literature. Then, for each of such matched motifs, we verify whether the motif on the other side of the corresponding motif pairs can be found in the

Table 4. Motif coincidence with the mutagenesis method.

Our Motif	Mutagenesis Motif	PMID of Mutagenesis Motif
GSGKT	GxGKT	10464259
ALETS	LETS	11435317
P[IV]DL	PVDLS	11373277
L[DN]LL	LLDLL	11451993
K[DE]K[EK]	KEKE	10748065

proteins known to bind the particular motif/peptide. An example is shown in Table 5. Tumbarello et al. (2002) studied the binding sites of protein paxillin and its binding proteins. The binding site of paxillin is in the form of LDxLLxxL. Our method discovered similar motifs as shown in the first column of Table 5. The other side of the corresponding motif pairs are shown in the second column of the table, which have been found to exist in the binding proteins reported in the literature (Tumbarello et al., 2002). The fully matched binding proteins or roughly matched motifs are shown in the last column of the table. More examples are detailed in *Table S2* and *Table S3* of the *supplementary information*.

Table 5. The coincidence between our discovered motif pairs and the binding sites of paxillin and its binding proteins, where the binding site of paxillin is motif LDxLLxxL.

Left motif	Right Motif	Confirmed Proteins
D[IL]L[IL]	[ST]D[EK]A	Vinculin,FAK
[IL][DG][IV]LD	D[EK]EGI	PYK2(D[EK]EG)

L[FL]VLK	L[FL]VLK	Vinculin(L[FL]VL)
		PYK2(L[FL]VL)

Finally, we give full details of one of the 765 stable and significant motif pairs to see how it is discovered, where is its origin, and what is its biological significance. This stable motif pair is

$$\{L[DN]LL, [EK][LV]GDG\}$$

denoted by $MPr_{example} = \{M_L, M_R\}$, where $M_L = L[DN]LL$ and $M_2 = [EK][LV]GDG$.

Its origin is located at the so-called pdb3daa protein complex. Specifically, the motif $M_L = L[DN]LL$ is evolved from the the segment $LNLL$ at the chain A of the pdb3daa complex. These four amino acids are indexed from 147th to 150th residues in the chain A, denoted by $[a_{147}, a_{150}]$ with sequence $LNLL$.

The motif $M_R = [EK][LV]GDG$ is rooted at the segment $YQFGDG$ at the chain B of the pdb3daa complex. These six amino acids are indexed from 24th to 29th residues in the chain B, denoted by $[b_{24}, b_{29}]$ with sequence $YQFGDG$.

The segment pair, $([a_{147}, a_{150}], [b_{24}, b_{29}])$ with sequence (LNLL,YQFGDG) between chain A and chain B, is a maximal contact segment pair.

This maximal contact segment pair $(LNLL, YQFGDG)$ is then generalized to the following starting motif pair MPr_{start}

$$MPr_{start} = \{L[DN]LL, [EK][LV]GDG\}$$

for the function f.

Interestingly, we found that $f(MPr_{start}) = MPr_{start} = MPr_{example}$. That is, this starting motif pair MPr_{start} itself is a stable motif pair.

We found that this stable motif pair $MPr_{example}$ is statistically significant after examining its support level and P-score against random motif pairs. The support of motif $L[DN]LL$ is 265 in PrtnDB, the support of motif $[EK][LV]GDG$ is 13 with respect to the same protein set PrtnDB. The support of $MPr_{example}$ as a pair is 58 in the protein interaction sequence data set $\mathcal{D}$. Then, we generated 1000 random motif pairs according to $MPr_{example}$, where each random motif pair is generated by substituting every residue in $MPr_{example}$ with a random residue. So, the random motif pairs have the same length as $MPr_{example}$. The distribution of the randomly generated residues follows the same distribution of all the residues in the whole yeast genome. For these 1000 random motif pairs, the average support of the random motifs corresponding to $L[DN]LL$ is 32.91, the average support of the random motifs corresponding to $[EK][LV]GDG$ is 4.41. The average support for those 1000 motif pairs is 1.83 in the protein interaction sequence data set $\mathcal{D}$. The P-score of $MPr_{example}$ as a pair is 6.15 with respect to protein interaction sequence data set $\mathcal{D}$, while the average P-score for these 1000 random motif pairs is 2.63 with respect to the same $\mathcal{D}$. From these statistical numbers of $MPr_{example}$ and its equal-length 1000 random motif pairs, we can see that $MPr_{example}$ has occurrence much more than its random expectation either in single motifs or in pairs. So, the stable motif pair $MPr_{example}$ is not a random result indeed.

We also found many biological significance of the motif pair $MPr_{example}$. In biology, Doray and Kornfeld (2001) found a protein motif $M_{DK} = LLDLL$, a functional variant of the $LLNLD$ motif within the beta 1 subunit of AP-1, was biologically confirmed to bind to the terminal domain of the clathrin heavy chain. From the sequence of this terminal domain, we find that there exists a segment $ELGD$ near the end part of this domain. Comparing these biological results and our computational results, we can see that

- $M_{DK} = LLDLL$ is similar to the left motif $L[DN]LL$ of our motif pair $MPr_{example}$.
- The segment $ELGD$ matches well with our right motif $[EK][LV]GDG$ of $MPr_{example}$. The precise position of the segment $ELGD$ is from positions 462 to 465 at the end of the globular terminal domain (from 1th to 479th) of clathrin heavy chain 1 of human.
- Besides, our left motif $L[DN]LL$ is similar to $LLDLL$ and $LLNLD$ both of which share the same functions.

DISCUSSIONS AND SUMMARY

In this paper, we model binding sites using a mathematical concept—stable patterns. Our random experiments have shown that stable motif pairs are more likely to be significant than random motif pairs. It is interesting to examine the theoretical aspect of this concept in future.

In this paper, we use P-scores to evaluate the significance of motif pairs. In fact, more complicated score schemes can be tried. For example, Ng et al. (2003) used a similar score with a slight difference only in the calculation of expectations. Deng et al. (2002) used maximum likelihood to estimate the scores. Note that these two approaches are quite expensive in computation. Moreover, Ng's formula also has the divergence problem. So, both of them are not suitable to search significant motif pairs in a huge pattern space. Therefore, how to combine the strength of these score schemes is still a future research effort of us.

We validate our discovered motif pairs with those determined by experimental methods from literature, both for individual motifs and motif pairs. We show some examples for the coincidence between them. Nevertheless, there still a lot efforts to make in the future. We intend to collect a comprehensive database about experimental determined binding motif pairs through text mining methods in addition to manual check. Then we could perform a systematic validation for our discovered patterns.

Finally, we summarize the main results achieved in this paper. We used motif pairs to model the binding sites between proteins with two intuitions: (1) the motif pairs should satisfy the stability; (2) The motif pairs should be statistically significant, for both single motifs and their co-occurrence as pairs. We presented efficient algorithms to identify meaningful starting motif pairs, and to find a convergence route for stable motif pairs, as well as to compute the significance of motif pairs. As the search for all possible stable and significant motif pairs from a sequence dataset of interacting protein pairs is a challenging problem, in this paper we turned to look for a special subset of them. The discovery of this subset of stable and significant motif pairs is guided by binding sites identified from a biologically reliable dataset of protein complexes. For this, we extract maximal contact segment pairs from the complexes dataset, then generalize them to become our crucial patterns—starting motif pairs that lead to stable motif pairs by a refinement process.

Our comprehensive comparison results have shown that our discovered motif pairs are much more statistically significant than random motif pairs, a result from the choice of starting motif pairs. Some of our discovered motif pairs are also highly matched with real binding motifs reported in literature.

ACKNOWLEDGMENTS

We are grateful to Pierre Nicodeme, Frederic Chyzak and Bruno Salvy for providing the source code and library of the powerful software package *Algolib* to let us compute the statistical scores of motif pairs. We are thankful for the time of Soon-Heng Tan for helping us validate the patterns in comparison with literature results.

REFERENCES

Atteson, K. (1998). Calculating the exact probability of language-like patterns in biomolecular sequences. In *Proceeding of he 6th International Conference on Intelligent Systems for Molecular Biology(ISMB)*, pp. 17–24.

Azarya-Sprinzak, E., D. Naor, H. J. Wolfson, and R. Nussinov (1997). Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Eng 10*(10), 1109–22.

Botstein, D. and D. Shortle (1985). Strategies and applications of in vitro mutagenesis. *Science* (4719), 1193–1201.

Clemmons, D. R. (2001). Use of mutagenesis to probe igf-binding protein structure/function relationships. *Endocr Rev 22*(6), 800–17.

Deng, M., S. Mehta, F. Sun, and T. Chen (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res 12*(10), 1540–8.

Doray, B. and S. Kornfeld (2001). Gamma subunit of the ap-1 adaptor complex binds clathrin: implications for cooperative binding in coated vesicle assembly. *Mol Biol Cell 12*(7), 1925–35.

Fariselli, P., F. Pazos, A. Valencia, and R. Casadio (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem 269*, 1356–1361.

Josephson, K., N. J. Logsdon, and M. R. Walter (2001). Crystal structure of the il-10/il-10r1 complex reveals a shared receptor binding site. *Immunity 15*(1), 35–46.

Li, H., J. Li, S. H. Tan, and S. K. Ng (2004). Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. In *Proceeding of the Ninth Pacific Symposium on Biocomputing (PSB)*, Hawaii USA, pp. 312–323.

Meng, S. W., Z. Zhang, and J. Li (2004). Twelve c2h2 zinc finger genes on human chromesone 19 can be each translated into the same type of protein after frameshifts. *Bioinformatics 20*, 1–4.

Mohamed, A. K. and A. K. William (2001). *An introduction to metric spaces and fixed point theory*. John Wiley & Sons.

Nevill-Manning, C. G., T. D. Wu, and D. L. Brutlag (1998). Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci 95*(11), 5865–71.

Ng, S. K., Z. Zhang, and S. H. Tan (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics 19*(8), 923–9.

Nicodeme, P., B. Salvy, and P. Flajolet (2002). Motif statistics. *Theoretical Computer Science 287*, 593–618.

Rodi, D. J., G. E. Agoston, and et al. (2001). Identification of small molecule binding sites within proteins using phage display technology. *Comb Chem High Throughput Screen 4*(7), 553–72.

Sidhu, S. S., W. J. Fairbrother, and K. Deshayes (2003). Exploring protein-protein interactions with phage display. *Chembiochem 4*(1), 14–25.

Smith, G. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science 228*(4705), 1315–7.

Sprinzak, E. and H. Margalit (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol 311*(4), 681–92.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology(ISMB)*, pp. 262–271.

Tumbarello, D. A., M. C. Brown, and C. E. Turner (2002). The paxillin ld motifs. *FEBS Lett 513*(1), 114–8.

von Mering, C., R. Krause, and et al. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature 417*(6887), 399–403.

APPENDIX A

For a motif $M = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, the expectation in overlapping model is as follows:

$$\begin{aligned}
p(M) &= \prod_{i=1}^{k} \frac{|\mathcal{A}_i|}{|\Sigma|} \\
&= \frac{\prod_{i=1}^{k} |\mathcal{A}_i|}{|\Sigma^k|} \\
exp(M, \mathsf{PrtnDB}) &= p(M) * \sum_{i=1}^{m}(|P^i| - k + 1) \\
&= p(M) * (\sum_{i=1}^{m} |P^i| - m * (k-1)) \\
&= p(M) * (|\mathsf{PrtnDB}| - m * (k-1))
\end{aligned} \tag{11}$$

APPENDIX B

DEFINITION 7. **[Contact segment pairs]** *Given two proteins $P_a = (a_1, \dots a_i, \dots a_u)$ and $P_b = (b_1, \dots b_j, \dots b_v)$, where a_i and b_j are corresponding residue ids on its protein, a segment pair $([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}])$ is a contact segment pair if $\forall a_i \in [a_{i_1}, a_{i_2}]$, $\exists b_j \in [b_{j_1}, b_{j_2}]$ such that contact(a_i, b_j), and $\forall b_j \in [b_{j_1}, b_{j_2}]$, $\exists a_i \in [a_{i_1}, a_{i_2}]$ such that contact(a_i, b_j). Residue a_i and b_j is contacted if one of their atom pairs having Euclidean distance less than a threshold.*

The following proposition is useful for the efficient discovery of all maximal contact segment pairs from a complex dataset.

PROPOSITION 1. **[Containing property]** *A segment pair $([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}])$ in protein P_a and P_b is a contact segment pair iff the coverage of any of the two segments contains the other segment, i.e. $Contact([a_{i_1}, a_{i_2}], [b_{j_1}, b_{j_2}]) \iff (Cov([a_{i_1}, a_{i_2}]) \supseteq [b_{j_1}, b_{j_2}]) \wedge (Cov([b_{j_1}, b_{j_2}]) \supseteq [a_{i_1}, a_{i_2}])$.*

where Cov is short for coverage, $Cov(a_i) = \{b_j \mid contact(a_i, b_j), b_j \in P_b\}$, and $Cov([a_{i_1}, a_{i_2}]) = \bigcup_{a_i \in [a_{i_1}, a_{i_2}]} Cov(a_i)$; $Cov(b_j)$ and $Cov([b_{j_1}, b_{j_2}])$ are similarly defined.

We use a top-down divide-and-conquer strategy to make use of this proposition for the discovery of all the maximal contact segment pairs. First, we start with the two entire segments, then we check whether the containing property exists between these two segments. If yes, stop. Otherwise, we split the coverage of one segment into several discontinued sub-segments and form several new segment pairs. This process goes recursively until all the segment pairs fulfill the containing property. By this way, we also guarantee that we only target on maximal contact segment pairs. The detailed proof of the proposition and the formal description of the algorithm can be found in our previous work (Li et al., 2004).