BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale

Haiquan Li [a,b], Jinyan Li [a], Limsoon Wong [a,b]

[a]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613, [b]School of Computing, National University of Singapore, Singapore, 117543

ABSTRACT

Motivation: Protein–protein interaction, mediated by protein interaction sites, is intrinsic to many functional processes in the cell. In this paper, we propose a novel method to discover patterns in protein interaction sites. We observed from protein interaction networks that there exist a kind of significant substructures called *interacting protein group pairs*, which exhibit an *all-versus-all* interaction between the two protein-sets in such a pair. The full-interaction between the pair indicates a common interaction mechanism shared by the proteins in the pair, which can be referred as an interaction type. Motif pairs at the interaction sites of the protein group pairs can be used to represent such interaction type, with each motif derived from the sequences of a protein group by standard motif discovery algorithms. The systematic discovery of all pairs of interacting protein groups from large protein interaction networks is a computationally challenging problem. By a careful and sophisticated problem transformation, the problem is solved by using efficient algorithms for mining frequent patterns, a problem extensively studied in data mining.

Results: We found 5349 pairs of interacting protein groups from a yeast interaction dataset. The expected value of sequence identity within the groups is only 7.48%, indicating non-homology within these protein groups. We derived 5343 motif pairs from these group pairs, represented in the form of blocks. Comparing our motifs with domains in the BLOCKS and PRINTS databases, we found that our blocks could be mapped to an average of 3.08 correlated blocks in these two databases. The mapped blocks occur 4221 out of total 6794 domains (protein groups) in these two databases. Comparing our motif pairs with iPfam consisting of 3045 interacting domain pairs derived from PDB, we found 47 matches occurring in 105 distinct PDB complexes. Comparing with another putative domain interaction database InterDom, we found 203 matches.

Availability: `http://research.i2r.a-star.edu.sg/BindingMotifPairs/resources`

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

Supplementary information: `http://research.i2r.a-star.edu.sg/BindingMotifPairs`

1 INTRODUCTION

Protein-protein interactions carry out many biological processes in the cells such as gene expressions, signal transduction and intercellular communication. Protein interactions are usually mediated by short sequences of residues, which form the contact interfaces between two interacting proteins, referred to as interaction sites (Sheu et al., 2005). These interaction sites are often geometrically complementary and electric-statically compatible (Jones and Thornton, 1996). They are also highly conserved (Keskin et al., 2004, 2005) and co-evolved (Pazos et al., 1997) and only limited interaction templates exist, which are termed as interaction types (Aloy and Russell, 2004). Unraveling these interaction sites is helpful for understanding the mechanism of protein recognition and protein function, and is beneficial to the design of drug-aimed protein-protein interactions (Loregian and Palu, 2005).

Protein interaction sites can be determined by various experimental methods, including X-Ray crystallographic screening (Garman et al., 2000), NMR-based methods (Swanson et al., 1995; Takahashi et al., 2000), site-directed mutagenesis (Clemmons, 2001) and phage display (DeLano et al., 2000). Experimental methods are generally laborious and expensive. Consequently, only a small number of interaction types have been determined so far. It was estimated that it would take more than 20 years to accomplish all interaction types by using current experimental techniques (Aloy and Russell, 2004).

On the other hand, computational methods play an important role in the determination of interaction sites due to their low cost. Recently, protein-protein docking, which predicts the structures of protein complexes based on solved or modeled structures of the component proteins (Terwilliger, 2004), has made significant progress since the proposal of CAPRI assessment in 2001 (Mendez et al., 2005). Protein interaction sites can be pinpointed during the course of docking. However, about 40% of proteins cannot be modeled for putative structures (Aloy et al., 2005). This leaves a critical gap in this docking approach.

Another approach is based on the conservation characteristics of interaction sites among homologous sequences, also referred to as binding motif discovery algorithms such as PROTOMAT (Henikoff and Heinikoff, 1991) and MEME (Bailey and Elkan, 1995). Correlations between the binding motifs can be measured by an expectation maximization (EM) model as shown by Wang et al. (2005). The intrinsic deficiency of this approach lies in the difficulty to distinguish the folding and binding motifs as binding and folding are both interrelated (Kumar et al., 2000). The third category of computational methods are machine learning oriented approaches. The features utilized in the learning are some known characteristics about interaction sites such as hydrophobicity (Gallet et al., 2000), the sequence segments (Ofran and Rost, 2003) or the spatial patches (Jones and Thornton, 1997). SVMs (Yan et al., 2004) and neural networks (Zhou and Shan, 2001) are two commonly used machine learning methods. Drawbacks in this approach include the difficulty in finding discriminating features and the unattractive performance in accuracy. Overall, current computational methods for interaction site prediction are far from perfect.

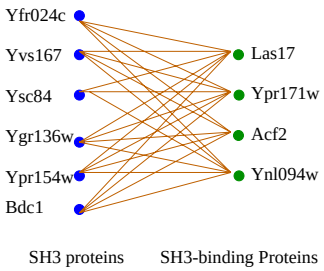

Fig. 1. An all-versus-all predicted interaction subnetwork (most are confirmed by experiments) consisting of two groups of proteins, where one group contains 6 proteins with SH3 domains and the other contains 4 proteins with SH3-binding motifs. (Tong et al., 2002).

In this paper, we propose a novel approach to the discovery of interaction sites on a proteome-wide scale. This approach uses only protein interaction data and the associated sequence data. As mentioned earlier, interaction sites are highly conserved (Keskin et al., 2005). Conserved interaction sites are favorable interfacial scaffolds that have been repeatedly used in the evolution process by proteins with different sequence, structure and function (Keskin and Nussinov, 2005). An example can be seen from *cipa* (PDB code 1aoh) and *Dsred* (PDB code 1g7k), two complexes which have similar interfaces between their component chains A and B (Keskin et al., 2004), but which have dissimilar global structures and functions. (See supplementary information.) A set of conserved interaction sites corresponds to an interaction type (Aloy and Russell, 2004) as they share some common binding mechanism. Whenever the interaction type occurs in a novel protein pair regardless of their homology, the two proteins are likely to interact—This principle has been used by Tong et al. (2002) and Aytuna et al. (2005) to predict protein interactions with an acceptable performance. Such an interaction type implies a most-versus-most and even an *all-versus-all* interaction subnetwork between two groups of proteins in a protein network, with each protein group corresponding to one side of the interaction type. Figure 1 shows an example of such a subnetwork (Tong et al., 2002).

Interestingly, if a large enough subnetwork with all-versus-all interactions between two protein groups is found in a protein network, an interaction type with conserved interaction sites can be predicted. That is because most proteins only contain a small number of interaction sites (usually, $2 \sim 6$ for typical proteins (Liang et al., 1998)). Due to the constraints of all-versus-all interactions between these two groups, it is expected that there exists two groups of interaction sites from these two protein groups which interact with each other for at least some occurrences. The interaction sites within the same group should hold similar structures and possibly have a sequence motif as they have similar interaction partners. These two groups of interaction sites and their corresponding motifs can be easily identified using standard motif discovery methods from the sequence data of the corresponding protein group. Then, an interacting motif pair (Li et al., 2004) is formed, with which to represent the corresponding interaction sites of the interaction type.

We term the above two protein groups that exhibit an all-versus-all interaction as a pair of *interacting protein groups*. It is a challenging problem to discover all pairs of interacting protein groups from a proteome-wide protein interaction network by a naive way as the number of combinations of proteins is exponential. However, we found that this problem of mining interacting protein groups can be transformed into the classical problem of mining *frequent patterns* (Agrawal and Srikant, 1994). As frequent pattern mining has been extensively studied in the data mining field, many existing algorithms can be directly used to efficiently find all pairs of frequent interacting protein groups from large datasets of protein interactions.

To assess the performance of our proposed method for mining motif pairs from a large yeast interaction dataset, we propose a systematic validation experiment on comprehensive domain databases and domain–domain interaction databases. We compare our single motifs with the domains in specific domain databases to study the relationship between our motifs and domains. Even more importantly, we study the relationship between motif pairs at interaction sites and interacting domain pairs, by mapping our motif pairs into domain–domain interacting pairs and analyzing the amount of overlaps between our mapped domain pairs and those in domain–domain interaction databases.

2 INTERACTING PROTEIN GROUPS

We fix PrtAll to be a set of m proteins: $\{P_i, i = 1, \ldots, m\}$, and PairDB to be all n interacting protein pairs of the proteins in PrtAll. That is, PairDB $= \{PP_i = \{P_i, Q_i\}, i = 1, \ldots, n, P_i \in \text{PrtAll}, Q_i \in \text{PrtAll}$, where P_i and Q_i have interactions$\}$. PrtAll and PairDB are used throughout the paper.

DEFINITION 1. *[Neighborhood of a protein] The* neighborhood $\beta(P)$ *of a protein* $P \in$ PrtAll *is defined as the set of proteins in* PrtAll *that interact with* P*. That is,* $\beta(P) = \{Q \mid Q \in \text{PrtAll}, Q$ *interacts with* $P\}$.

This neighborhood notion can be generalized by replacing one protein with a protein set. Then the definition can capture a partial all-versus-all relation between two protein-sets.

DEFINITION 2. *[Neighborhood of a protein set] The neighborhood* $\beta(\mathbf{S})$ *of a protein set* $\mathbf{S} \subseteq$ PrtAll *is the intersection of the neighborhoods of all proteins in* $\mathbf{S}$*. In other words, it is the set of proteins that interact with all proteins in* $\mathbf{S}$*. That is,* $\beta(\mathbf{S}) = \bigcap_{P \in \mathbf{S}} \beta(P)$*. In particular, we define* $\beta(\emptyset) =$ PrtAll.

If a protein interacts with all proteins in $\mathbf{S}$, it must be in the neighborhood set $\beta(\mathbf{S})$. However, if a protein interacts with all proteins in $\beta(\mathbf{S})$, it may not be in $\mathbf{S}$. Our next definition gives a maximal all-versus-all neighborhood-relation between two protein-sets.

DEFINITION 3. *[A pair of interacting protein groups] Let* $\mathbf{A}, \mathbf{B} \subseteq$ PrtAll *be two protein sets. If* $\beta(\mathbf{A}) = \mathbf{B}$ *and* $\beta(\mathbf{B}) = \mathbf{A}$*, then we call* $\mathbf{A}$ *and* $\mathbf{B}$ *a pair of* interacting protein groups*. If* $|\mathbf{A}| \geq \tau$ *and* $|\mathbf{B}| \geq \tau$*, we call* $\mathbf{A}$ *and* $\mathbf{B}$ *a pair of* frequent *interacting protein groups, where* τ *is a positive user-defined threshold.*

Definition 3 also says that if two proteins are a pair of interacting protein groups, then every protein in one set ($\mathbf{A}$ or $\mathbf{B}$) interacts with all proteins in the other set, and vice versa. Note that not every protein set is an interacting protein group because the partner interacting protein group may not exist.

Our definition of interacting protein group pairs is closely related to that of maximal complete bipartite subgraphs in graph theory (Eppstein, 1994). For details about their theoretical issues, please refer to our previous paper (Li et al., 2005).

We require a protein group to be large enough ($\geq \tau$) in our definition. This is because it is rather hard to determine whether a motif is significant if there are only a few of proteins in a group.

Problem statement: Let a set PrtAll, its PairDB, and the sequence data of all the interacting protein pairs be given. The problem is to find all pairs of frequent interacting protein groups $\mathbf{A}$ and $\mathbf{B}$, such that $|\mathbf{A}| \geq \tau$ and $|\mathbf{B}| \geq \tau$, and then to identify "good" motif pairs from the pairs of frequent interacting protein groups.

3 METHODS

Our algorithm consists of two steps: The first step is to find all pairs of interacting protein groups from PairDB where this problem is transformed to the problem of mining *frequent patterns* (Agrawal and Srikant, 1994); The second step is to identify motif pairs from the pairs of protein groups discovered in the first step.

3.1 Mining interacting protein groups

The classic problem of mining frequent patterns in the data mining field is: Given a set I of *items* and a set TDB of *transactions* T_i, $i = 1, \cdots, x$, where a transaction T_i is a subset of I (i.e. $T_i \subseteq I$), the problem is to find all frequent patterns of TDB—all *itemsets* $I' \subseteq I$ such that the number of transactions that contain I' is no less than a threshold τ, where τ is a user-specified threshold. Here, the number of the transactions that contain I' is called the *support* of I' in TDB. The set of the identities (ids) of the transactions that contain I' is called the *occurrence set* of I', denoted by $g(I') = \{id(T_i) \mid T_i \in TDB, I' \subseteq T_i\}$, where $id(T_i)$ is the identity of T_i.

To transform the problem of mining frequent interacting protein groups to the problem of mining frequent patterns, a crucial step is to determine what is an item and what is a transaction. We map **a protein** to an item—therefore, PrtAll is I. Thus, a protein set is a transaction. The next crucial step is to determine which and how many protein sets (transactions) are in TDB. We define a transaction in TDB as **the neighborhood of a protein** in PrtAll. Thus TDB is the set of all the neighborhoods of all the proteins in PrtAll. This TDB is specially denoted as TDB^{PrtAll}. So, this special TDB contains m transactions, namely $T_i = \beta(P_i)$, $i = 1, \cdots, m$, where m is the total number of proteins in PrtAll. The identity of each transaction T_i ($i = 1, \cdots, m$) is P_i, namely $id(T_i) = P_i$.

Let $\mathbf{X}$ be a frequent pattern of TDB^{PrtAll}. Let the support of $\mathbf{X}$ be k, and its occurrence set be $g(\mathbf{X}) = \{P_1, P_2, \cdots, P_k\}$. Then, the meaning of $\mathbf{X}$ is (of course) a protein set in which every protein interacts with all proteins in the occurrence set $g(\mathbf{X})$. This is because $\mathbf{X}$ is a subset of $\beta(P_i)$ for all $i = 1, \cdots, k$. In other words, all proteins in $\mathbf{X}$ interact with every protein in $g(\mathbf{X})$. Therefore, all proteins in $g(\mathbf{X})$ must be in the neighborhood of $\mathbf{X}$ (i.e. $\beta(\mathbf{X})$).

Furthermore, there is no protein other than P_i ($i = 1, \cdots, k$) that interacts with all proteins in $\mathbf{X}$. This is due to the definition of support of a pattern. We explain it using a contradiction. Suppose there is a protein $Q \in$ PrtAll other than P_i ($i = 1, \cdots, k$) that interacts with all proteins in $\mathbf{X}$, then $\beta(Q)$—a transaction in TDB^{PrtAll}—contains $\mathbf{X}$. So, the support of $\mathbf{X}$ would be $k + 1$. But, the support of $\mathbf{X}$ is only k. Here is a contradiction. Therefore, there are exactly only P_i, $i = 1, \cdots, k$, that interact with all proteins in $\mathbf{X}$. This can be re-written as $\beta(\mathbf{X}) = g(\mathbf{X})$. That is, the neighborhood of a frequent pattern $\mathbf{X}$ of TDB^{PrtAll} is the occurrence set of $\mathbf{X}$. All these ideas and discussions can be in the important theorem below.

THEOREM 1. *Let $\mathbf{X}$ be a frequent pattern of TDB^{PrtAll}. Then $\beta(\mathbf{X}) = g(\mathbf{X})$, and $g(\mathbf{X})$ is a frequent protein set.*

Let $f_\beta(\mathbf{X}) = \beta(\beta(\mathbf{X}))$. If $|f_\beta(\mathbf{X})| \geq \tau$, then $\beta(\mathbf{X})$ and $f_\beta(\mathbf{X})$ is a pair of frequent interacting protein groups.

PROOF. Denote $\beta(\mathbf{X}) = \mathbf{A}$ and $f_\beta(\mathbf{X}) = \mathbf{B}$.
(I) Obviously, $\beta(\mathbf{A}) = f_\beta(\mathbf{X}) = \mathbf{B}$.
(II) As $\mathbf{B}$ is the neighborhood of $g(\mathbf{X})$, then

$$\mathbf{B} = \bigcap_{P \in g(\mathbf{X})} \beta(P)$$

So, $\mathbf{B}$ is an itemset of TDB^{PrtAll} with support at least $|g(\mathbf{X})|$. On the other hand, it is a superset of $\mathbf{X}$ as $\mathbf{X}$ is contained in every $\beta(P)$ for $P \in g(\mathbf{X})$, so, its support is at most $|g(\mathbf{X})|$. Therefore, $\mathbf{B}$ and $\mathbf{X}$ have the same level of support in TDB^{PrtAll}, and also have the same occurrence set, namely the $g(\mathbf{X})$. This means $\beta(\mathbf{B}) = g(\mathbf{X}) = \mathbf{A}$.

Combining (I) and (II), we get $\beta(\mathbf{X})$ and $f_\beta(\mathbf{X})$ is a pair of frequent interacting protein groups if $|f_\beta(\mathbf{X})| \geq \tau$.

This theorem indicates that every frequent pattern of TDB^{PrtAll} corresponds to a candidate for a pair of frequent interacting protein groups.

Is there any other patterns of TDB^{PrtAll} that could correspond to a pair of frequent interacting protein groups? The answer is no. This is because for an infrequent pattern $\mathbf{X}$ of TDB^{PrtAll}, its occurrence set $\beta(\mathbf{X})$ is infrequent, i.e. $|\beta(\mathbf{X})| < \tau$. So, no matter what is the size of $f_\beta(\mathbf{X})$, $\beta(\mathbf{X})$ and $f_\beta(\mathbf{X})$ is not a pair of frequent interacting protein pairs.

We also conjecture that some frequent patterns of TDB^{PrtAll} can lead to the same pair of interacting protein groups. In fact, all frequent patterns $\mathbf{X}$ in an *equivalence class* (Nicolas et al., 1999) share the same pair of interacting protein groups (Li et al., 2005). The resulted $f_\beta(\mathbf{X})$ defined in above theorem one-to-one matches an equivalence class, often termed as a closed pattern (Nicolas et al., 1999). So, it is unnecessary for us to identify all frequent patterns from TDB^{PrtAll} for a given threshold τ, instead, one can just discover all frequent closed patterns from TDB^{PrtAll} using an efficient algorithm such as FPClose* (Grahne and Zhu, 2003).

3.2 Generating a motif pair from a pair of interacting protein groups

Given a protein group and its sequence data, we can get a motif (possibly containing flexible gaps) by using standard motif discovery algorithms such as PROTOMAT (Henikoff and Heinikoff, 1991) and MEME (Bailey and Elkan, 1995). So, we can easily obtain a motif pair from a pair of interacting protein groups by executing the motif discovery algorithm twice. In this paper, we choose PROTOMAT (Henikoff and Heinikoff, 1991) as the motif discovery algorithm because it is believed to be a good method to find local conserved regions from a group of related proteins. PROTOMAT is also a key method to construct BLOCKS database (Pietrokovski et al., 1996)—a comprehensive database of highly conserved regions for homologous protein groups (domains).

4 RESULTS

To assess the performance of our proposed method for mining motif pairs, we performed several experiments on a PC with a CPU clock rate of 3.2GHz and 2GB of main memory. The protein interaction set PairDB used in the experiments was downloaded from

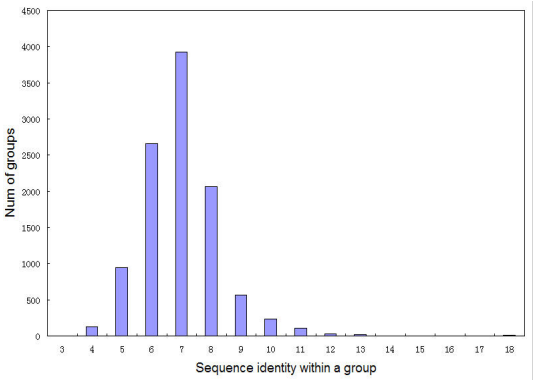

Fig. 2. The distribution of the sequence identities within our 10698 groups.

DIP (database of interacting proteins) on Oct. 23, 2005, consisting of 17511 experimentally determined interactions in saccharomyces cerevisiae (yeast) among 4959 proteins. We select 10640 physical interactions by excluding 6871 interactions determined only by complex level experiments such as Tandem Affinity Purification (TAP) and immunoprecipitation (the full list of excluded experiments can be found in the supplementary information). To discover frequent closed patterns by FPClose* (Grahne and Zhu, 2003), we set the threshold $\tau = 5$, an average number of interactions per protein in the yeast genome (Grigoriev, 2003). Default parameters are used for PROTOMAT (Henikoff and Heinikoff, 1991). To facilitate our analysis, we further term the motifs induced from closed patterns as left motifs (left blocks), while the ones induced from the occurrence sets of the closed patterns as right motifs (right blocks).

The FPClose* algorithm outputs a total of 5349 non-redundant pairs of interacting protein groups, by taking 4.35 seconds on our machine (including the transformation). The mining based on the transformation idea is very efficient compared to a naive search method which needs about 33 minutes (455-fold more than the efficient approach) to find all the protein groups. The implementation of the naive search contains some optimization techniques.

The homology property within a group is an interesting issue. It can be estimated simply by the sequence identity within the group—A value less than 15% is often considered as a good indicator for non-homology (Doolittle, 1981). We calculate all pairwise sequence identities within a same protein group using CLUSTAL W package with default parameters (Thompson et al., 1994). Then we use the average value of these pairwise sequence identities as the sequence identity within the group. The distribution of the sequence identities within the 10698 groups is shown in Figure 2, with more details in the supplementary information. The expected value of the sequence identities within the groups is 7.48%, with a standard deviation 1.33%. This is a good value indicating the non-homology within these groups. Therefore, these groups and their underlying sequence motifs are unlikely to detect by standard methods based on sequence homology (Sauder et al., 2000).

The PROTOMAT method outputs 5343 motif pairs from these 5349 pairs of interacting protein groups by taking 3 hours. 85% of protein groups generate two or three blocks. (Note that a group in BLOCKS contains 6.91 blocks on average.) Only 4 left groups and 2 right groups failed to produce any valid motif, with a failure rate $< 0.2\%$. Totally, there are 11948 left blocks and 13004 right blocks. The average length of these blocks is 11.05, with a standard deviation 5.06. Compared with BLOCKS where the average length of blocks is 25.337 and the standard deviation is 12.897, our blocks are more specific and match better with current knowledge about interaction sites, that is, 10-20 residues in length (Sheu et al., 2005).

Table 1. Databases used in our validation experiments.

	BLOCKS	PRINTS	Pfam	iPfam
Version	14.0	37.0	16.0	18.0
Num. of domains	4944	1850	7677	2145
Num. of entries	24294	11170	7677	3045

We treat the whole set of blocks generated by PROTOMAT from a protein group rather than each individual block as a motif to reflect the cooperation among these blocks. We expect that some interactions happen among the blocks from different sides of the motif pair, but do not study the detailed interactions among these blocks in this paper. In our results, the average number of blocks per motif is 2.33, with a standard deviation 0.73. The average number of proteins per motif is 7.01, with a standard deviation 2.59. More details are in the supplementary information.

4.1 Validations

Currently, comprehensive databases for motif–motif interactions (motif pairs) are hard to find but there are a handful of databases for domain–domain interactions such as iPfam (Finn et al., 2005), 3did (Stein et al., 2005) and InterDom (Ng et al., 2003). Since domains are known to involve in protein interactions and are closely related to motifs, we compare our motif pairs with these domain pairs. The following two steps are employed to illustrate the effectiveness of our algorithm.

- Compare all single motifs in our discovered motif pairs with all domains in specific domain databases to obtain overall matches, i.e. to determine the number of motifs that can be mapped to these domains and the overall correlation in the portions that are mapped.
- Map our motif pairs into domain–domain interacting pairs to determine the number of overlaps between our mapped domain pairs and those in the domain–domain interaction database.

4.1.1 Validations for single motifs As our motifs are in the form of blocks, we need domain databases also in the form of blocks for comparison. Currently, there are two major domain databases in the form of blocks: BLOCKS (Pietrokovski et al., 1996) and PRINTS (Attwood and Beck, 1994). Some information of these two domain databases are shown in the first two columns of Table 1, where an entry corresponds to a block.

The comparison is conducted by a program called *Local Alignment of Multiple Alignments* (LAMA) (Pietrokovski, 1996). It utilizes Smith-Waterman algorithm (Smith and Waterman, 1981) to determine the optimal local alignments for pairs of position-specific scoring matrices (PSSMs) (Gribskov et al., 1987) of the corresponding blocks. To estimate the alignment scores with different lengths and to filter out the coincidental matches, LAMA uses the *Z-score* as a significance measurement, where a Z-score between a pair of PSSMs is defined as the number of standard deviations away from the mean score generated by millions of shuffled blocks in the BLOCKS database.

In our study, we used the default threshold 5.6 for Z-score in LAMA to compare our blocks with those in BLOCKS and PRINTS. If 95% of the positions of a block are in the optimal alignment between this block and another block and the Z-score is no less than

Table 2. Statistics of mappings from our blocks to blocks in the BLOCKS and PRINTS databases.

	# of our blocks	# of mappings to BLOCKS blocks	# of mappings to PRINTS blocks	Average correlation
Left blocks	11948	29357	8632	54.31
Right blocks	13004	30220	8738	53.42

Table 3. Statistics of blocks or domains in the BLOCKS or PRINTS databases that can be mapped from our blocks or motifs.

	Mapped / total # in BLOCKS	Mapped / total # in PRINTS	Mapped / total # in ANY
Blocks	6408 / 24294	2174/ 11170	8582 / 35464
Domains	3128 / 4944	1093/ 1850	4221 / 6794

Table 4. Statistics of blocks or motifs in our motif pairs that can be mapped to blocks or domains in BLOCKS or PRINTS databases.

	total #	# mapped to BLOCKS	# mapped to PRINTS	# mapped to ANY
Blocks	24952	13859	8010	14620
Motifs	10686	8879	6464	9153

the threshold, we say there is a *mapping* from the former block to the latter one. If there is a mapping from any block of a motif to any block of a domain, we say the motif can be mapped to the domain. We have following results from this experiment:

- On average, each of our blocks maps to about 3.08 blocks in the BLOCKS or PRINTS databases. See more detailed report in the columns 2 and 3 of Table 2.
- The average correlation between the columns of our blocks and the columns from the database in the optimal alignments is as high as 53.88%. See column 4 of Table 2 for details.
- Our motifs can be mapped to 4221 domains out of a total of 6794 domains in these two databases, having a coverage of 62%. See Table 3. This result is interesting as our blocks can only be mapped to 8582 blocks out of the total 35464 blocks in these two databases, having a coverage $< 24\%$. The interpretation from a biological perspective is that most domains have about 40% of blocks as their interaction sites, while others may be related to folding.
- Although only 59% (14620 out of 24952) of our blocks can be mapped to blocks in BLOCKS and PRINTS, as high as 86% (9153 out of 10686) of motifs can be mapped to domains in these two databases. See Table 4 for details.

Note that our groups and groups in BLOCKS and PRINTS are constructed in quite different ways and their homology properties are also different. However, our comparison results reveal high correlation between their resulted blocks. This correlation may origin from the common involvement of interactions for both our motifs and their domains. This confirms the effectiveness of our method in some way.

4.1.2 Validations for motif pairs To assess whether our discovered motif pairs are indeed interaction sites, we compare them with domain–domain interacting pairs. If our motif pairs represent interaction sites, they should be mapped to some domain–domain interacting pairs in some databases. We choose iPfam (Finn et al., 2005) for this purpose. It consists of 3045 interacting pairs among 2145 Pfam domains derived from protein complexes in PDB.

Table 5. Occurrences of our mapped domains in different databases.

	BLOCKS	PRINTS	Combined
BLOCKS/PRINTS domains	3128	1093	4221
Pfam domains	2305	144	2338
iPfam domains	975	87	997

The cross-links between our motif pairs and the domain–domain pairs in iPfam is complicated. A reason is that the domain–domain pairs are represented by Pfam entries. To find the cross-links, we (i) firstly map our motifs to domains (protein groups) in the BLOCKS or PRINTS database, as shown in Section 4.1.1; (ii) we then map a protein group of BLOCKS to a protein group of InterPro (Apweiler et al., 2001) as there exists a one-to-one mapping between an entry of BLOCKS and an entry of InterPro; (iii) then we use existing cross-links between protein groups of InterPro and domains of Pfam to determine the cross-link between our motifs and Pfam domains. By this roadmap, we can map our motif pairs into domain–domain pairs with Pfam domain entries. Note that the association between PRINTS and Pfam is clear. Also note that the cross-linking mapping between motif pairs and domain–domain pairs is not a one-to-one mapping.

Using the above cross-link mapping, we compared our 5343 motif pairs with the 3045 domain-domain pairs in the iPfam database, 47 motif pairs can be mapped to 18 distinct domain pairs among 22 domains occurring in PDB complexes for 172 times (totally 105 distinct protein complexes).

Though the overlapping proportion seems modest, we assert that the result is significant because:

- We read only interacting protein sequence pairs, while some predictions about interaction sites can be confirmed by domain-domain interactions in PDB complexes.
- iPfam is a rather incomplete database, containing merely 3045 pairs among 2145 domains. Moreover, only 997 out of 4221 of our mapped domains are studied in iPfam, as shown in Table 5.
- The motif pairs we discovered are taken only from the yeast genome while iPfam covers a variety of species.
- Comparing with Interdom with 30037 putative interacting domain pairs (Ng et al., 2003), our motif pairs can be mapped to 203 domain pairs, including 94 high-confidence ones.

4.2 A case study

The 5343 motif pairs that we discovered can be ranked according to their correlation score in the mapping. Most of top-ranked motif pairs can be confirmed by protein complexes. Here we report details of one such pair. Our purpose is to check whether some block pairs in the motif pair can be aligned with a segment pair in a complex containing the mapped domain pair, and then check whether the segment pair has some contacts among their residues.

This motif pair is generated from the first pair of interacting protein groups. This protein group pair generates three blocks on the left and one block on the right. The first left block 1xxxxxxA contains 24 positions, while the right block 1xright contains 36 positions, as shown in Table 6.

Through the approach depicted in section 4.1.2, we map the block pair (1xxxxxxA, 1xright) into domain pair (PF01423,PF01423) in iPfam. Pfam database indicates that PF01423 is a LSM domain, and iPfam shows that one LSM domain interacts with another LSM domain densely in 20 complexes such as pdb1mgq, pdb1h64. We

Table 6. Left block 1xxxxxxA aligning with the chain A and right block 1xright aligning with the chain B of complex 1mgq, where capital letters are well aligned and low-case letters are skipped in the alignment.

```
AC 1xxxxxxA; distance from previous block=(18,243)
BL LLE motif=[4,0,17] motomat=[1,80,-10] width=24 seqs=4
   DIP:1330N ( 58)  LRDGRMLFGVLRTFD QY A NLI LQD
   DIP:2570N ( 206) TLE GRE I MI RNLSTE LL D  ENLLRE
   DIP:848N  ( 19)  LKNGE I  I QG ILT NVDNWM NLTLSN
   DIP:883N  ( 244) LQS GRR SKRDLS PEE  QR R  LQI RHA

   pdb1mgq_A(  30) LKg dRE  f r GVLk SFD Lh M  NLv Ln D

AC 1xright; distance from previous block=(6,52)
BL GNL motif=[3,0,17] motomat=[1,80,-10] width=36 seqs=5
   DIP:1417N ( 12)  IDK TI N QKVLI V LQS NRE FEG TLV GFD DFV NVI LED
   DIP:1418N ( 53)  LSD I I G KTV NVKLAS GLL YSG RLE S I D GFM NVALSS
   DIP:1419N ( 22)  LAKYKDSK I RVK LMGGKL VI G VLKGYD QLMNLV LDD
   DIP:794N  ( 7)   FKT LVD QEV VVE LKN DI E  I KG TLQ SVD QFL  NLKLDN
   DIP:903N  ( 24)  LKDYLN KRV V I I KVDGEC LI A SLN GFD KNT NLF I TN

   pdb1mgq_A(  18) Lg n s LN S p V i I KLKG DRE Fr G VLKSFD  l h MNLV Ln D
```

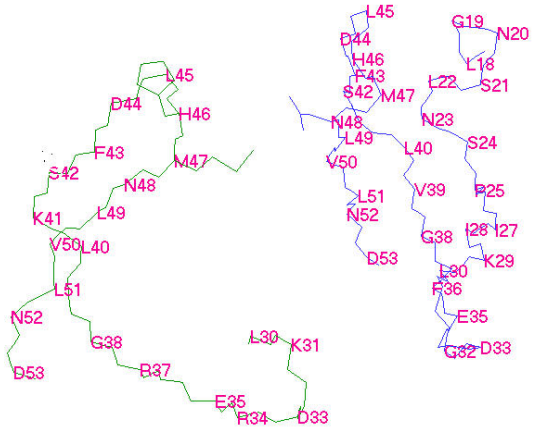

Fig. 3. Interactions between segment [30L, 53D] of the chain LSM A and segment [18L,53D] of the chain LSM B in the complex pdb1mgq (showing only the backbone).

take the complex pdb1mgq as an example to explain what we found. It has 7 chains each containing a LSM domain. The 3-D structure of these 7 chains and their interactions can be found in our Supplementary Information and also in the reference (`http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1mgq`). We observed the following details:

- Our left block 1xxxxxxA can be well aligned at positions 30 to 53 within the LSM domain of the chain A at the complex pdf1gmgq, and our right block 1xright can be well aligned at positions 18 to 53 of the chain B also within the LSM domain at the same complex. See Table 6 for alignment details.
- The residue 47M (residue M at position 47) of the chain A interacts with residue 48N of the chain B in pdb1mgq; another pair between residue 46H of the chain A and residue 48N of the chain B is also spatially close. See Figure 3 for details about the interactions between this segment pair (`http://www.sanger.ac.uk/cgi-bin/Pfam/detailed_interaction_view.pl?acc=PF01423&partner=PF01423&pdb=1mgq`).
- The interaction pair (47M,48N) is well conserved in the complex pdb1mgq—it occurs in 7 chain interactions out of a total of 9 chain interactions. The 7 interactions are between chain A and chain B, between chain B and chain C, ..., and between chain G to chain A. Interestingly, this residue interaction located in the middle of the domain is also highly conserved in other complexes containing LSM domains, for example in the complex pdb1h64.

5 DISCUSSION AND CONCLUSION

Our motif pairs are conceptually similar with correlated sequence-signatures proposed by Sprinzak and Margalit (2001). But their correlated sequence-signatures are modeled as over-represented domain pairs, which are essentially longer than our motif pairs and can not derive novel binding motifs since their domains are predefined. On the other hand, our interacting protein group pairs are structurally similar with interacting domain profile pairs proposed by Wojcik and Schachter (2001). But each of their domain profiles is the summarization of a domain cluster, which is a set of domains sharing significant sequence similarity and interacting with the same region of a certain protein. This approach replies on protein-protein interactions with domain interaction annotations, which are not widely available.

In our model, we require that pairs of interacting protein groups should always have an all-versus-all relationship. This is a bit strict as it is vulnerable to handle incomplete dataset. As a future direction, we will consider *most-versus-most* relationship.

Other future work include new evaluation methods. For example, the predicted interaction sites in the blocks of motif pairs can be compared with known interaction sites in some protein-protein interaction databases (Rain et al., 2001) or compared with interaction sites in interface databases (Keskin et al., 2004). Also our motif pairs can be compared with those learned from non-interacting protein pairs or from random protein pairs, to study their statistical significance, as done in our previous study (Li and Li, 2005).

Finally, we summarize the main results achieved in this work. We have used the concept of motif pairs to model protein interaction sites and studied the mining problem based on the sequence data of interacting protein pairs. We have proposed the new concept of interacting protein groups for the discovery, where a protein group may share a common interaction motif and a pair of protein groups may share a motif pair at their interaction sites. We transformed the mining of interacting protein groups into the mining of frequent closed patterns. We used standard motif discovery algorithms onto these discovered interacting protein groups to generate motif pairs in form of blocks. The high efficiency of this two-step approach is due to: (1) In the discovery of interacting protein groups, we examine only interacting protein pairs without checking their sequences, thereby dramatically reduce the complexity of the problem; (2) By producing protein groups firstly, the discovery of interaction motifs is greatly accelerated as we need not execute the NP-hard motif discovery algorithm on insignificant candidates of protein sets.

The systematic validation results of the discovered motif pairs indicate that our discovered motifs have high correlation with domains in the existing domains databases. Our discovered motif pairs can also be mapped into the domain–domain interacting pairs in an experimentally validated domain–domain database with good matches.

ACKNOWLEDGEMENT

We are grateful to Hao Han and Soon Heng Tan for their comments on the validation methods, biological concepts and nomenclature in the paper. We are also thankful to Benjamin Schuster-B*ö*ckler for providing the iPfam database. We appreciate Donny Soh and Ling Li for polishing the initial version of the manuscript.

REFERENCES

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules. In *Proc. of the 20th Int'l Conference on Very Large Databases*, Chile, pp. 487–499.

Aloy,P., Pichaud,M. and Russell,R.B. (2005) Protein complexes: structure prediction challenges for the 21st century. *Curr. Opin. Struct. Biol.*, 15,15–22.

Aloy,P. and Russell,R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.*, 22,1317–1321.

Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E. and et al. (2001) The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29,37–40.

Attwood,T.K. and Beck,M.E. (1994) Prints–a protein motif fingerprint database. *Protein Eng.*, 7,841–848.

Aytuna, A.S., Gursoy,A. and Keskin,O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21,2850–2855.

Bailey, T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21,51–80.

Clemmons,D.R. (2001) Use of mutagenesis to probe igf-binding protein structure/function relationships. *Endocr. Rev.*, 22,800–817.

DeLano,W.L., Ultsch,M.H., de Vos,A.M. and Wells,J.A. (2000) Convergent solutions to binding at a protein-protein interface. *Science*, 287,5456.

Doolittle,R. F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, 214,149–159.

Eppstein,D. (1994) Arboricity and bipartite subgraph listing algorithms. *Inf. Process. Lett.*, 51,207–211.

Finn,R.D., Marshall,M. and Bateman,A. (2005) ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21,410–412.

Gallet, X., Charloteaux,B., Thomas,A. and Brasseur R. (2000) A fast method to predict protein protein interaction sites from sequences. *J. Mol. Biol.*, 302,917–926.

Garman,S.C., Wurzburg,B.A., Tarchevskaya,S.S., Kinet,J.P., and Jardetzky,T.S. (2000) Structure of the fc fragment of human ige bound to its high-affinity receptor fc epsilonri alpha. *Nature*, 406,259–266.

Grahne,G. and Zhu,J. (2003) Efficiently using prefix-trees in mining frequent itemsets. In *Workshop on Frequent Itemset Mining Implementations(FIMI)*, USA.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84,4355–4358.

Grigoriev,A. (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.*, 31,4157–4161.

Henikoff,S. and Heinikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, 19,6565–6572.

Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, 93,13–20.

Jones,S. and Thornton,J.M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, 272,133–143.

Keskin,O., Ma,B. and Nussinov,R. (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, 345,1281–1294.

Keskin,O. and Nussinov,R. (2005) Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng. Des. Sel.*, 18,11–24.

Keskin,O., Tsai,C.J., Wolfson,H. and Nussinov,R. (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.*, 13,1043–1055.

Kumar,S., Ma,B., Tsai,C.J., Sinha,N. and Nussinov,R. (2000) Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.*, 9,10–19.

Li,H. and Li,J.(2005) Discovery of stable and significant binding motif pairs from pdb complexes and protein interaction datasets. *Bioinformatics*, 21, 314–324.

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

Li,J., Li,H., Soh,D., and Wong,L. (2005) A correspondence between maximal complete bipartite subgraphs and closed patterns. In *9 th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Porto, Portugal, pp. 146–156.

Liang,J., Edelsbrunner,H. and Woodward,C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, 7,1884–1897.

Loregian,A. and Palu,G. (2005) Disruption of protein-protein interactions: towards new targets for chemotherapy. *J. Cell. Physiol.*, 204,750–62.

Mendez,R., Leplae,R., Lensink,M.F., and Wodak,S.J. (2005) Assessment of capri predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60, 150–69.

Ng,S.K., Zhang,Z. and Tan,S.H. (2003) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19,923–929.

Nicolas,P., Yves,B., Rafik,T., and Lotfi,L. (1999) Discovering frequent closed itemsets for association rules. In *Proc. of the 7th ICDT*, Israel, pp. 398–416.

Ofran,Y. and Rost,B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.*, 544,236–239.

Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, 271, 511–523.

Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, 24,3836–3845.

Pietrokovski,S., Henikoff,J.G. and Henikoff,S. (1996) The blocks database–a system for protein classification. *Nucleic Acids Res.*, 24,197–200.

Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V., Chemama,Y., Labigne,A. and Legrain,P. (2001). The protein-protein interaction map of helicobacter pylori. *Nature*, 409, 211–215.

Sauder,J.M., Arthur,J.W., and Dunbrack,R.L. Jr. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, 40,6–22.

Sheu,S.H., Jr Lancia,D.R., Clodfelter,K.H., Landon,M.R., and Vajda,S. (2005) Precise: a database of predicted and consensus interaction sites in enzymes. *Nucleic Acids Res*, 33,D206–11.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, 147,195–197.

Sprinzak,E. and Margalit,H.(2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311,681–692.

Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28,405–420.

Stein,A., Russell,R.B. and Aloy,P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, 33,D413–D417.

Swanson,R.V., Lowry,D.F., Matsumura,P., McEvoy,M.M., Simon,M.I. and Dahlquist,F.W. (1995) Localized perturbations in chey structure monitored by nmr identify a chea binding interface. *Nat. Struct. Biol.*, 2,906–910.

Takahashi,H., Nakanishi,T., Kami,K., Arata,Y. and Shimada,I.(2000) A novel nmr method for determining the interfaces of large protein-protein complexes. *Nat. Struct. Biol.*, 7,220–223.

Terwilliger,T.C. (2004) Structures and technology for biologists. *Nat. Struct. Mol. Biol.*, 11,296–297.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673–4680.

Tong,A.H., Drees,B., Nardelli,G., Bader,G.D., Brannetti,B., Castagnoli,L., and et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295,321–324.

Wang,H., Segal,E., Ben-Hur,A., Koller,D. and Brutlag,D. (2004) Identifying protein-protein interaction sites on a genome-wide scale. In *Advances in Neural Information Processing Systems 17*, USA, pp. 1465–1472.

Wojcik,J. and Schachter,V.(2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17,S296–S305.

Yan,C., Dobbs,D. and Honavar,V. (2004) A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20,I371–I378.

Zhou H.X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44,336–343.