Modeling Protein Interacting Groups by Quasi-bicliques: Complexity, Algorithm and Application

Xiaowen Liu, Jinyan Li and Lusheng Wang* Member, IEEE

Abstract—Protein-protein interactions (PPIs) are one of the most important mechanisms in cellular processes. To model protein interaction sites, recent studies have suggested to find interacting protein group pairs from large PPI networks at the first step, and then to search conserved motifs within the protein groups to form interacting motif pairs. To consider noise effect and incompleteness of biological data, we propose to use quasi-bicliques for finding interacting protein group pairs. We investigate two new problems which arise from finding interacting protein group pairs: the maximum vertex quasi-biclique problem and the maximum balanced quasi-biclique problem. We prove that both problems are NP-hard. This is a surprising result as the widely known maximum vertex biclique problem is polynomial time solvable [1]. We then propose a heuristic algorithm which uses the greedy method to find the quasi-bicliques from PPI networks. Our experiment results on real data show that this algorithm has a better performance than a benchmark algorithm for identifying highly matched BLOCKS and PRINTS motifs. We also report results of two case studies on interacting motif pairs which map well with two interacting domain pairs in iPlam.


Index Terms—Protein-protein interactions, interaction sites, and quasi-bicliques.

1 INTRODUCTION

Proteins with interactions carry out most biological functions within living cells such as gene expression, enzymatic reactions, signal transduction, inter-cellular communications and immunoreactions. As the interactions are mediated by short sequence of residues among the long stretches of interacting sequences, these interacting residues or so-called interaction (binding) sites are at the central spot of proteome research. Although many imaging wet-lab techniques like X-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy and mass spectrometry have been developed to determine protein interaction sites, the solved amount of protein interaction sites constitute only a tiny proportion among the whole population due to high cost and low throughput. Computational methods are still considered as the major approaches for the deep understanding of protein binding sites, especially for their subtle 3-dimensional structure properties that are not accessible by experimental methods.

The classical graph concept—maximal biclique subgraph (also known as maximal complete bipartite subgraph)—has been emerged recently for bioinformatics research closely related to topological structures of protein interaction networks and biomolecular binding sites. For example, Thomas et al. introduced complementary domains in [2], and they showed that the complementary domains can form near complete bipartite subgraphs in PPI networks. A lock-and-key model has been proposed by Morrison et al. which is also based on the concept of maximal complete bipartite subgraphs [3]. Very recently, Andreopoulos et al. used clusters in PPI networks for identifying locally significant protein mediators [4]. Their idea is to cluster common-friend proteins, which are in fact complete-bipartite proteins, based on their similarity to their direct neighborhoods in PPI networks. Other computational methods studying bipartite structures of PPI networks include [5], [6] which were focused on protein function prediction.

To identify motif pairs at protein interaction sites, Li et al. introduced a novel method with the core idea related to the concept of complete bipartite subgraphs from PPI networks [7]. The first step of the algorithm in [7] finds large subnetworks with all-versus-all interactions (complete bipartite subgraphs) between a pair of protein groups. As the proteins within these protein groups have similar protein interactions and may share the same interaction sites, the second step of Li’s algorithm is to compute conserved motifs (possible interaction sites) by multiple sequence alignments within each protein group. Thus, those conserved motifs can be paired with motifs identified from other protein groups to model protein interaction sites. One of the novel aspects of the algorithm in [7] is that it combines two types of data: the PPI data and the associated sequence data for modeling binding motif pairs.

Each protein in the above PPI networks is represented
by a vertex and every interaction between two proteins is represented by an edge. Discovering complete bipartite subgraphs in PPI networks can thus be formulated as the following biclique problem: Given a graph, the biclique problem is to find a subgraph which is bipartite and complete. The objective is to maximize the number of vertices or edges in the bipartite complete subgraph. We note that the maximum vertex biclique problem is polynomial time solvable [1]. This problem is also equivalent to the maximum independent set problem on bipartite graphs which is known to be solvable by a minimum cut algorithm. However, the maximum vertex balanced biclique problem is NP-hard [8]. The maximum edge biclique problem is proved to be NP-hard as well [9].

In this paper, we consider incompleteness of biological data, as the interaction data of PPI networks is usually not fully available. On the other hand, within an interacting protein group pair, some proteins in one group may only interact with a proportion of the proteins in the other group. Therefore, many subgraphs formed by interacting protein group pairs are not perfect bicliques—They are more often near complete bipartite subgraphs. Therefore, methods of finding bicliques may miss many useful interacting protein group pairs. To deal with this problem, we use quasi-bicliques instead of bicliques to find interacting protein group pairs. With the quasi-biclique, even though some interactions are missing in a protein interaction subnetwork, we can still find the two interacting protein groups. In this paper, we introduce and investigate two theoretical problems: the maximum vertex quasi-biclique problem and the maximum balanced quasi-biclique problem. We show that both problems are NP-hard. We also propose a heuristic algorithm for finding large quasi-bicliques in PPI networks.

2 BICLIQUES AND QUASI-BICLIQUES

Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph, where each vertex represents a protein and there is an edge connecting two vertices if the two proteins have an interaction. Since $G$ is an undirected graph, any edge $(u, v) \in \mathcal{E}$ implies $(v, u) \in \mathcal{E}$. For a selected edge $(u, v)$ in $G$, in order to find the two groups of proteins having the similar pairs of binding sites, we translate the graph $G = (\mathcal{V}, \mathcal{E})$ into a bipartite graph. Let $X = \{x| (x, v) \in \mathcal{E}\}, Y_1 = \{y| (u, y) \in \mathcal{E} \& u \notin X\}$ and $Y_2 = \{w| (w, v) \in \mathcal{E} \& w \in X\}$. For a vertex $w \in Y_2$, $w$ is incident to both $u$ and $v$ in $G$. Thus both $X$ and $Y_2$ contain $w$. We keep $w$ in $X$ and replace $w$ in $Y_2$ with a new virtual vertex $\overline{w}$. After replacing all vertices $w$ in $Y_2$ with $\overline{w}$, we get a new vertex set $Y_2'$. Let $Y = Y_1 \cup Y_2'$ and $E = \{(x, y)|(x, y) \in \mathcal{E} \& x \in X \& y \in Y_1\} \cup \{(x, \overline{w})|(x, w) \in \mathcal{E} \& x \in X \& \overline{w} \in Y_2\}$. In this way, we have a bipartite graph $G = (X \cup Y, E)$. A biclique in $G$ corresponds to two subsets of vertices, say, subset $A$ and subset $B$, in $G$. In $G$, every vertex in $A$ is adjacent to all the vertices in $B$, and every vertex in $B$ is adjacent to all the vertices in $A$. Moreover, $A \cap B$ may not be empty. In this case, for any vertex $w \in A \cap B$, $(w, w) \in \mathcal{E}$. This is the case, where the protein has a self-loop. Self-loops are very common in practice. When a self-loop appears, one protein molecule interacts with another identical protein molecule. For example, two identical protein subunits can assemble together to form a homodimeric protein.

In the following, we focus on the bipartite graph $G = (X \cup Y, E)$. For a vertex $x \in X$ and a vertex set $Y' \subseteq Y$, the degree of $x$ in $Y'$ is the number of vertices in $Y'$ that are adjacent to $x$, denoted by $d(x, Y') = |\{y| y \in Y' \& (x, y) \in \mathcal{E}\}|$. Similarly, for a vertex $y \in Y$ and $X' \subseteq X$, we use $d(y, X')$ to denote $|\{x| x \in X' \& (x, y) \in \mathcal{E}\}|$. Now, we are ready to define the $\delta$-quasi-biclique.

Definition 1: For a bipartite graph $G = (X \cup Y, E)$ and a parameter $0 < \delta \leq \frac{1}{2}$, $G$ is called a $\delta$-quasi-biclique if for each $x \in X$, $d(x, Y) \geq (1 - \delta)|Y|$ and for each $y \in Y$, $d(y, X) \geq (1 - \delta)|X|$. Similarly, a $\delta$-quasi-biclique in $G$ corresponds to two subsets of vertices, say, subset $A$ and subset $B$, in $G$. In $G$, every vertex in $A$ is adjacent to at least $(1 - \delta)|B|$ vertices in $B$, and every vertex in $B$ is adjacent to at least $(1 - \delta)|A|$ vertices in $A$. Moreover, according to the translation and the definition, $A \cap B$ may not be empty. Again, if a protein appears in both sides of a $\delta$-quasi-biclique and there is an edge between the two corresponding vertices, the protein has a self-loop. In our experiments, we observe that about 22% of the $\delta$-quasi-bicliques produced by our program contain self-loop proteins.

In many applications, due to various reasons, some edges in a clique/biclique may be missing and a clique/biclique becomes a quasi-clique/quasi-biclique. Thus, finding quasi-cliques/quasi-bicliques is more important in practice. Here we show that large quasi-bicliques may not contain any large bicliques.

Theorem 1: Let $G = (X \cup Y, E)$ be a random graph with $|X| = |Y| = n$, where for each pair of vertices $x \in X$ and $y \in Y$, $(x, y)$ is chosen, randomly and independently, to be an edge in $E$ with probability $\frac{1}{2}$. When $n \to \infty$, with high probability, $G$ is a $\frac{1}{2}$-biclique, and $G$ does not contain any biclique $G' = (X' \cup Y', \mathcal{E}')$ with $|X'| \geq 2\log n$ and $|Y'| \geq 2\log n$. (The proof is in the Appendix.)

In the biological context, Theorem 1 indicates that it is possible that some large interacting protein groups cannot be obtained by simply finding a maximal biclique if a few (interaction) edges are missing. As large interacting protein groups are more useful, according to this theorem, we have to develop new computational algorithms to extract from PPI networks large interacting protein groups which form quasi-bicliques.

In terms of false positive edges, both quasi-biclique and biclique can handle spurious edges very well. If very few spurious edges are added, in most cases, an irreleative protein will not be included in the quasi-bicliques or bicliques unless $(1 - \delta)|A|$ spurious edges are simultaneously added to the protein that has no interaction with any of the proteins in $A$, where $A$ is one of the two interaction groups.
which is known to be NP-hard [10]. We will construct a bipartite graph with $|S| = 3m$ and $|T| = (3m + 3)^2/p$. The set $Y_B$ also has three disjoint subsets $Y_B = Y_a \cup Y_b \cup Y_c$, where $|Y_a| = |Y_b| = n$ and $|Y_c| = (m + 1)^2 - 2m$. The $X_a \cup Y_a$ induced subgraph and the $X_b \cup Y_b$ induced subgraph are two copies of $G_A$. There is no edge between any vertex in $X_a \cup Y_a$ and any vertex in $X_b \cup Y_b$.

Each vertex in $X_c$ is adjacent to all vertices in $Y_B$ and each vertex in $Y_c$ is adjacent to all vertices in $X_B$. The purpose for adding the vertices in $X_c$ and $Y_c$ is to get the ratio $p$. Then, we get a bipartite graph with $|X_B| = (3m + 3)^2/p$ and $|Y_B| = 2n + (m + 1)^2/p - 2m$. For $S = \{s_1, s_2, ..., s_6\}$, $T = \{\{s_1, s_2, s_3\}, \{s_4, s_5, s_6\}, \{s_1, s_3, s_5\}\}$ and $p = 1/2$, we can construct a graph with $|X| = 18$ and $|Y| = 8$ (Fig. 2).

Step 3: We construct an instance $G = (X \cup Y, E)$ based on $G_B$. The set $X$ is the same as $Y_B$, i.e., $Y = Y_B = Y_a \cup Y_b \cup Y_c$. The set $X$ contains $3m$ disjoint subsets $X = X_1 \cup X_2 \cup \ldots \cup X_{3m}$, where $|X_1| = |X_2| = \ldots = |X_{3m}| = (3m + 3)^2/p$. For each $X_i$, the $X_i \cup Y$ induced subgraph has the same structure with $G_B$. The set $X_i$ has three disjoint subsets $X_i = X_{i,a} \cup X_{i,b} \cup X_{i,c}$, which correspond to $X_a$, $X_b$ and $X_c$ in $B$, respectively. Each vertex $x_{i,k} \in X_i$ is associated with the vertex $x_k \in X_B$. For each pair of vertices $x_{i,k} \in X_i$ and $y_j \in Y$, $(x_{i,k}, y_j) \in E$ iff $(x_k, y_j) \in E_B$. In the graph $G$, $|X| = 3m(3m + 3)^2/p$ and $|Y| = 2n + (m + 1)^2/p - 2m$.

Lemma 1: Let $(X' \cup Y', E')$ be a maximum vertex $\frac{p}{q}$-quasi-biclique of the constructed bipartite graph $G(X \cup Y, E)$. Set $Y'_a = Y_a \cap Y'$, $Y'_b = Y_b \cap Y'$, and $Y'_c = Y_c \cap Y'$. If $|X'| \geq 3m(3m + 3)^2/p$ and $|Y'| = 2n + (m + 1)^2/p - 2m$, we have $|Y'_a| = |Y'_b| = m$, $Y'_c = Y_c$ and $X' = X$. (The proof is in the Appendix.)

Theorem 2: For any constant integers $p > 0$ and $q > 0$ such that $0 < \frac{p}{q} \leq \frac{1}{2}$, the maximum vertex $\frac{p}{q}$-quasi-biclique problem is NP-hard. (The proof is in the Appendix.)

3.2 The Balanced Quasi-Biclique Problem

A balanced quasi-biclique is a quasi-biclique in which the numbers of the vertices in both groups are similar. The maximum balanced quasi-biclique problem is defined as follows:

Definition 3: Given a bipartite graph $G = (X \cup Y, E)$ and $0 < \delta \leq \frac{1}{2}$, the maximum balanced $\delta$-quasi-biclique problem is to find $X' \subseteq X$ and $Y' \subseteq Y$ such that the $X' \cup Y'$ induced subgraph is a $\delta$-quasi-biclique and $|X'| = |Y'|$ is maximized.

We can also prove that the maximum balanced quasi-biclique problem is NP-hard. The reduction is from the $3 \times 3$-BIC problem (perfect 3-cover) problem. The instance of $3 \times 3$-BIC problem consists of a finite set $S$ of $3m$ elements, and a collection $T$ of 3-element subsets of $S$, where $|T| > 3m$. The objective is to determine whether $T$ contains a sub-collection $T' \subseteq T$ (perfect 3 cover) such that every element in $S$ occurs in exactly one triple in $T'$.
large quasi-bicliques. Consider a PPI network $G = (V,E)$. Our heuristic algorithm has two steps. First, we construct a bipartite graph from the graph $G$ based on a pair of interacting proteins $(u,v)$. Using the method described at the beginning of Section 2, we can get a bipartite graph $G = (X \cup Y, E)$. Second, we find quasi-bicliques in $G$. The bipartite graph $G$ contains all proteins that have interactions with $u$ and $v$ in the bipartite graph.

In the algorithm for finding quasi-bicliques in $G$, we have two parameters $\delta$ and $\tau$, which control the quality and sizes of the quasi-bicliques. We use a greedy method to get the seeds for finding large quasi-bicliques in $G$. At the beginning, we set $X' = \phi$ and $Y' = Y$. In each step, we find a vertex with the maximum degree in $X - X'$. The vertex is added into the biclique vertex set $X''$. We eliminate all vertices $y$ in $Y''$ such that $d(y,X') < (1 - \delta)|X'|$. We will continue this process until the size of $Y'$ is less than $\tau$. At each step, we get a seed for finding large quasi-bicliques.

The seeds may miss some possible vertices in the quasi-bicliques. We can extend the seeds to find larger quasi-bicliques. Let $X'' = X'$ and $Y'' = Y'$ be a pair of seed vertex sets. In the first step, we can find a vertex $x$ in $X - X''$ with the largest degree $d(x,Y'')$ in $X - X''$. If $d(x,Y'') \geq (1 - \delta)|Y''|$, we add the vertex $x$ to $X''$. In the second step, we can find a vertex $y$ in $Y - Y''$ with the largest $d(y,X'')$ in $Y - Y''$. If $d(y,X'') \geq (1 - \delta)|X''|$, we add the vertex $y$ to $Y''$. We repeat the above two steps until no vertex can be added. The whole algorithm is shown in Fig. 4. We can also exchange the two vertex sets $X$ and $Y$ to find more quasi-bicliques using the algorithm.

Let $n$ be the number of vertices in the bipartite graph $G$. In the greedy algorithm, the time complexity of Steps 3–5 and Step 10 is $O(n)$, and the time complexity of Steps 6–9 is $O(n^2)$. So the time complexity of Steps 3–10 is dominated by $O(n^2)$. Since Steps 3–10 is repeated $O(n)$ times, the time complexity of the whole algorithm is $O(n^3)$.
The Greedy Algorithm

<table>
<thead>
<tr>
<th>Input</th>
<th>A bipartite graph ((X \cup Y, E)) and two parameters (\delta) and (\tau).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>A set of (\delta)-quasi-bicliques ((X' \cup Y', E')) with (</td>
</tr>
<tr>
<td></td>
<td>Let (X' = \emptyset) and (Y' = Y).</td>
</tr>
<tr>
<td></td>
<td>while (</td>
</tr>
<tr>
<td></td>
<td>3. Find the vertex (x \in X - X') with the maximum degree (d(x, Y')).</td>
</tr>
<tr>
<td></td>
<td>4. Add (x) into (X', X' = X' \cup {x}), and delete from (Y') all vertices (y \in Y') such that (d(y, X') &lt; (1 - \delta)</td>
</tr>
<tr>
<td></td>
<td>5. (X'' = X') and (Y'' = Y').</td>
</tr>
<tr>
<td></td>
<td>repeat</td>
</tr>
<tr>
<td></td>
<td>7. Find the vertex (x \in X - X'') with the maximum degree (d(x, Y'')).</td>
</tr>
<tr>
<td></td>
<td>8. Find the vertex (y \in Y - Y'') with the maximum degree (d(y, X'')).</td>
</tr>
<tr>
<td></td>
<td>until no vertex is added in the steps 7 and 8.</td>
</tr>
<tr>
<td></td>
<td>10. if (</td>
</tr>
</tbody>
</table>

Evaluation of the quality of bicliques (interacting protein groups)

There are two ways to evaluate the quality of found interacting protein groups or group pairs (bicliques or quasi-bicliques) as suggested in [7]. For the first method, we look at the number of motifs found by the program that are also in the two block databases, BLOCKS [12] and PRINTS [13]. Blocks in these databases are multiply aligned segments of sequences corresponding to the most highly conserved regions of protein sequences. They are generated by doing the multiple alignment of homologous protein sequences and finding highly conserved regions in the multiple alignments. The blocks can be used to characterize protein families and to detect or verify protein sequence homology. To find motifs that are also in BLOCKS and PRINTS, we use LAMA (Local Alignment of Multiple Alignments) [12] which can optimally align two blocks. For the second method, we map the motif pairs found by our greedy algorithm into domain-domain interaction pairs in domain-domain interaction database iPfam [14]. The versions of the databases are shown in Table 1.

In the rest of the section, we compare our program PPIExtend with program FPClose* in [7] using the above two methods. We also give two interesting case studies on binding motif pairs.

### 5.1 Motif Mapping with BLOCKS, PRINTS, and iPfam

The protein interaction data was provided by the authors of [7]. The data includes 10640 experimentally determined physical interactions of 4959 proteins in *Saccharomyces cerevisiae* (yeast). This protein interaction set was originally downloaded by Li et al. [7] from DIP (database of interacting proteins) on October 23, 2005, consisting of 17511 experimentally determined interactions. It was then filtered by removing 6871 interactions determined only by complex level experiments such as Tandem Affinity Purification (TAP) and immunoprecipitation. We set \(\delta = 0.1\) and \(\tau = 5\) and \(\alpha = 5\). Our greedy algorithm produced 59,124 interacting protein group pairs. In all the protein group pairs, 13266 pairs (about 22%) contain self-loops, and a large number of protein group pairs (about 78%) do not contain self-loops.

### Distributions of the sizes of protein groups

For a protein group pair \((A, B)\) (a quasi-biclique) found by our algorithm, we can assume that \(|A| \leq |B|\). We first analyze the distribution of the protein group pairs based on \(|A|\) and \(|B|\). In our experiment, we only output quasi-bicliques with \(|A| \geq 5\) and \(|B| \geq 5\). From the analysis of the 59124 quasi-bicliques, we found that

---

**Table 1**

<table>
<thead>
<tr>
<th>Databases used in the experiments.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>BLOCKS</td>
</tr>
<tr>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Version</td>
</tr>
<tr>
<td>Number of domains</td>
</tr>
<tr>
<td>Number of entries</td>
</tr>
</tbody>
</table>
all the biclques have |A| \leq 15. The distribution of the 59124 protein group pairs is shown in Table 2. In this
table, each row has a specific value a of |A| and each
column has a range [b, c] of |B|. Each cell is the number
of quasi-bicliques with |A| = a and |B| \in [b, c].

From the table, we can see that the most common sizes
of the quasi-bicliques are 5 \leq |A| \leq 15 and 10 \leq |B| \leq 20.

Now, we analyze the size distribution of the quasi-
bicliques based on |A||B|. The distribution of the sizes
|A||B| of the protein group pair is shown in Fig. 5. From
the figure, we can see that the range 180 ~ 189 contains
the largest number of protein group pairs.

**Comparison of PPIExtend and FPClose**

We have examined the overlaps between the protein
group pairs found by PPIExtend and the protein group
pairs found by FPClose* in [7], [15]. Let G_1 = (A_1, B_1)
and G_2 = (A_2, B_2) be two protein group pairs. If A_1 \cup B_1
contains more than 90% proteins in A_2 \cup B_2, we say
that G_1 covers G_2. We found that only 38 out of
the 5,349 protein group pairs found by FPClose* can not
be covered by the protein group pairs found by our
program PPIExtend. However, there are 38,305 protein
group pairs found by PPIExtend that cannot be covered
by any protein group pairs found by FPClose*.

Now, we look at the quality of the found protein
group pairs. We use the multiple sequence alignment
program PROTOMAT [12] to find the conserved motifs
within each of the protein groups. By using the default
parameters, PROTOMAT outputs 220,393 motifs from
59,124 pairs of interacting protein groups obtained from
PPIExtend. To compare our program PPIExtend with
FPClose* [7], [15], we use the two evaluation methods.

**The first evaluation method:** We look at the numbers of
motifs found by the programs PPIExtend and FPClose*
that are also in the two block databases, BLOCKS [12]
and PRINTS [13]. The LAMA program [12] is used
to find the local optimal alignment of two blocks (the
motif output by PPIExtend/FPClose* and a block in the
databases), where the Z-score is computed to measure
the alignments. The default threshold of Z-score was
used in the experiments. The results are reported in Table
3. From this table, we can see that our method has more
mappings to BLOCKS and PRINTS than FPClose* [7],
[15].

**The second evaluation method:** We look at the numbers of
motif pairs found by the two programs PPIExtend and
FPClose* that can be mapped between the same protein
group pairs in the database BLOCKS or PRINTS; (2) we map a protein group
of BLOCKS to a protein group of InterPro based on the one-
to-one mapping between an entry of BLOCKS and an
entry of InterPro; (Note that both PRINTS and Pfam
are member databases of InterPro, and the mapping between
PRINTS and Pfam is clear.) (3) we use existing cross-links
between protein groups of InterPro and domains of Pfam
to determine the cross-links between the motifs found by
PPIExtend/FPClose* and Pfam domains. In this way, we
can map our motif pairs into domain pairs with Pfam
domain entries. Note that the mapping between motif
pairs and domain pairs is not one-to-one.

We observed that the motif pairs found by PPIExtend
can map to 81 distinct domain pairs in Pfam. However,
only 18 domain pairs were reported in [7]. This is a
significant improvement and the main reason is the use
of quasi-bicliques. In the 81 domain pairs, 48 pairs are
domain-domain interactions on one protein (self-loops)
and 33 pairs are domain-domain interactions on different
proteins. Although the self-loops form a large portion,
we still find many other domain-domain interactions
that are not self-loops.

### 5.2 Protein Interaction Sites: Two Case Studies

In this section, we present detailed information about
two binding motif pairs that can be mapped to interacting
domain pairs. The first motif pair is derived from
a protein group pair in which the left protein group
contains 7 proteins and the right protein group contains
10 proteins. There are 66 interactions between the two
groups of proteins. Using the hypergeometric probability
model, the p-value of the protein group pair is less
than 1.57 \times 10^{-191}. PROTOMAT finds two left blocks
and two right blocks in this protein group pair. The
second left block contains 20 positions and the first right
block contains 12 positions. By the mapping method, the
positions 1 ~ 19 of the second left block can be aligned
with the positions 9 ~ 27 of block IPB001425B in BLOCKS,
and the positions 4 ~ 12 of the first right block can be
aligned with the positions 1 ~ 9 of block IPB003660A
in BLOCKS. Block IPB001425B is in the Bac rhodopsin
domain, and block IPB003660A is in the HAMP domain.

See Table 4 for more details. Our binding motif pair
can map into the domain pair (PF00672, PF01036) in
Pfam. Pfam shows that the HAMP domain interacts
with the Bac rhodopsin domain in protein complexes
such as l2s. l2s is the complex of *Natronobacterium
pharaonis* sensory rho-dopsin II (sRII) with receptor-
binding domain of HtrII. The X-ray structure of l2s
was obtained at 1.93 Å resolution [18] and it provided
an atomic picture of the first step of the signal transduction.
The interactions in the sRII-HtrII complex have been intensively investigated to find the signal relay mechanism
from the receptor to the transducer [19], [20], [21]. The
3D structure of the interactions is shown in Fig. 6(a)
and 6(b), which are generated by Protein Explorer [22].
TABLE 2
The distribution of the 59124 interacting protein group pairs in terms of $|A|$ and $|B|$. Each cell is the number of interacting group pairs for different $|A|$ and $|B|$.

| $|B|$ in [5, 9] | $[10, 14]$ | $[15, 19]$ | $[20, 24]$ | $[25, 29]$ | $[30, 34]$ | $[35, 39]$ | $\geq 40$ | SUM |
|----------------|------------|------------|------------|------------|------------|------------|----------|------|
| $|A|=5$        | 340        | 991        | 730        | 810        | 642        | 767        | 800      | 387  |
| 6             | 2860       | 2673       | 3321       | 1769       | 1770       | 1775       | 675      | 73   | 14920 |
| 7             | 511        | 1931       | 1946       | 1852       | 1755       | 268        | 0        | 0    | 8263  |
| 8             | 292        | 1569       | 2059       | 1837       | 939        | 7          | 0        | 0    | 5467  |
| 9             | 36         | 1417       | 2390       | 940        | 353        | 0          | 0        | 0    | 5136  |
| 10            | 0          | 1733       | 2998       | 1372       | 88         | 0          | 0        | 0    | 6191  |
| 11            | 0          | 1720       | 2615       | 554        | 0          | 0          | 0        | 0    | 4889  |
| 12            | 0          | 1303       | 2115       | 77         | 0          | 0          | 0        | 0    | 3695  |
| 13            | 0          | 875        | 1214       | 0          | 0          | 0          | 0        | 0    | 2088  |
| 14            | 0          | 204        | 528        | 0          | 0          | 0          | 0        | 0    | 732   |
| 15            | 0          | 0          | 1040       | 0          | 0          | 0          | 0        | 0    | 1040  |

Fig. 5. The distribution of the 59124 interacting protein group pairs ($A, B$) in terms of the size $|A||B|$. The size $|A||B|$ is divided into bins of 10 each. Each bar represents the total number of interacting protein group pairs in the bin.

TABLE 3
The mappings between the motifs and the two databases: BLOCKS and PRINTS. FPClose* uses BLOCKS 14.0 and PRINTS 37.0. Our PPIExtend method uses BLOCKS 14.3 and PRINTS 38.0. Each entry $a/b$ means the motifs are mapped to $a$ blocks(domains) in all $b$ blocks(domains) in the databases.

<table>
<thead>
<tr>
<th></th>
<th>BLOCKS</th>
<th>PRINTS</th>
<th>BOTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>blocks</td>
<td>domains</td>
<td>blocks</td>
</tr>
<tr>
<td>FPClose*</td>
<td>6408/24294</td>
<td>3128/4944</td>
<td>2174/11170</td>
</tr>
<tr>
<td>PPIExtend</td>
<td>9325/29767</td>
<td>4191/6149</td>
<td>2423/11435</td>
</tr>
</tbody>
</table>

The shortest residue-residue distance between the two motifs in a pair are also interesting. In protein complex 1h2s, there are two chains: chain A (1h2s_A) and chain B (1h2s_B). The left motif is located at positions 168–186 of 1h2s_A, and the right motif is located at positions 61–69 of 1h2s_B (Table 4). We downloaded the coordinate information of 1h2s from http://www.ebi.ac.uk/msd-srv/msdlite/atlas/summary/1h2s.html, and computed the residue-residue distances between the two motifs. The shortest residue-residue distance is 4.07 Å between atom 1346 of residue 177 in 1h2s_A and atom 2018 of residue 69 in protein 1h2s_B (Fig. 6(b)). The average shortest residue-residue distance is 9.17 Å. From these calculation and information, we may conclude that the positions 1–19 of the second left block and the positions 4–12 of the first right block are possibly interaction sites.

The second motif pair is derived from a protein group pair in which the left protein group contains 6 proteins and the right protein group contains 8 proteins. There are 43 interactions between the two groups of proteins. The $p$-value of the protein group pair is less than $1.09 \times 10^{-12}$. One left block containing 30 positions and one right block containing 32 positions in the protein group pair are reported by PROTOMAT.
The positions 3 – 29 of the left block can be well aligned with the positions 3 – 29 of block IPB005255E in BLOCKS, and the positions 5 – 31 of the right block can be well aligned with the positions 3 – 29 of block IPB005255E in BLOCKS. The two blocks IPB005255E and IPB005255E can be mapped to the interacting PdxA domain pair (PF04166, PF04166) in iPam. The domain pair has protein-protein interactions in protein complexes such as 1ps6. The structure of *Escherichia coli* PdxA (1ps6) was reported by Sivaraman et al. [23]. PdxA is an important enzyme involved in the biosynthesis pathway of pyridoxal 5’-phosphate, the catalytically active form of vitamin B<sub>6</sub>, which is an essential cofactor of numerous metabolic enzymes. In the pathway for the *de novo* synthesis of pyridoxal 5’-phosphate utilized by *Escherichia coli*, the fourth step is catalyzed by PdxA. PdxA is very important in the study of the *de novo* synthesis of vitamin B<sub>6</sub> [24], [25]. The 3D structure of PdxA was downloaded from http://www.ebi.ac.uk/msd-srv/msd-lite/atlas/summary/1ps6.html. The shortest residue-residue distance between the two motifs is 2.3 Å between atom 2038 of residue 275 in 1ps6_A and atom 4455 of residue 276 in protein 1ps6_B (Fig. 6(d)). The average shortest residue-residue distance is 10.15 Å. The mapping shows that the left block and the right block are another pair of possible interaction sites.

6 Conclusion and Open Problem

We have proved that both the maximum vertex quasi-biclique problem and the maximum balanced quasi-biclique problem are NP-hard. The NP-hardness result for the first problem is surprising since the maximum vertex biclique problem is polynomial time solvable. In this paper, we have demonstrated the usefulness of the topology information of PPI networks for finding the binding motifs at interaction sites.

In particular, we are very interested in the ranking of motif pairs and biological interpretation of top-ranked patterns. This problem is currently open as there are several difficult factors to consider. For example, one factor is whether the length of the pattern is more important than the occurrence of the pattern. This needs a comprehensive study on the sensitivity and specificity of the top-ranked motif pairs against lab-confirmed protein binding sites. As lab-confirmed protein binding sites are incomplete, we have much space left for the future evaluation of the proposed method.

The other open problem is to integrate other information sources, such as topology structures in PPI networks, protein functions, and gene ontology localization information for identifying possible interaction sites.

### Table 4

<table>
<thead>
<tr>
<th>Domain</th>
<th>Left block l18493xB</th>
<th>Right block r18493xA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>distance from previous block = (4, 396)</td>
<td>distance from previous block = (7, 177)</td>
</tr>
<tr>
<td></td>
<td>DE none</td>
<td>DE none</td>
</tr>
<tr>
<td></td>
<td>BL IIK motif = [6, 0, 17] motifmat = [1, 1, -10]</td>
<td>BL LLL motif = [6, 0, 17] motifmat = [1, 1, -10]</td>
</tr>
<tr>
<td></td>
<td>width = 20 seqs = 7</td>
<td>width = 12 seqs = 8</td>
</tr>
<tr>
<td></td>
<td>DIP: 8095N (206) VIGILISYTKATCDMLAGK</td>
<td>DIP: 7371N (356) MILILAQFWA1APIGEGK</td>
</tr>
<tr>
<td></td>
<td>DIP: 4973N (536) MIIILIAQFWA1APIGEGK</td>
<td>DIP: 4973N (536) MIIILIAQFWA1APIGEGK</td>
</tr>
<tr>
<td></td>
<td>DIP: 5150N (417) LKIDEINNOKGDADKYYK</td>
<td>DIP: 3571N (384) I11ALIVT1LMFLMGLN1AK</td>
</tr>
<tr>
<td></td>
<td>DIP: 676N (402) VI1AIFVEVSTIVSGVK</td>
<td>DIP: 676N (402) VI1AIFVEVSTIVSGVK</td>
</tr>
<tr>
<td>PdxA</td>
<td>Bac_rhodopsin:</td>
<td>Bac_rhodopsin:</td>
</tr>
<tr>
<td></td>
<td>VVLW1YPPVW1LGPGPV</td>
<td>VVLW1YPPVW1LGPGPV</td>
</tr>
<tr>
<td></td>
<td>pdb 1h2s_A, B</td>
<td>pdb 1h2s_A, B</td>
</tr>
<tr>
<td></td>
<td>distance from previous block = (4, 396)</td>
<td>distance from previous block = (7, 177)</td>
</tr>
<tr>
<td></td>
<td>DE none</td>
<td>DE none</td>
</tr>
<tr>
<td></td>
<td>BL IIK motif = [6, 0, 17] motifmat = [1, 1, -10]</td>
<td>BL LLL motif = [6, 0, 17] motifmat = [1, 1, -10]</td>
</tr>
<tr>
<td></td>
<td>width = 20 seqs = 7</td>
<td>width = 12 seqs = 8</td>
</tr>
<tr>
<td></td>
<td>DIP: 8095N (206) VIGILISYTKATCDMLAGK</td>
<td>DIP: 7371N (356) MILILAQFWA1APIGEGK</td>
</tr>
<tr>
<td></td>
<td>DIP: 4973N (536) MIIILIAQFWA1APIGEGK</td>
<td>DIP: 4973N (536) MIIILIAQFWA1APIGEGK</td>
</tr>
<tr>
<td></td>
<td>DIP: 5150N (417) LKIDEINNOKGDADKYYK</td>
<td>DIP: 3571N (384) I11ALIVT1LMFLMGLN1AK</td>
</tr>
<tr>
<td></td>
<td>DIP: 676N (402) VI1AIFVEVSTIVSGVK</td>
<td>DIP: 676N (402) VI1AIFVEVSTIVSGVK</td>
</tr>
</tbody>
</table>

### Appendix A

#### Proof of Theorem 1

**Theorem 1:** Let $G = (X \cup Y, E)$ be a random graph with $|X| = |Y| = n$, where for each pair of vertices $x \in X$ and $y \in Y$, $(x, y)$ is chosen, randomly and independently, to be an edge in $E$ with probability $\frac{2}{3}$. When $n \to \infty$, with high probability, $G$ is a $\frac{1}{2}$-quasi-biclique, and $G$ does not contain any biclique $G’ = (X’ \cup Y’, E’)$ with $|X’| \geq 2 \log n$ and $|Y’| \geq 2 \log n$.

**Proof:** For each subset $X’ \subset X$ with size $2 \log n$, let random variable $Z$ be the number of vertices in $Y$ that are adjacent to all vertices in $X’$. The expectation of $Z$ is

$$
E(Z) = n \left( \frac{2}{3} \right)^{2 \log n} \leq n^{-0.16}
$$

and the variance of $Z$ is

$$
Var(Z) = n \left( \frac{2}{3} \right)^{2 \log n} \cdot \left( 1 - \left( \frac{2}{3} \right)^{2 \log n} \right) \leq n^{-0.16}.
$$

When $n \to \infty$, by the central limit theorem, the probability that $Z \geq 2 \log n$ is

$$
Pr(Z \geq 2 \log n) \leq Q(2n^{0.08} \log n) \leq n^{-2n^{0.16} \log n}.
$$
Fig. 6. (a) The 3D structure (best viewed in color) of the interactions between the Bac_rhodopsin domain and the HAMP domain in 1h2s. The left part is chain A and contains the Bac_rhodopsin domain. The right part is chain B and contains the HAMP domain. (b) The backbone structure of the interactions between segment [168V,186A] in 1h2s_A and segment [61V,69I] in 1h2s_B. (c) The 3D structure of the interactions between two PdxA domains in chain A and chain B in 1ps6. The top part is chain A and the bottom part is chain B. (d) The backbone structure of the interactions between segment [261V, 287G] in 1ps6_A and segment [261V,287G] in 1ps6_B.
where \( Q(x) \) is the \( Q \)-function defined as the probability that a standard normal random variable exceeds \( x \), and the last inequality is from \( Q(x) \leq e^{-x^2/2} \).

There are \( \binom{n}{2\log n} \) different ways of choosing \( X' \) from \( X \), the probability that \( G \) does not contain any biclique \( G' = (X' \cup Y', E') \) with \( |X'| \geq 2 \log n \) and \( |Y'| \geq 2 \log n \) is greater than \( 1 - \binom{n}{2\log n} n^{-2n^{0.16} \log n} \geq 1 - n^{-2 \log n (n^{0.16} - 1)} \geq e^{-\frac{1}{32}}. \)

For each vertex \( x \in X \), let random variable \( D \) be the degree of \( x \). The expectation of \( D \) is \( \frac{2}{3}n \) and the variance of \( D \) is \( \frac{2}{3}(1 - \frac{2}{3})n = \frac{4}{9}n \). When \( n \to \infty \), by the central limit theorem, the probability that \( D < \frac{4}{9}n \) is

\[
Pr(D < \frac{1}{2}n) \leq Q(\frac{1}{4} \sqrt{n}) \leq e^{-\frac{1}{32}}.
\]

There are \( 2n \) different vertices in \( X \) and \( Y \), the probability that all degrees of all vertices are greater than \( \frac{1}{2}n \) is greater than \( 1 - 2ne^{-\frac{1}{32}} \).

From the above analysis, the probability that \( G \) is a \( \frac{1}{4} \)-quasi-biclique and the degrees \( D \) of all vertices are greater than \( \frac{1}{2}n \) is greater than \( 1 - 2ne^{-\frac{1}{32}} \).

**APPENDIX B**

**PROOF OF LEMMA 1**

**Lemma 1**: Let \( (X' \cup Y', E') \) be a maximum vertex \( \frac{2}{3} \)-quasi-biclique of the constructed bipartite graph \( G(X \cup Y, E) \). Set \( Y'_a = Y_a \cap Y', \ Y'_b = Y_b \cap Y', \) and \( Y' = Y_a \cup Y_b \). If \( |X'| + |Y'| \geq 3m(3m+3) \frac{q}{p} + (m+1) \frac{q}{p} \), we have \( |Y'_a| = |Y'_b| = m \), \( Y'_c = Y_c \) and \( X' = X \).

**Proof**: (1) Since each vertex in \( Y_c \) is adjacent to all vertices in \( X \), we have \( Y'_c = Y_c \). (If a vertex \( v \in Y_c \) is not in \( Y' \), then we can add \( v \) into the optimal solution.)

(2) Now, we prove that \( |Y'_a| = |Y'_b| = m \). We need two steps.

(2.1) First, we prove \( |Y'_a| + |Y'_b| = 2m \) by contradiction. There are two cases.

**Case 1**: \( |Y'_a| + |Y'_b| < 2m \). Then

\[
|X'| + |Y'| \leq |X| + |Y'_a| + |Y'_b| + |Y_c| < 3m(3m+3) \frac{q}{p} + (m+1) \frac{q}{p}.
\]

It is a contradiction.

**Case 2**: \( |Y'_a| + |Y'_b| > 2m \). From the assumption \( |Y'_a| + |Y'_b| > 2m \), we can find an integer \( k \geq 1 \) such that if \( |Y'_a| + |Y'_b| \) is even, then \( |Y'_a| + |Y'_b| = 2(m+k), \) whereas if \( |Y'_a| + |Y'_b| \) is odd, then \( |Y'_a| + |Y'_b| = 2(m+k) - 1 \). That is, we can find an integer \( k \geq 1 \) such that \( 2(m+k) - 1 \leq |Y'_a| + |Y'_b| \leq 2(m+k) \). Suppose \( |Y'_a| \leq |Y'_b| \), we have \( |Y'_a| \leq m+k \).

Now we consider \( X'_1 \), one of the 3m disjointed subsets of \( X \). Let \( X'_{1,a} = X_{1,a} \cap X' \). We can prove \( X'_{1,a} \neq X_{1,a} \) by contradiction. Suppose \( X'_{1,a} = X_{1,a} \). For each \( x \in X'_{1,a}, \)

\[
d(x, Y') \geq (1 - \frac{p}{q})|Y'|
\]

\[
\geq (1 - \frac{p}{q})((m+1) \frac{q}{p} + 2k - 1)
\]

\[
\geq (m+1) \frac{q}{p} - (m+1) + k - \frac{1}{2}.
\]

The last inequality is from \( \frac{p}{q} \leq \frac{1}{2} \). Since the degree \( d(x, Y') \) is an integer and \( (m+1)p \), we have \( d(x, Y') \geq (m+1) \frac{q}{p} - (m+1) + k \). From the construction of \( G \), \( d(x, Y'_a) = (m+1) \frac{q}{p} - 2m \) and \( d(x, Y'_b) = 0 \). Thus, \( d(x, Y'_a) \geq m + k - 1 \). From the assumption, \( |X'_{1,a}| = |X_{1,a}| = 3m \), there are at least \( 3m(m+k-1) = m^2 + 6k - 3m \) edges between \( X'_{1,a} \) and \( Y'_a \). However, \( |Y'_a| \leq m+k \) and for each vertex \( y \in Y'_a \), \( d(y, X'_{1,a}) = 3m - 3. \) From this point, there are at most \( (3m-3)(m+k) = m^2 + 2m - 3k \) edges between \( X'_{1,a} \) and \( Y'_a \). It is a contradiction. Therefore, \( X'_{1,a} \neq X_{1,a}. \)

From the above discussion, at least one vertex in \( X_{1,a} \) is not in \( X'_{1,a}. \) In the 3m subsets \( X_1, X_2, \ldots, X_{3m} \), at least \( 3m \) vertices in \( X \) are not in \( X' \). By the assumption \( n < 2m \), we have \( |Y'| < 2m + (m+1) \frac{q}{p} \). Therefore, we get a contradiction that \( |X'| + |Y'| < 3m(3m+3) \frac{q}{p} + (m+1) \frac{q}{p} \).

(3) By \( |Y'| = |Y'_a| + |Y'_b| + |Y'_c| = (m+1) \frac{q}{p}, \) we have \( |X'| = 3m(3m+3) \frac{q}{p} \). That is, \( X' = X \).
Suppose that there exists a vertex $x$ in $G$. Then $|x|^2$-quasi-biclique problem is NP-hard. By Lemma 1, we have $|x|^2 = m$, $Y_a = Y_c$ and $X = Y$. For each vertex $x_i \in X$, the degree $d(x_i, Y_a) \geq (m + 1)^2 - m$. From the construction of $G$, $d(x_i, Y_a) = (m + 1)^2 - 2m$ and $d(x_i, Y_a) = 0$. Thus, $d(x_i, Y_a) = m + 1$. Since $G$ is adjacent to all vertices in $Y_c$ and $3m - 3$ vertices in $Y_a$, we can select the $3m$ triples corresponding to $Y_a$ as the solution $T'$. For each $x \in X$, $d(x, Y_a) = 3m - 3$. Therefore, $T'$ is a perfect 3 cover of $S$.

**ACKNOWLEDGMENTS**

We thank Dr. Haiquan Li for providing us the protein interaction data. We also thank Dr. Shmuel Pietrokowski for giving us the LAMA program.

Lusheng Wang is fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 121207), and was the corresponding author. Jinyan Li is fully supported a NTU (Nanyang Technological University) Tier-1 start-up grant (ref. no. RG66/07).

**References**


[25] Lusheng Wang received the PhD degree from McMaster University, Hamilton, Ontario, Canada, in 1995. Currently, he is an associate professor in the Department of Computer Science, City University of Hong Kong. His research interests include algorithms, bioinformatics, and computational biology. He is a member of the IEEE.

Jinyan Li received the PhD degree in computer science from the University of Melbourne in 2001. He is an associate professor in the School of Computer Engineering, Nanyang Technological University. He has published about 80 articles in the fields of bioinformatics, data mining, and machine learning. His research focus is currently on protein interaction networks and protein binding in computational biology, on statistically important discriminative patterns in data mining, and on interaction subgraphs in graph theory. Detailed research work and publications can be found in his website: www.ntu.edu.sg/home/jyli.

Xiaowen Liu received the PhD degree in computer science from City University of Hong Kong, Hong Kong, P.R. China, in 2007. He is now a Postdoctoral Fellow at Department of Computer Science, University of Western Ontario, London, Ontario, Canada. His research interests include bioinformatics and algorithms.