Quasi-bicliques: Complexity and Binding Pairs

Xiaowen Liu[1,3], Jinyan Li[2], and Lusheng Wang[1,⋆]

[1] Department of Computer Science,
City University of Hong Kong, Kowloon, Hong Kong
lwang@cs.cityu.edu.hk
[2] School of Computer Engineering,
Nanyang Technological University, Singapore 639798
[3] Department of Computer Science,
University of Western Ontario, Canada

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 [16]. 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.

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 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.

⋆ Corresponding author.

X. Hu and J. Wang (Eds.): COCOON 2008, LNCS 5092, pp. 255–264, 2008.

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 [15], and they showed that the complementary domains can form near complete bipartite subgraphs in PPI networks. Morrison *et al.* proposed a lock-and-key model which is also based on the concept of maximal complete bipartite subgraphs [11]. Very recently, Andreopoulos *et al.* used clusters in PPI networks for identifying locally significant protein mediators [1]. 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 focused on protein function prediction [4,8].

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 [10]. The first step of the algorithm [10] 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 [10] 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 [16]. 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 [6]. The maximum edge biclique problem is proved to be NP-hard as well [12].

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 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ 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. It is allowed that an edge $(u, v) \in \mathcal{E}$ and $u = v$, which is called a self-loop. Since $\mathcal{G}$ is an undirected graph, any edge $(u, v) \in \mathcal{E}$ implies $(v, u) \in \mathcal{E}$. For a selected edge (u, v) in $\mathcal{G}$, in order to find the two groups of proteins having the similar pairs of binding sites, we translate the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into a bipartite graph. Let $X = \{x|(x, v) \in \mathcal{E}\}$, $Y_1 = \{y|(u, y) \in \mathcal{E} \& y \notin X\}$ and $Y_2 = \{w|(u, w) \in \mathcal{E} \& w \in X\}$. For a vertex $w \in Y_2$, w is incident to both u and v in $\mathcal{G}$, so both X and Y_2 contain w. We remain 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 $\overline{Y_2}$. Let $Y = Y_1 \cup \overline{Y_2}$ and $E = \{(x, y)|(x, y) \in \mathcal{E} \& x \in X \& y \in Y_1\} \cup \{(x, \overline{w})|(x, w) \in \mathcal{E} \& x \in X \& \overline{w} \in \overline{Y_2}\}$. 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 $\mathcal{G}$. In $\mathcal{G}$, 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 self-loops. Self-loops are very common in practice. When self-loop appears, the protein contains two complementary motifs simultaneously.

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 $d(x, Y') = |\{y|y \in Y' \text{ and } (x, y) \in E\}|$. Similarly, for a vertex $y \in Y$ and $X' \subseteq X$, we use $d(y, X')$ to denote $|\{x|x \in X' \ (x, y) \in E\}|$. Now, we are ready to define the *δ-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 δ-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 δ-quasi-biclique in G corresponds to two subsets of vertices, say, subset A and subset B, in $\mathcal{G}$. In $\mathcal{G}$, 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 δ-quasi-biclique and there is an edge between the two corresponding vertices, the protein contains two complementary motifs simultaneously. In our experiments, we observe that

about 22% of the δ-quasi-bicliques produced by our program contain self-loop proteins.

In many applications, due to various reasons, some edges in a biclique may missing and a biclique becomes a quasi-biclique. Thus, finding quasi-bicliques is more important in practice. The following theorem shows 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 $\frac{2}{3}$. 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 $|Y'| \geq 2 \log n$.*

In the biological context, Theorem 1 indicates that some large interacting protein groups cannot be obtained simply by finding a maximal biclique. 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.

3 NP-Hardness

In this section, we study the following two problems: the maximum vertex quasi-biclique problem and the maximum balanced quasi-biclique problem.

3.1 The Maximum Vertex Quasi-biclique Problem

The maximum vertex quasi-biclique problem is defined as follows.

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

The maximum vertex biclique problem, where $\delta = 0$, can be solved in polynomial time [16]. We can prove that the maximum vertex δ-quasi-biclique problem when $\delta > 0$ is NP-hard. The reduction is from $X3C$ (Exact Cover by 3-Sets), which is known to be NP-hard [9].

Exact Cover by 3-Sets
Instance: A finite set S of $3m$ elements, and a collection T of n triples (3-element subsets of S).
Objective: To determine whether T contains an exact cover of S, i.e. a sub-collection $T' \subseteq T$ such that every element of S occurs in exactly one triple in T'.

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.*

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* δ*-quasi-biclique problem is to find* $X' \subseteq X$ *and* $Y' \subseteq Y$ *such that the* $X' \cup Y'$ *induced subgraph is a* δ*-quasi-biclique and* $|X'| = |Y'|$ *is maximized.*

We can also prove that the maximum balanced quasi-biclique problem is NP-hard.

Theorem 3. *For any constant integers* $p > 0$ *and* $q > 0$ *such that* $0 < \frac{p}{q} \leq \frac{1}{2}$, *the maximum balanced* $\frac{p}{q}$*-quasi-biclique problem is NP-hard.*

4 The Heuristic Algorithm

In practice, we need to find large quasi-bicliques in PPI networks. Here, we propose a heuristic algorithm to find large quasi-bicliques. Consider a PPI network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Our heuristic algorithm has two steps. First, we construct the bipartite graph 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 $G = (X \cup Y, E)$ from $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and an edge (u, v).

In the algorithm, we have two parameters δ and τ, which control the quality and sizes of the quasi-bicliques. We can 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', and 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 τ. 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. 1. 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 Step $3 - 5$ and Step 10 is $O(n)$, and the time complexity of Step $6 - 9$ is $O(n^2)$. So the time complexity of Step $3 - 10$ is $O(n^2)$. Step $3 - 10$ is repeated $O(n)$ times. Therefore, the time complexity of the whole algorithm is $O(n^3)$. To speed up the algorithm, we can do multiple vertex

The Greedy Algorithm

Input A bipartite graph $(X \cup Y, E)$ and two parameters δ and τ.
Output A set of δ-quasi-bicliques $(X' \cup Y', E')$ with $|X'| \geq \tau$ and $|Y'| \geq \tau$.

1. Let $X' = \phi$ and $Y' = Y$.
2. **while** $|Y'| \geq \tau$ and $X' \neq X$ **do**
3. Find the vertex $x \in X - X'$ with the maximum degree $d(x, Y')$.
4. Add x into X', $X' = X' \cup \{x\}$, and delete from Y' all vertices $y \in Y'$ such that $d(y, X') < (1 - \delta)|X'|$.
5. $X'' = X'$ and $Y'' = Y'$.
6. **repeat**
7. Find the vertex $x \in X - X''$ with the maximum degree $d(x, Y'')$. If $d(x, Y'') \geq (1 - \delta)|Y''|$, add x to X'', $X'' = X'' \cup \{x\}$.
8. Find the vertex $y \in Y - Y''$ with the maximum degree $d(y, X'')$. If $d(y, X'') \geq (1 - \delta)|X''|$, add y to Y'', $Y'' = Y'' \cup \{y\}$.
9. **until** no vertex is added in the step 7 and 8.
10. **if** $|X''| \geq \tau$, $|Y''| \geq \tau$, for each $x \in X''$, $d(x, Y'') \geq (1 - \delta)|Y''|$, and for each $y \in Y''$, $d(y, X'') \geq (1 - \delta)|X''|$, output $(X'' \cup Y'')$ as a quasi-biclique.

Fig. 1. The greedy algorithm

addition in the algorithm. To do multiple vertex addition, we have an integer parameter $\alpha > 0$, and change the algorithm as follows. In Step 3, we select the best α vertices in $X - X'$ and add all the α vertices into X' in Step 4.

5 Experiments

We implemented the heuristic algorithm in JAVA. The software is called PPIExtend. As shown in the last step of the algorithm, some vertices in X'' may be adjacent to less than $(1 - \delta)|Y''|$ vertices in $|Y''|$, but the average degree of the vertices in X'' is no less than $(1 - \delta)|Y''|$. Similarly, some vertices in Y'' may be adjacent to less than $(1 - \delta)|X''|$ vertices in $|X''|$, but the average degree of the vertices in Y'' is no less than $(1 - \delta)|X''|$. In our experiments, these quasi-bicliques are still output to get more useful quasi-bicliques. Our algorithm consists of two steps: (i) find quasi-interacting protein group pairs, then (ii) use the sequence data of these protein groups to find conserved motifs. To evaluate the motif pairs found by our algorithm, we follow the two validation methods in [10]. First, we compare the single motifs with two block databases: BLOCKS [13] and PRINTS [3]. Second, we map our motif pairs into domain-domain interaction pairs in domain-domain interaction database *i*Pfam [5]. We also study the overlapping between the protein group pairs found by PPIExtend and the protein group pairs found by FPClose* in [10]. Two interesting case studies on binding motif pairs are then followed.

5.1 Motif Mapping with BLOCKS, PRINTS, and *i*Pfam

The protein interaction data of *Saccharomyces cerevisiace* (yeast) was downloaded from http://research.i2r.a-star.edu.sg/BindingMotifPairs/resources. The

Table 1. 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 that the motifs are mapped to a blocks(domains) in all b blocks(domains) in the databases.

	BLOCKS		PRINTS		BOTH	
	blocks	domains	blocks	domains	blocks	domains
FPClose*	6408/24294	3128/4944	2174/11170	1093/1850	24.1%	62.1%
PPIExtend	9325/29767	4191/6149	2423/11435	1160/1900	28.5%	66.4%

data includes 10640 experimentally determined physical interactions of 4959 proteins in *Saccharomyces cerevisiace*. We set $\delta = 0.1$ and $\tau = 5$ and $\alpha = 5$. The greedy algorithm produced $59,124$ interacting protein group pairs. In all the protein group pairs, $13,266$ pairs (about 22%) contain self-loops, and a large number of protein group pairs (about 78%) do not contain self-loops. We then used PROTOMAT [13] to find the conserved motifs within the protein groups, as PROTOMAT is an algorithm that can find the multi-alignment of a group of sequences and can output the conserved motifs. By using the default parameters, PROTOMAT output $220,393$ motifs from our $59,124$ pairs of interacting protein groups.

The LAMA program [13] is a dynamic programming method that can find the optimal local alignment of two blocks where the Z-score is computed to evaluate the alignments. We make use of it to compare the block databases with our motifs. The default threshold of Z-score was used in the experiments. The mappings between the motifs and the two databases are also compared between our method and the FPClose* method [7] that was used in [10]. The comparison results are reported in Table 1. From this table, we can see that our method has more mappings to BLOCKS and PRINTS than FPClose* does. This indicates that the use of quasi-bicliques is effective to find more number of motif pairs at interaction sites.

The *i*Pfam database is built on top of the Pfam database [14] which stores the information of protein domain-domain interactions. To examine whether our binding motif pairs can match some pairs of interacting domains in *i*Pfam, we map our binding motif pairs through the integrated protein family database InterPro [2] which integrates a number of databases. We first map our motifs to domains in the BLOCKS and PRINTS databases. Then, we map the protein groups in BLOCKS and PRINTS to protein groups in InterPro. Finally, we map the protein groups in InterPro to domains in the Pfam database. In this way, our motif pairs can be well mapped to Pfam domain pairs. In fact, we strictly follow the procedure as suggested in [10] to map motif pairs to domain pairs.

In the experiments, we used Pfam 20.0 and *i*Pfam 20.0. We observed that the motif pairs found by our PPIExtend method can map to 81 distinct domain pairs in *i*Pfam. This is a much bigger number than 18 number of domain pairs that can be mapped from the motif pairs reported in [10]. This significant increase is mainly attributed to the use of quasi-bicliques because using quasi-bicliques can find many interacting protein group pairs with larger sizes. (See Theorem 1.)

Table 2. Left block l18493xB and right block r18493xA (output of PROTOMAT for protein group pair No.18493) aligning with the Bac_rhodopsin domain and the HAMP domain respectively. For brevity, only 5 sequences in each of the two blocks are shown. In the Bac_rhodopsin domain and HAMP domain, the capital letters are the amino acids with the highest frequency in each position. Pdb 1h2s_A and pdb 1h2s_B are chain A and chain B in protein complex 1h2s, respectively.

```
AC  l18493xB;
    distance from previous block=(4,396)
DE  none
BL  IIK motif=[6,0,17] motomat=[1,1,-10]
    width=20 seqs=7
 DIP:8095N  ( 206) VIGILIISYTKATCDMLAGK
 DIP:4973N  ( 536) MILILIAQFWVAIAPIGEGK
 DIP:5150N  ( 417) LIKDEINNDKKDNADDKYIK
 DIP:5371N  ( 384) IILALIVTILWFMLRGNTAK
 DIP:676N   ( 402) VIVAWIFFVVSFVTTSSVGK
 ...
 pdb 1h2s_A ( 168) VILWAIYPFIWLLGPPGVA
 Bac_rhodopsin:    VVLWLAYPVVWLLGPEGIG
```

```
AC  r18493xA;
    distance from previous block=(7,177)
DE  none
BL  LLL motif=[6,0,17] motomat=[1,1,-10]
    width=12 seqs=8
 DIP:7371N  (  10) LALIILYLSIPL
 DIP:8128N  (  35) LSLRFLALIFDL
 DIP:4176N  ( 106) LVLTSLSLTLLL
 DIP:7280N  (  11) LSLFLPPVAVFL
 DIP:5331N  ( 178) LSFFVLCGLARL
 ...
 pdb 1h2s_B (  61)    VSAILGLII
 HAMP:                IALLLALLL
```

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 is a large portion of the pairs of domain interactions found by our method, we still found many other domain-domain interactions that are not self-loops.

We also examined the overlapping between the protein group pairs found by PPIExtend and the protein group pairs found by FPClose* in [10]. The overlapping criteria is that if one protein group pair G_1 contains more than 90% proteins of another protein group pair G_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 our protein group pairs. However, there are $38,305$ protein group pairs found by PPIExtend that cannot be covered by any protein group pairs found by FPClose*. This result demonstrates that our method not only can find much more number of bicliques and with larger size, also can find more binding motif pairs than the method presented in [10].

5.2 Case Studies

In this section, we present 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 1.57×10^{-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 2 for more details. Our motif pair can map into the domain pair (PF00672, PF01036) in *i*Pfam. *i*Pfam shows that the HAMP domain interacts with the Bac_rhodopsin domain in protein complexes such as lh2s.

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 1.09×10^{-122}. The motif pair can be mapped to the interacting PdxA domain pair (PF04166, PF04166) in *i*Pfam. The domain pair has protein-protein interactions in protein complexes such as 1ps6.

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. However, the hardness of the maximum edge quasi-biclique problem is still an open problem. In this paper, we have shown the usefulness of the topology information of PPI networks for modeling the binding motifs at interaction sites. In future work, we will focus on how to integrate other information sources, such as protein functions and gene ontology localization information for identifying possible interaction sites.

Acknowledgments. We thank Dr. Haiquan Li for providing us the protein interaction data. We also thank Dr. Shmuel Pietrokovski 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].

References

1. Andreopoulos, B., An, A., Wang, X., Faloutsos, M., Schroeder, M.: Clustering by Common Friends Finds Locally Significant Proteins Mediating Modules. Bioinformatics 23(9), 1124–1131 (2007)
2. Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J., Zdobnov, E.M.: The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites. Nucleic Acids Research 29(1), 37–40 (2001)
3. Attwood, T.K., Beck, M.E.: PRINTS-a Protein Motif Fingerprint Database. Protein Engineering, Design and Selection 7, 841–848 (1994)
4. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., Chen, R.: Topological Structure Analysis of the Protein-Protein Interaction Network in Budding Yeast. Nucleic Acids Research 31(9), 2443–2450 (2003)

5. Finn, R.D., Marshall, M., Bateman, A.: iPfam: Visualization of Protein-Protein Interactions in PDB at Domain and Amino Acid Resolutions. Bioinformatics 21(3), 410–412 (2005)
6. Garey, M.R., Johnson, D.S.: Computers and Intractability, A Guide to the Theory of NP-Completeness. Freeman, San Francisco (1979)
7. Grahne, G., Zhu, J.: Efficiently using Prefix-Trees in Mining Frequent Itemsets. In: Proceedings of the Workshop on Frequent Itemset Mining Implementations (FIMI) (2003)
8. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of Prediction Sccuracy of Protein Gunction From Protein–Protein Interaction Data. Yeast 18(6), 523–531 (2001)
9. Karp, R.M.: Reducibility among Combinatorial Problems. In: Miller, R.E., Thatcher, J.W. (eds.) Complexity of Computer Computations, pp. 85–103 (1972)
10. Li, H., Li, J., Wang, L.: Discovering Motif Pairs at Interaction Sites from Protein Sequences on a Proteome-Wide Scale. Bioinformatics 22(8), 989–996 (2006)
11. Morrison, J.L., Breitling, R., Higham, D.J., Gilbert, D.R.: A Lock-and-Key Model for Protein-Protein Interactions. Bioinformatics 22(16), 2012–2019 (2006)
12. Peeters, R.: The Maximum Edge Biclique Problem is NP-Vomplete. Discrete Applied Mathematics 131(3), 651–654 (2003)
13. Pietrokovski, S.: Searching Databases of Conserved Sequence Regions by Aligning Protein Multiple-Alignments. Nucleic Acids Research 24, 3836–3845 (1996)
14. Sonnhammer, E.L.L., Eddy, S.R., Durbin, R.: Pfam: A Vomprehensive Database of Protein Domain Families Based on Seed Alignments. Proteins: Structure, Function and Genetics 28, 405–420 (1997)
15. Thomas, A., Cannings, R., Monk, N.A.M., Cannings, C.: On the Structure of Protein-Protein Interaction Networks. Biochemical Society Transactions 31(Pt 6), 1491–1496 (2003)
16. Yannakakis, M.: Node Deletion Problems on Bipartite Graphs. SIAM Journal on Computing 10, 310–327 (1981)