Protein complex prediction via improved verification methods using constrained domain-domain matching

Yang Zhao,†1 Morihiro Hayashida,†1 Jose Nacher,†2
Hiroshi Nagamochi†3 and Tatsuya Akutsu†1

In the field of functional genomics, identifying protein complexes within large-scale protein-protein interaction networks is one of the central research objectives due to limited availability of known protein complexes. Many approaches such as MCL and MCODE have been developed, and enabled researchers to detect protein complexes from protein-protein interaction networks. However, structural constraints of proteins have not been taken into consideration, and as a result, incorrect proteins were often extracted in predicted complexes. In order to prevent generation of too many erroneous complexes, Ozawa et al. proposed a verification method of protein complexes by introducing a constraint that a domain interacts with at most one other domain.

In this technical report, we propose an improved integer programming-based method based on the idea that a candidate complex should not be divided into many small complexes. Furthermore, we enhance this method by combining with maximal components and extreme sets in graph theory. Comparison with the method by Ozawa et al. is conducted to show the advantage of our methods. The results suggest that the proposed methods outperform their method.

1. Introduction

Enormous amounts of protein-protein interaction (PPI) data are available for researchers to understand important principles of cellular organization and biological function with the rapid development of cell biology and systems biology, An inevitable consequence of this wealth of data goes to the need for efficient methods to identify important portions of these data. Protein complexes are known as clusters of multiple proteins linked by non-covalent physical protein-protein interactions that generally correspond to dense regions within PPI networks. As PPI data grows rapidly, identifying protein complexes within PPI networks becomes necessary and important due to limited availability of known protein complexes.

Recent approaches enable researchers to detect known and unknown protein complexes within PPI networks. We give a brief overview of state-of-the-art methods in identification of protein complexes. These methods often extract dense subgraphs in PPI networks as protein complexes since proteins in complexes are highly interactive with each other. Most methods for predicting protein complexes have been developed based on graph theory. The MCL algorithm as a novel graph clustering approach categorizes member proteins within large databases based on precomputed sequence similarity information[1]. Another graph theoretic clustering algorithm, MCODE, detects densely connected regions as molecular complexes in large PPI networks based on connectivity data[2]. Maruyama et al. proposed NWE (Node-Weighted Expansion of clusters of proteins) by introducing a random walk with restarts with a cluster of proteins[3].

However, one problem that current methods face is that they detect dense regions as protein complexes without taking into account of structural constraints of proteins. Therefore, methods considering multiple domains of proteins and topology of PPIs are desired to improve the precision of predicting protein complexes, where the precision of prediction methods is important for understanding biological systems because protein complexes often play crucial roles in cellular mechanism. So far, several computational methods have been proposed to verify protein complexes. These methods have assessed the validation of individual interaction based on the topology of PPI networks. However, almost all of the existing methods have paid no attention to the structural constraint of proteins in PPI networks, which resulted in low precision. The method proposed by[4] has verified and reconstructed the topology of domain-domain interactions in PPI networks. This method makes use of the concept that proteins in candidates each of whose domains participates only in a single interaction can form a valid protein complex. In terms of this concept, this approach seeks for optimal combinations of domain-domain interactions (DDIs) in the complex candidates

†1 Bioinformatics Center, Institute for Chemical Research, Kyoto University
†2 Department of Complex and Intelligent Systems, Future University-Hakodate
†3 Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University

predicted from other existing methods, by using integer linear programming. As a result, this optimization problem extracts subgraphs from complex candidates that contain more than one proteins connected by more than one DDI as verified protein complexes. Although this approach has achieved a relatively high precision, it still outputs a number of false positives.

In this technical report, we propose a novel formulation of integer programming based on the idea that a candidate complex should not be divided into many small complexes, and improve the method by Ozawa et al. for verifying candidate complexes predicted by graph clustering methods. In addition, we use maximal components and extreme sets that are defined based on edge connectivity in graph theory[5]. Since the internal proteins of a maximal component are connected more strongly with each other than with any other external proteins as well as an extreme set, they are expected to be useful to further increase the precision. We implement this improved IP-based method and the combination methods with maximal components and extreme sets, and perform several computational experiments. Comparison with the existing method is also conducted to confirm the advantage of our methods. Finally, we discuss the results of our proposed methods.

2. Methods

As mentioned in the previous section, Ozawa et al. proposed an integer programming (IP)-based method for verifying candidate complexes by maximizing the number of protein-protein interactions[4]. In this technical report, we propose a novel formulation of integer programming based on the idea that a candidate complex should not be divided into many small complexes, and improve their method. Since the problem of maximizing the size of a connected component as well as that of maximizing the number of protein-protein interactions can be proved as NP-hard[6], we use integer programming for solving the problem. However, we use an approximate reduction method because it is difficult to compactly formulate the problem as an integer program. Furthermore, we propose combinations of the improved method with maximal components and extreme sets[5].

2.1 Improved integer programming IPc

The original IP-based method by Ozawa et al. verifies an interaction between two proteins depending on the presence of interactions between domains included in the proteins. It is assumed that a domain interacts with at most one other domain. If a domain can interact with multiple domains, only one domain is selected as the partner. In the original IP-based method, such pairs of domains are selected by maximizing the number of interacting protein pairs. However, candidate proteins should be connected as much as possible because the proteins are selected as a complex by prediction methods such as MCL, and MCODE. Therefore, we consider the problem of finding the largest set of proteins that are connected to each other under the condition that a domain interacts with at most one domain.

Let $\mathcal{P}$ and $\mathcal{D}$ be a set of candidate proteins for constituting a complex, and a set of domains included in the proteins of $\mathcal{P}$, respectively, where each domain $i.k \in \mathcal{D}$ is distinguished by the protein i that the domain k belongs to. Let $\mathcal{I}_{\mathcal{P}}$, $\mathcal{I}_{\mathcal{D}}$, and $\mathcal{I}_{\mathcal{D}i,j}$ be a set of potentially interacting protein pairs, a set of potentially interacting domain pairs, and a set of potentially interacting domain pairs between proteins i and j, respectively. Then, we approximate the problem of maximizing the size of a connected component of proteins into that of maximizing the number of connected components with size three. This approximated problem can be simply transformed into the following integer program.

Maximize $\sum_{i,j,k \in \mathcal{P}} x_{i,j,k}$,

Subject to

$$\sum_{\{(i.k,j.l) \in \mathcal{I}_{\mathcal{D}} | i.k=m.n \text{ or } j.l=m.n\}} d_{i.k,j.l} \leq 1 \qquad \text{for all } m.n \in \mathcal{D}, \qquad (1)$$

$$p_{i,j} \leq \sum_{(i.k,j.l) \in \mathcal{I}_{\mathcal{D}i,j}} d_{i.k,j.l} \qquad \text{for all } (i,j) \in \mathcal{I}_{\mathcal{P}}, \qquad (2)$$

$$x_{i,j,k} \leq \frac{1}{2}(p_{i,j} + p_{j,k} + p_{i,k}) \qquad \text{for all } i,j,k \in \mathcal{P}. \qquad (3)$$

In the above inequalities, each variable of $x_{i,j,k}$, $p_{i,j}$, and $d_{i.k,j.l}$ takes 0 or 1. $x_{i,j,k} = 1$ if and only if proteins i, j, and k are connected. $p_{i,j} = 1$ if and only if proteins i and j interact with each other. $d_{i.k,j.l} = 1$ if and only if domains $i.k$ and $j.l$ interact with each other. It should be noted that for variables $x_{i,j,k}$, we

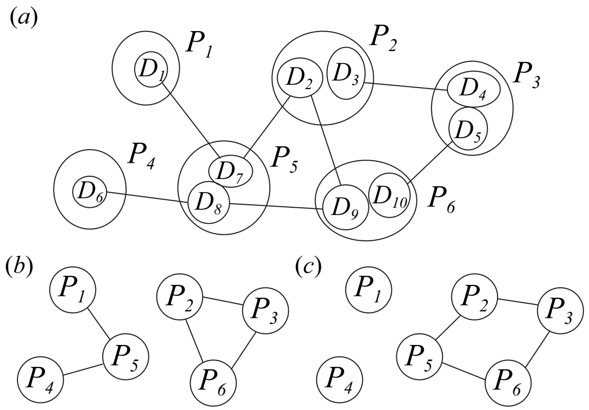


Fig. 1 Example of verification by two IP-based methods, IPo and IPc. (a) Example of a protein interaction network and domain-domain interactions. (b) The optimal solution by the IP of Ozawa et al.[4]), IPo. Each solid line denotes a protein-protein interaction. Two protein complexes are generated. (c) The optimal solution by our proposed IP, IPc. A larger protein complex is generated.

do not need to treat all combinations of three proteins, and need only proteins can be connected. Thus, the number of variables $x_{i,j,k}$ is at most $\binom{|\mathcal{I}_\mathcal{P}|}{2}$. The inequalities (1) and (2) are also included in the original IP by Ozawa et al. The meaning of each inequality is as follows:

(1) The number of domains that interact with domain $m.n$ is at most one.

(2) Proteins i and j interact if and only if there is at least one interacting domain pair $(i.k, j.l)$.

(3) Proteins i, j, and k are connected if and only if there are at least two interacting protein pairs from (i, j), (j, k), and (i, k).

It should be noted that the topology of protein-protein interaction networks is taken into account in Eq. (2). We call the original IP proposed by Ozawa et al. and our improved IP, 'IPo' and 'IPc', respectively.

Figure 1 shows an example of verification by these IP-based methods. Figure 1(a) shows an example of a protein interaction network and domain-domain interactions. There are six proteins $P_1, \ldots, P_6$ that contain one or two domains, $\{D_1\}$, $\{D_2, D_3\}$, $\{D_4, D_5\}$, $\{D_6\}$, $\{D_7, D_8\}$ and $\{D_9, D_{10}\}$, respectively. There are seven potentially interacting domain pairs $\mathcal{I}_\mathcal{D}$, and seven potentially interacting protein pairs $\mathcal{I}_\mathcal{P}$. Then, Fig. 1(b) shows the optimal solution by IPo. A candidate complex is divided into two complexes $\{P_1, P_4, P_5\}$ and $\{P_2, P_3, P_6\}$. The value of the objective function of IPo, that is, the maximum number of verified interacting protein pairs is 5. On the other hand, the optimal solution by IPc is shown by Fig. 1(c). A protein complex $\{P_2, P_3, P_5, P_6\}$ is generated. Then, the values of the objective functions of IPo and IPc are 4 and 4, respectively. Though the optimal score of IPo is better than that of IPc, we can see from this example that IPc outputs more reasonable results than IPo because a larger cluster remains in the solution by IPc.

We assume that each complex consists of at least three proteins as well as the original IP-based method. If only two proteins are obtained as a complex from the integer programs, the complex is ignored.

2.2 Maximal components and extreme sets

As mentioned before, we use maximal components and extreme sets in graph theory to enhance the verification ability of the proposed IP-based method. Maximal components and extreme sets are defined by using edge connectivity. Let $G(V, E)$ be an undirected edge-weighted graph with a set of vertices V and a set of edges E, where each edge e has a non-negative real weight $w_G(e)$. The *local edge-connectivity* $\lambda_G(u, v)$ between two nodes u and v is defined as follows[5]).

$$\lambda_G(u, v) = \min_{\{X \subset V \mid u \in X, v \in V - X\}} d_G(X),$$

where $d_G(X)$ denotes the cut size of $\{X, V - X\}$, that is, $\sum_{u \in X, v \in V - X} w_G(u, v)$. For two vertices u and v, if the local edge-connectivity $\lambda_G(u, v)$ between u and v is large, it is considered that the relationship between them is also strong.

A subset X of V is called a *maximal component* of a graph G if it satisfies the following conditions,

$$\begin{aligned} &\lambda_G(u, v) \geq l && \text{for } \forall u, v \in X, \\ &\lambda_G(u, v) < l && \text{for } \forall u \in X,\ \forall v \in V - X, \end{aligned}$$

where $l = \min_{u,v \in X} \lambda_G(u, v)$. It means that the internal vertices of a maximal component are connected more strongly with each other than with any other external vertices.

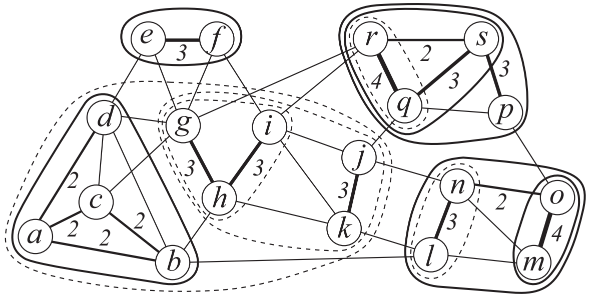


Fig. 2 Illustration of maximal components and extreme sets. Each dashed (solid) curve corresponds to a maximal component (an extreme set and a maximal component).

Furthermore, a nonempty proper subset X of V is called an *extreme set* of a graph G if it satisfies the following condition,

$$d_G(X) < d_G(Y) \qquad \text{for } \forall\, Y \subset X.$$

It is known that every extreme set is a maximal component, and there exists an $O(mn + n^2 \log n)$ time algorithm for a graph with n vertices and m edges that computes maximal components and extreme sets[5].

Figure 2 illustrates maximal components and extreme sets. Each dashed (solid) curve corresponds to a maximal component (an extreme set and a maximal component).

For verifying protein complexes, we let $w_G(u, v) = 1$ for each protein-protein interaction, and calculate maximal components and extreme sets.

3. Computational experiments

For evaluating our proposed IP-based method and the combination methods with maximal components and extreme sets, we performed several computational experiments, and compared with the original IP-based method that is considered to be the best existing method for verifying protein complexes[7].

3.1 Data and implementation

We used WI-PHI[8] and BioGRID[9] as data of protein-protein interactions, which includes 5,907 and 4,603 yeast proteins identified by UniProt database (Release 2011_03)[10], and 49,847 and 30,853 interacting protein pairs, respectively. For each protein, we extracted Pfam domains[11] included in the protein using the UniProt database. We used iPfam database (version 21.0)[12] as data of potential domain-domain interactions, which includes 2,837 Pfam domains and 4,030 interacting Pfam domain pairs. For obtaining candidate protein complexes, we applied MCL[1] with several parameters of 'inflation' and MCODE[2] with several parameters of 'node score cutoff', respectively, to both of the WI-PHI and BioGRID protein-protein interaction data.

To evaluate the performances of verification methods, we used a known comprehensive catalog of yeast protein complexes CYC2008[13], which includes 408 curated complexes. The precision and the recall of each method, also used in[4),14], for a set of verified protein complexes $\mathcal{C}$ and a set of known protein complexes $\mathcal{K}$ were calculated as follows:

$$\mathit{precision} = \frac{|\{c \in \mathcal{C} | \exists k \in \mathcal{K}\ \mathit{concordance}(c, k) \geq 0.5\}|}{|\mathcal{C}|},$$

$$\mathit{recall} = \frac{|\{k \in \mathcal{K} | \exists c \in \mathcal{C}\ \mathit{concordance}(c, k) \geq 0.5\}|}{|\mathcal{K}|},$$

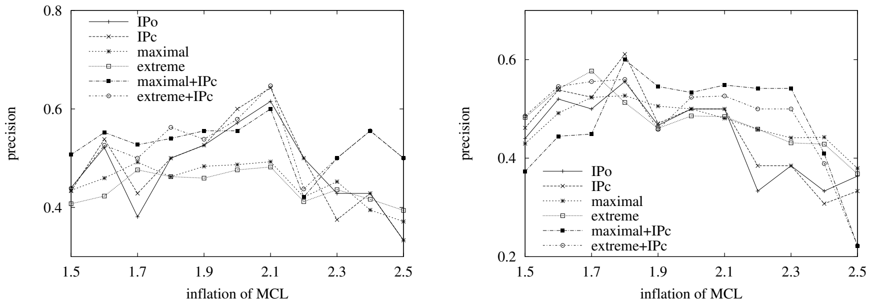
where $\mathit{concordance}(c, k)$ denotes the concordance rate between sets of proteins c and k, which is defined as $\frac{|c \cap k|}{\sqrt{|c| \cdot |k|}}$. From the definition, multiple predicted complexes may correspond to the same known complex. The *accuracy* is defined as the geometrical mean of the precision and the recall, that is, $\mathit{accuracy} = \sqrt{\mathit{precision} \cdot \mathit{recall}}$.

We used IBM ILOG CPLEX (version 12.1) to solve the integer programs. All of the computational experiments were conducted on a PC with a Xeon CPU 3.33 GHz and 10 GB memory under the linux OS (version 2.6.16).

3.2 Results

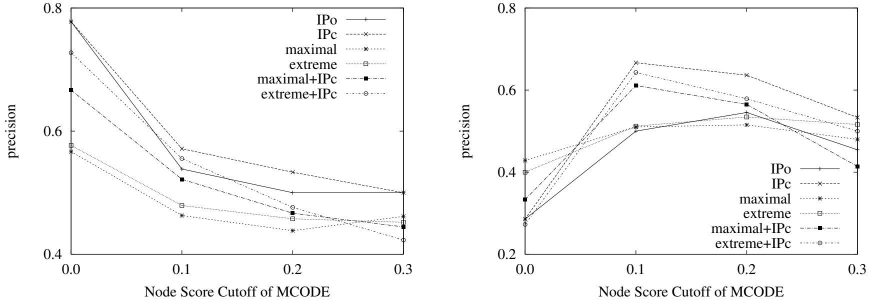
For comparing verification performances of the original IP-based method and our proposed methods, we performed computational experiments using results by MCL[1] as candidate protein complexes because MCL was reported to outperform other prediction methods for protein complexes[7] and has often been used for that purpose. In addition to results by MCL, we used those by MCODE[2].

Figure 3 shows the results of the precision by the original IP-based method (IPo), our improved IP-based method (IPc), maximal components, extreme sets, and the combination methods of IPc with maximal components (maximal+IPc)



(a) WI-PHI (b) BioGRID

Fig. 3 Results of the precision by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidates obtained from (a) WI-PHI (b) BioGRID by MCL with varying the inflation parameter from 1.5 to 2.5. 'maximal+IPc' and 'extreme+IPc' denote that IPc is applied after the calculation of maximal components and extreme sets, respectively. Each method was applied to candidate protein complexes.



(a) WI-PHI (b) BioGRID

Fig. 4 Results of the precision by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidates obtained from (a) WI-PHI (b) BioGRID by MCODE with varying the node score cutoff parameter from 0.0 to 0.3.

and extreme sets (extreme+IPc) for candidate protein complexes obtained from the WI-PHI and BioGRID protein-protein interaction data, by MCL with varying the inflation parameter from 1.5 to 2.5. In the combination methods, IPc is applied after the calculation of maximal components and extreme sets, respectively. Each method was applied to candidate protein complexes obtained by MCL. For the WI-PHI data, the precision of IPc was better than that of IPo except for inflation=2.3, and in almost all methods, the precision was the best for inflation=2.1. For the BioGRID data, the precision of IPc was better than or comparable to that of IPo, and among all methods, the precision of IPc for inflation=1.8 was the best.

Table 1 Results of the precision, the recall, and the accuracy by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidate protein complexes obtained from the WI-PHI data by MCL with inflation 2.1.

method	precision	recall	accuracy
IPo	0.6154	0.0319	0.1400
IPc	0.6429	0.0343	0.1485
maximal	0.4928	0.0882	0.2085
extreme	0.4821	0.0858	0.2034
maximal+IPc	0.6000	0.0368	0.1485
extreme+IPc	0.6471	0.0343	0.1490

Figure 4 shows the results of the precision by the original IP-based method (IPo), our improved method (IPc), maximal components, extreme sets, maximal+IPc, and extreme+IPc for candidate protein complexes obtained from the WI-PHI and BioGRID protein-protein interaction data, by MCODE with varying the inflation parameter from 0.0 to 0.3. For both protein-protein interaction data, the precision of IPc was better than or comparable to that of IPo, and among all methods, the precision of IPc was the best except for the BioGRID data with cutoff=0.0.

Table 1 shows the results of the precision, recall, and accuracy by the original IP-based method (IPo), our improved IP-based method (IPc), maximal components, extreme sets, maximal+IPc, and extreme+IPc for candidate protein complexes obtained from the WI-PHI protein-protein interaction data by MCL with inflation 2.1. The recalls and accuracies of our methods were better than those of IPo, and the precision of extreme+IPc for inflation=2.1 was the best. Though the recalls of IPo and IPc were low, Ozawa et al. also reported that the recall of their method that used domain-domain interaction data of iPfam database (version 21.0) and MCL was low. However, it is important to enhance the precision in order to avoid generation of too many erroneous predictions. These results suggest that our proposed IP-based methods, especially extreme+IPc, considerably outperform the original IP-based method both in recall and precision. The maximum execution times of IPo and IPc for a candidate protein complex by MCL with inflation 2.1 were about 0.04 and 0.84 seconds, respectively, where both methods took less than 0.01 second per complex in most cases. Though IPc took longer CPU time than IPo did, it is still acceptable. Since it is more important to achieve a better precision than to have shorter CPU time, we can conclude that IPc is better than IPo.

4. Conclusions

We have addressed the problem of verification of candidate protein complexes, and proposed an improved integer programming (IP)-based method by introducing the size of a connected component. In addition to the IP-based method, we proposed the combination methods with maximal components and extreme sets, which partition vertices based on the connectivity between two vertices graph-theoretically. The results of several computational experiments suggest that our proposed methods outperform the existing IP-based method.

As a future work, it remains to find a compact formulation of the problem of maximizing the size of a connected component because we solved this problem approximately. Other future work includes developing a method with a better recall while keeping the precision, and improving the efficiency factor to a higher range.

Acknowledgements

This work was partially supported by Grants-in-Aid #22240009 and #21700323 from MEXT, Japan.

References

1) Enright, A.J., Dongen, S.V. and Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Research*, Vol.30, No.7, pp.1575–1584 (2002).

2) Bader, G.D. and Hogue, C. W.V.: An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*, Vol.4, p.2 (2003).

3) Maruyama, O. and Chihara, A.: NWE: Node-weighted expansion for protein complex prediction using random walk distances, *Proteome Science*, Vol.9, p.S14 (2011).

4) Ozawa, Y., Saito, R., Fujimori, S., Kashima, H., Ishizaka, M., Yanagawa, H., Miyamoto-Sato, E. and Tomita, M.: Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions, *BMC Bioinformatics*, Vol.11, p.350 (2010).

5) Nagamochi, H.: Graph algorithms for network connectivity problems, *Journal of the Operating Research Society of Japan*, Vol.47, pp.199–223 (2004).

6) Zhao, Y., Hayashida, M., Nacher, J., Nagamochi, H. and Akutsu, T.: Protein complex prediction via improved verification methods using constrained domain-domain matching, *Proc. The tenth Asia-Pacific Bioinformatics Conference*, pp.394–406 (2012).

7) Brohée, S. and van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks, *BMC Bioinformatics*, Vol.7, p.488 (2006).

8) Kiemer, L., Costa, S., Ueffing, M. and Cesareni, G.: WI-PHI: A weighted yeast interactome enriched for direct physical interactions, *Proteomics*, Vol.7, pp.932–943 (2007).

9) Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M.: BioGRID: a general repository for interaction datasets, *Nucleic Acids Research*, Vol.34, pp.D535–D539 (2006).

10) The UniProt Consortium: Ongoing and future developments at the Universal Protein Resource, *Nucleic Acids Research*, Vol.39, pp.D214–D219 (2011).

11) Bateman, A., Coin, L., Durbin, R., Finn, R.D., Holich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L.L., Studholme, D.J., Yeats, C. and Eddy, S.R.: The Pfam protein families database, *Nucleic Acids Research*, Vol.32, pp.D138–D141 (2004).

12) Finn, R.D., Marshall, M. and Bateman, A.: iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions, *Bioinformatics*, Vol.21, No.3, pp.410–412 (2005).

13) Pu, S., Wong, J., Turner, B., Cho, E. and Wodak, S.J.: Up-to-date catalogues of yeast protein complexes, *Nucleic Acids Research*, Vol.37, pp.825–831 (2009).

14) Chua, H.N., Ning, K., Sung, W.K., Leong, H.W. and Wong, L.: Using indirect protein-protein interactions for protein complex prediction, *Journal of Bioinformatics and Computational Biology*, Vol.6, No.3, pp.435–466 (2008).