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

Yang Zhao[1], Morihiro Hayashida[1], Jose C. Nacher[2], Hiroshi Nagamochi[3], Tatsuya Akutsu[1]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
[2] Department of Complex and Intelligent Systems, Future University-Hakodate, 116-Kamedankano, Hakodate, Hokkaido 041-8655, Japan
[3] Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, 606-8501, Japan

Abstract

Identification of protein complexes within protein-protein interaction networks is one of the important objectives in functional genomics. Ozawa et al. proposed a verification method of protein complexes by introducing a structural constraint.

In this paper, we propose an improved integer programming-based method based on the idea that a candidate complex should not be divided into many small complexes, and combination methods with maximal components and extreme sets. The results of computational experiments suggest that our methods outperform the method by Ozawa et al. We prove that the verification problems are NP-hard, which justifies the use of integer programming.

1 Introduction

With the rapid development of cell biology and systems biology, enormous amounts of protein-protein interaction (PPI) data are available for researchers to understand important principles of cellular organization and biological function. 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 (Enright et al., 2002). Another graph theoretic clustering algorithm, MCODE, detects densely connected regions as molecular complexes in large PPI networks based on connectivity data (Bader and Hogue, 2003). SPC (superparamagnetic clustering), MC, and PCP also make use of topological properties that proteins in each complex are densely connected (Spirin and Mirny, 2003; Chua et al., 2008). RNSC algorithm efficiently clusters PPI networks by using a cost function (King et al., 2004). Qi et al. (2008) modeled each complex subgraph by a probabilistic Bayesian network using topological patterns and biological properties. Maruyama and Chihara (2011) proposed NWE (Node-Weighted Expansion of clusters of proteins) by introducing a random walk with restarts with a cluster of proteins. Wu et al. (2011) used tandem affinity purification (TAP) data that is obtained from an experimental method for detecting multi-protein interactions.

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. As Singh et al. (2010) proposed a structure-based method for predicting protein-protein interactions, it is important to consider protein structures in addition to protein-protein interaction networks. Proteins composed of multiple structural domains, which are in most cases autonomous folding units, may not adequately be treated by these methods and result in an incorrect group of proteins in extracted complexes. Another reason that causes the incorrect proteins perhaps comes from no consideration of topology of PPIs in the networks. 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. Chen et al. (2005) proposed an algorithm called AlternativePathFinder which considers a complex candidate to be valid if it is included in a closed loop. A clustering-based method, CMC, maximizes cliques from the weighted PPI networks to detect protein complexes. In addition, Habibi et al. (2010) proposed an algorithm that seeks for protein complexes from the PPI networks based on finding maximal k-connected subgraphs. 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 Ozawa et al. (2010) 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 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 paper, 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 (Nagamochi, 2004). 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. Furthermore, we prove that the problem of maximizing the number of protein-protein interactions, considered by Ozawa et al., is NP-hard, and also prove that our problem of maximizing the size of a connected component given by verified protein-protein interactions is NP-hard. 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. (2010) proposed an integer programming (IP)-based method for verifying candidate complexes by maximizing the number of protein-protein interactions. In this paper, 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 will be proved as NP-hard in the next section, 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 (Nagamochi, 2004).

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 MCODE, MCL, and RNSC. 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

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

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

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

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 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 $\begin{pmatrix} |\mathcal{I}_{\mathcal{P}}| \\ 2 \end{pmatrix}$. 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 $\mathcal{I}_{\mathcal{D}} = \{(D_1, D_7), (D_2, D_7), (D_2, D_9), (D_3, D_4), (D_5, D_{10}), (D_6, D_8), (D_8, D_9)\}$, and seven potentially interacting protein pairs $\mathcal{I}_{\mathcal{P}} = \{(P_1, P_5), (P_2, P_5), (P_2, P_6), (P_2, P_3), (P_3, P_6), (P_4, P_5), (P_5, P_6)\}$. Then, Figure 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 Figure 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 (Nagamochi, 2004).

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

where $d_G(X)$ denotes the cut size of $\{X, V - X\}$, that is, $\sum_{u \in X, v \in V - X} w_G(u, v)$.

Figure 2 illustrates a cut $\{X, V - X\}$ that determines the local edge-connectivity $\lambda_G(g, h)$ between vertices g and h, where the graph G contains the set of 19 vertices $V = \{a, b, \ldots, s\}$ and the set of edges E, each number in this figure denotes the weight w_G of the edge, and the edges without a number are weighted by 1. For the set $X = \{a, b, e, f, g, l, m, k, q\}$, $\sum_{u \in X, v \in V - X} w_G(u, v) = 6$. Then, X gives one of the minimum (g, h)-cuts, and $\lambda_G(g, h) = 6$ (All the minimum (g, h)-cuts of G are shown in Figure 3). 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.

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 (Nagamochi, 2004).

Figure 4 shows the maximal components and the extreme sets for the graph of Figure 2. 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 Hardness results

In this section, we prove that the problem formulated as an integer program by Ozawa et al. is NP-hard. It maximizes the number of protein-protein interactions under the condition that a domain interacts with at most one other domain. In addition, we prove that the problem of maximizing the size of a connected component in a protein complex is NP-hard.

Problem 1 *Protein Complex Verification Problem (PCVP)*

Given a set of proteins $\mathcal{P}$ as a candidate complex, a set of domains $\mathcal{D}$ included in the proteins, a set of potential protein-protein interactions $\mathcal{I}_\mathcal{P}$, and a set of potential domain-domain interactions $\mathcal{I}_\mathcal{D}$, maximize the number of verified protein-protein interactions, where a domain can interact with at most one domain, and a protein-protein interaction is said to be verified if the corresponding pair of proteins contains at least one interacting domain pair.

Then, we have the following theorem.

Theorem 3.1 *The protein complex verification problem (PCVP) is NP-hard.*

Proof. We show a polynomial-time reduction from 3-dimensional matching.

Problem 2 *3-dimensional matching (3DM)*

Given $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ are finite mutually disjoint sets with $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}| = n$, find $\mathcal{M} \subseteq \mathcal{S}$ such that $|\mathcal{M}| = n$ and $\{X, Y, Z \mid (X, Y, Z) \in \mathcal{M}\} = \mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}$.

Suppose that an instance of 3DM is given as $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$, $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ and $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_n\}$ are three mutually disjoint sets with n elements, respectively.

We transform the instance of 3DM to an instance of PCVP as follows (see Figure 5).

$$
\begin{aligned}
\mathcal{P} &= \{\mathcal{P}_{\mathcal{X}}\} \cup \bigcup_{j=1}^{n} \{\mathcal{P}_{Y_j}\} \cup \bigcup_{k=1}^{n} \bigcup_{l=1}^{n-1} \{\mathcal{P}_{Z_k}^{(l)}\}, \\
\mathcal{D} &= \bigcup_{i=1}^{n} \{X_i\} \cup \bigcup_{j=1}^{n} \bigcup_{k=1}^{n} \{D_{Y_j}^{Z_k}\} \cup \bigcup_{k=1}^{n} \bigcup_{l=1}^{n-1} \{Z_k^{(l)}\}, \\
\mathcal{I}_{\mathcal{D}} &= \{(X_i, D_{Y_j}^{Z_k}) \mid (X_i, Y_j, Z_k) \in \mathcal{S}\} \cup \bigcup_{j=1}^{n} \bigcup_{k=1}^{n} \bigcup_{l=1}^{n-1} \{(D_{Y_j}^{Z_k}, Z_k^{(l)})\}
\end{aligned}
$$

where $\mathcal{P}_{\mathcal{X}}$ contains domains $X_1, X_2, \ldots$, and X_n ($\in \mathcal{X}$), $\mathcal{P}_{Y_j}$ ($Y_j \in \mathcal{Y}$) contains domains $D_{Y_j}^{Z_1}, D_{Y_j}^{Z_2}, \ldots$, and $D_{Y_j}^{Z_n}$, $\mathcal{P}_{Z_k}^{(l)}$ ($Z_k \in \mathcal{Z}, l \in \{1, \ldots, n-1\}$) contains a domain $Z_k^{(l)}$. If $(\alpha, \beta) \in \mathcal{I}_{\mathcal{D}}$ and α and β are respectively included in proteins γ and δ, then we let $(\gamma, \delta) \in \mathcal{I}_{\mathcal{P}}$.

Then, we can see in the following way that there exists a 3-dimensional matching if and only if the maximum number of verified protein-protein interactions is exactly $(n + n \cdot (n-1)) = n^2$, and $(X_i, Y_j, Z_k) \in \mathcal{M}$ holds if and only if $(X_i, D_{Y_j}^{Z_k})$ is selected. For each protein pair $(\mathcal{P}_{\mathcal{X}}, \mathcal{P}_{Y_j})$, exactly one domain pair is selected because protein $\mathcal{P}_{\mathcal{X}}$ contains n domains, and a domain interacts with at most one domain. If more than one domain pair are selected for a protein pair, the number of proteins that interact with $\mathcal{P}_{\mathcal{X}}$ is less than n, that is, the maximum cannot achieve n^2. Furthermore, among n domains $D_{Y_j}^{Z_k}$ for each Z_k ($\in \mathcal{Z}$), exactly one domain $D_{Y_j}^{Z_k}$ is selected for the interaction with $\mathcal{P}_{\mathcal{X}}$, and the other domains interact with $(n-1)$ domains of $Z_k^{(l)}$ ($l = 1, \ldots, n-1$), respectively. If more than one domain are selected for the interaction with $\mathcal{P}_{\mathcal{X}}$, there exists the protein $\mathcal{P}_{Z_k}^{(l)}$ for some $l \in \{1, \ldots, n-1\}$ that is not able to interact with any protein.

Furthermore, we can derive a solution for an instance of 3DM in polynomial time from a solution for the transformed instance of PCVP. An instance of 3DM can also be transformed in polynomial time to the corresponding instance of PCVP. Therefore, the protein complex verification problem (PCVP) is NP-hard. □

From Theorem 3.1, we can also prove that the problem of maximizing the size of a connected component consisting of proteins is NP-hard.

Problem 3 *Connected Protein Complex Verification Problem (CPCVP)*

Given a set of proteins $\mathcal{P}$ as a candidate complex, a set of domains $\mathcal{D}$ included in the proteins, a set of potential protein-protein interactions $\mathcal{I}_{\mathcal{P}}$, and a set of potential domain-domain interactions $\mathcal{I}_{\mathcal{D}}$, maximize the size of a connected component given by verified protein-protein interactions, where a domain can interact with at most one domain.

Theorem 3.2 *The connected protein complex verification problem (CPCVP) is NP-hard.*

Proof. We use the same reduction as in the proof of Theorem 1. Then, we can see that there exists a 3-dimensional matching if and only if all the proteins in $\mathcal{P}$ are connected, and $(X_i, Y_j, Z_k) \in \mathcal{M}$ holds if and only if $(X_i, D_{Y_j}^{Z_k})$ is selected. Furthermore, the derivation of a solution of 3DM and the transformation of an instance of 3DM can be done in polynomial time. Therefore, the connected protein complex verification problem (CPCVP) is NP-hard. □

These results justify the use of integer programming both in our method and in the method by Ozawa et al. (2010). It is to be noted that maximizing the number of domain-domain interactions (instead of protein-protein interactions) can be done in polynomial time by using maximum matching.

4 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 (Brohée and van Helden, 2006).

4.1 Data and implementation

We used WI-PHI (Kiemer et al., 2007) and BioGRID (Stark et al., 2006) as data of protein-protein interactions, which includes 5,907 and 4,603 yeast proteins identified by UniProt database (Release 2011_03) (The UniProt Consortium, 2011), and 49,847 and 30,853 interacting protein pairs, respectively. For each protein, we extracted Pfam domains (Bateman et al., 2004) included in the protein using the UniProt database. We used iPfam database (version 21.0) (Finn et al., 2005) 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 (Enright et al., 2002) with several parameters of 'inflation' and MCODE (Bader and Hogue, 2003) 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 (Pu et al., 2009), which includes 408 curated complexes. The precision and the recall of each method,

also used in (Ozawa et al., 2010; Chua et al., 2008), for a set of verified protein complexes $\mathcal{C}$ and a set of known protein complexes $\mathcal{K}$ were calculated as follows:

$$\begin{aligned} precision &= \frac{|\{c \in \mathcal{C} | \exists k \in \mathcal{K}\ concordance(c,k) \geq 0.5\}|}{|\mathcal{C}|}, \\ recall &= \frac{|\{k \in \mathcal{K} | \exists c \in \mathcal{C}\ concordance(c,k) \geq 0.5\}|}{|\mathcal{K}|}, \end{aligned}$$

where $concordance(c,k)$ 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, $accuracy = \sqrt{precision \cdot recall}$.

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).

4.2 Results

For comparing verification performances of the original IP-based method and our proposed methods, we performed computational experiments using results by MCL (Enright et al., 2002) as candidate protein complexes because MCL was reported to outperform other prediction methods for protein complexes (Brohée and van Helden, 2006) and has often been used for that purpose. In addition to results by MCL, we used those by MCODE (Bader and Hogue, 2003).

Figures 6 and 7 show 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) and extreme sets (extreme+IPc) for candidate protein complexes obtained from the WI-PHI and BioGRID protein-protein interaction data, respectively, 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.

Figures 8 and 9 show 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, respectively, 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: Results of the precision, the recall, and the 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 data by MCL with inflation 1.9.

method	precision	recall	accuracy
IPo	0.526316	0.036765	0.139104
IPc	0.526316	0.036765	0.139104
maximal	0.483516	0.098039	0.217724
extreme	0.459459	0.090686	0.204124
maximal+IPc	0.555556	0.039216	0.147602
extreme+IPc	0.538462	0.036765	0.140700

Table 2: Results of the precision, the recall, and the 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 data by MCL with inflation 2.0.

method	precision	recall	accuracy
IPo	0.571429	0.031863	0.134934
IPc	0.600000	0.034314	0.143486
maximal	0.486842	0.093137	0.212939
extreme	0.476190	0.085784	0.202113
maximal+IPc	0.555556	0.036765	0.142915
extreme+IPc	0.578947	0.034314	0.140946

Table 3: Results of the precision, the recall, and the 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 data by MCL with inflation 2.1.

method	precision	recall	accuracy
IPo	0.615385	0.031863	0.140028
IPc	0.642857	0.034314	0.148522
maximal	0.492754	0.088235	0.208514
extreme	0.482143	0.085784	0.203372
maximal+IPc	0.600000	0.036765	0.148522
extreme+IPc	0.647059	0.034314	0.149007

Tables 1,2, and 3 show 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 1.9, 2.0, and 2.1, respectively. For inflation=1.9, the precision of IPo was equal to that of IPc and less than that of maximal+IPc

and extreme+IPc, the precision of maximal+IPc was the best, and IPc output the same solution as IPo. For inflation=2.0, 2.1, the precisions of IPc and extreme+IPc were better than that of IPo. 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.

5 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.

Furthermore, we proved that the problem of maximizing the number of protein-protein interactions under the condition that a domain interacts with at most one other domain, considered by Ozawa et al., is NP-hard, and also proved that the problem of maximizing the size of a connected component given by verified protein-protein interactions under the same condition is NP-hard. These results justify the use of integer programming both in our method and in the method by Ozawa et al. (2010).

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.

References

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

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. (2004) 'The Pfam protein families database', *Nucleic Acids Research*, Vol. 32, pp. D138–D141.

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

Chen, J., Hsu, W., Lee, M. L., and Ng, S.-K. (2005) 'Discovering reliable protein interactions from high-throughput experimental data using network topology', *Artificial Intelligence in Medicine*, pp. 37–47.

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

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

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

Habibi, M., Eslahchi, C., and Wong, L. (2010) 'Protein complex prediction based on k-connected subgraphs in protein interaction network', *BMC Systems Biology*, Vol. 4, pp. 129.

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

King, A. D., Pržulj, N., and Jurisica, I. (2004) 'Protein complex prediction via cost-based clustering', *Bioinformatics*, Vol. 20, No. 17, pp. 3013–3020.

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

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

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

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

Qi, Y., Balem, F., Faloutsos, C., Klein-Seetharaman, J., and Bar-Joseph, Z. (2008) 'Protein complexes identification by supervised graph local clustering', *Bioinformatics*, Vol. 24, pp. i250–i258.

Singh, R., Park, D., Xu, J., Hosur, R., and Berger, B. (2010) 'Struct2Net: a web service to predict protein-protein interactions using a structure-based approach', *Nucleic Acids Research*, Vol. 2010, pp. 1–8.

Spirin, V. and Mirny, L. A. (2003) 'Protein complexes and functional modules in molecular networks', *Proc. Natl. Acad. Sci. USA*, Vol. 100, pp. 12123–12128.

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006) 'BioGRID: a general repository for interaction datasets', *Nucleic Acids Research*, Vol. 34, pp. D535–D539.

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

Wu, M., Li, X.-L., Kwoh, C.-K., Ng, S.-K., and Wong, L. (2011) 'Discovery of protein complexes with core-attachment structures from tandem affinity purification (TAP) data', *Journal of Computational Biology*, Vol. 18, pp. 1–16.

Figure 1: Example of verification by two IP-based methods, IPo and IPc. (a) Example of a protein interaction network and domain-domain interactions. There are six proteins that contain one or two domains, seven potentially interacting domain pairs, and seven potentially interacting protein pairs, where these protein-protein interactions are not shown. (b) The optimal solution by the IP of Ozawa et al. (2010), 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.

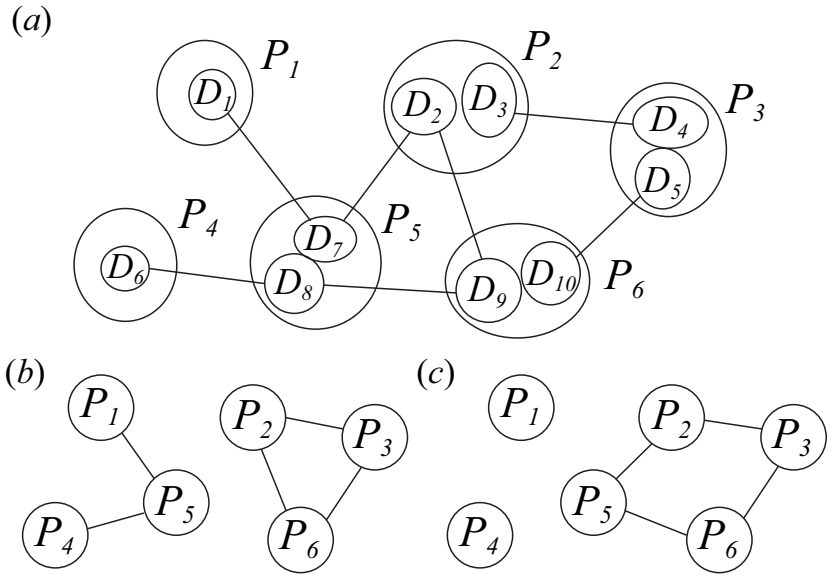

Figure 2: Illustration of a cut $\{X, V - X\}$ that determines the local edge-connectivity $\lambda_G(g, h)$ between vertices g and h, where the graph G contains the set of 19 vertices $V = \{a, b, \ldots, s\}$ and the set of edges E, each number in this figure denotes the weight w_G of the edge, and the edges without a number are weighted by 1. For the set $X = \{a, b, e, f, g, l, m, k, q\}$, $\sum_{u \in X, v \in V - X} w_G(u, v) = 6$. Then, X gives one of the minimum (g, h)-cuts, and $\lambda_G(g, h) = 6$.

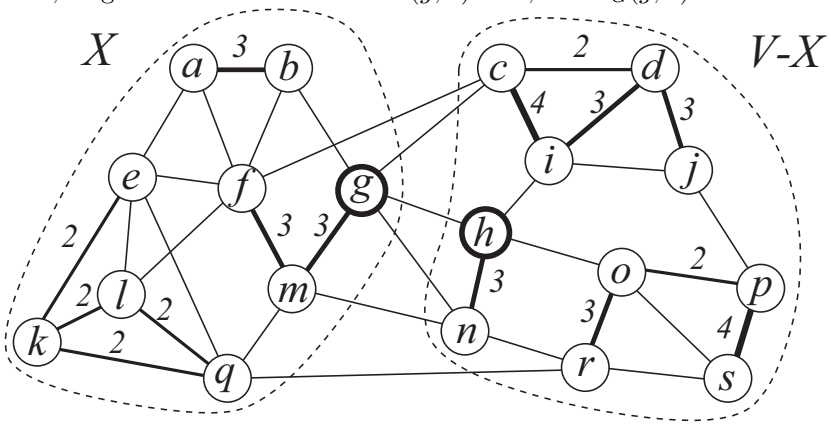

Figure 3: Minimum (g, h)-cuts of the graph G in Figure 2. There are four cuts given by the sets including the vertex h, $X_1 = \{h\}$, $X_2 = \{h, n\}$, $X_3 = \{h, n, o, p, r, s\}$, and $X_4 = \{c, d, h, i, j, n, o, p, r, s\}$.

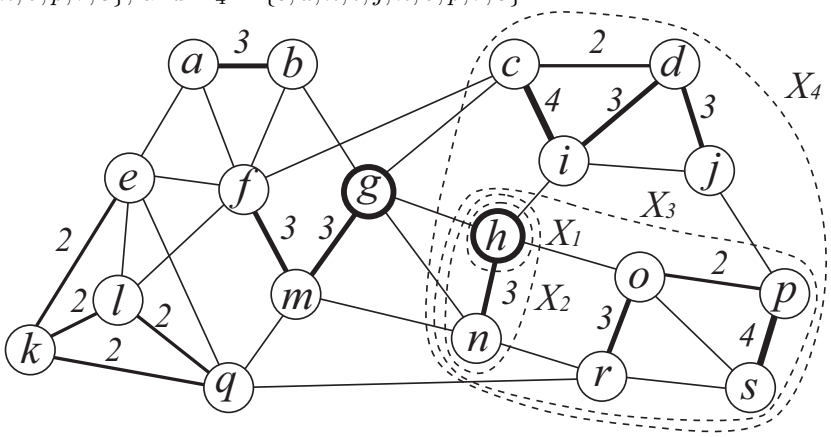

Figure 4: Illustration of maximal components and extreme sets. The maximal components and the extreme sets of the graph G in Figure 2. Each dashed (solid) curve corresponds to a maximal component (an extreme set and a maximal component).

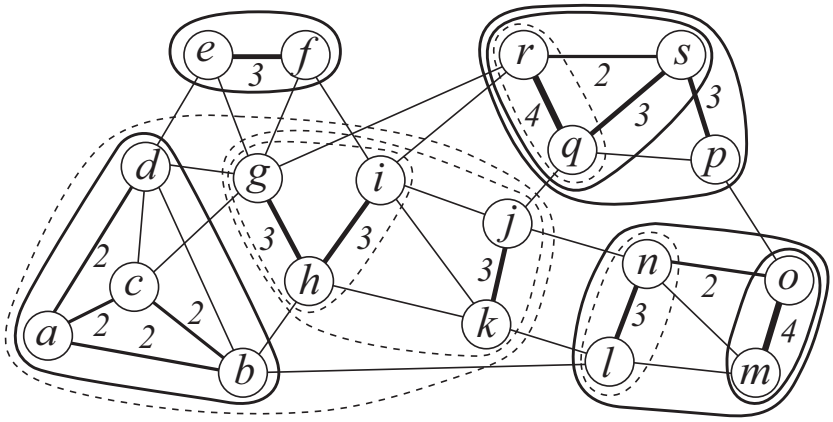

Figure 5: Illustration of the reduction from 3-dimensional matching (3DM) to the protein complex verification problem (PCVP) in the case of $n = 26$, where $\mathcal{X} = \{A, B, \ldots, Z\}$, $\mathcal{Y} = \{1, 2, \ldots, 26\}$, and $\mathcal{Z} = \{a, b, \ldots, z\}$. The large and small circles denote proteins and domains, respectively. The solid and dotted lines denote potential domain-domain interactions, and the solid lines are selected.

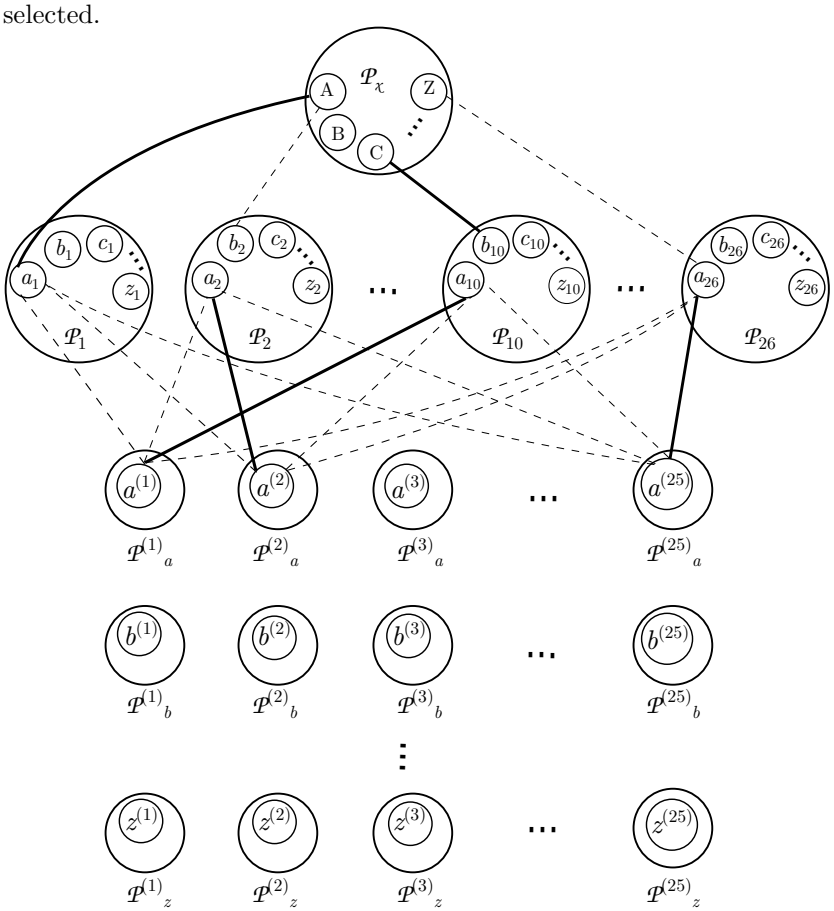

Figure 6: Results of the precision by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidates obtained from WI-PHI by MCL with varying the inflation parameter from 1.5 to 2.5. 'maximal+IPc' and 'extreme+IPc' denote that IPc is applied after the calculation of maximal components and extreme sets, respectively. Each method was applied to candidate protein complexes.

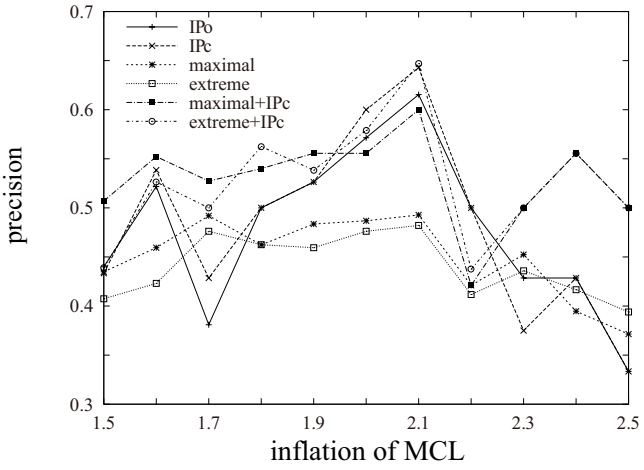

Figure 7: Results of the precision by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidates obtained from BioGRID by MCL with varying the inflation parameter from 1.5 to 2.5.

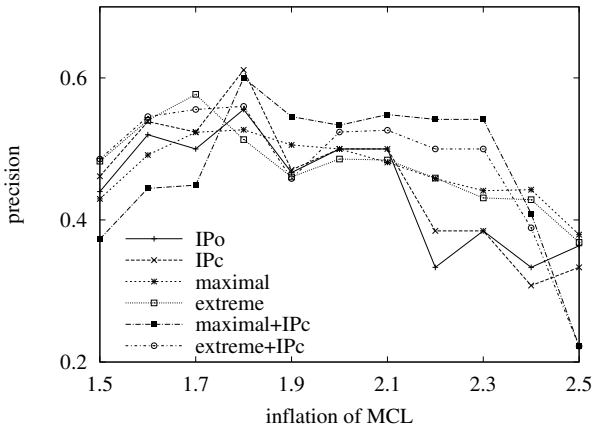

Figure 8: Results of the precision by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidates obtained from WI-PHI by MCODE with varying the node score cutoff parameter from 0.0 to 0.3.

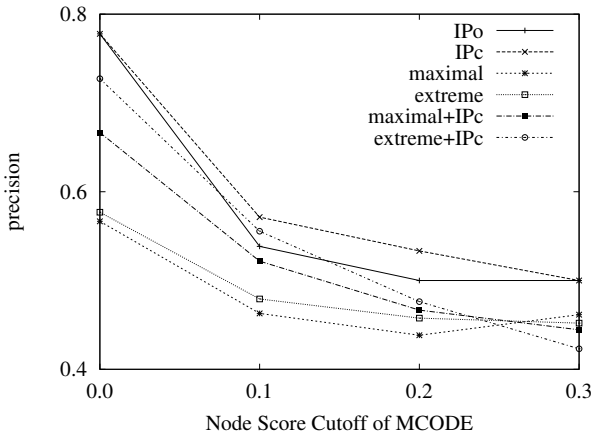

Figure 9: Results of the precision by IPo, IPc, maximal, extreme, maximal+IPc, and extreme+IPc for candidates obtained from BioGRID by MCODE with varying the node score cutoff parameter from 0.0 to 0.3.