Proc. ASONAM 2015, pp. 1068-1075, Paris, 25-28 Aug 2015

Regularizing predicted complexes by mutually exclusive protein-protein interactions

Osamu Maruyama
Institute of Mathematics for Industry
Kyushu University
Fukuoka 812-8581, Japan
Email: om@imi.kyushu-u.ac.jp

Limsoon Wong
School of Computing
National University of Singapore
Singapore 117417
Email: wongls@comp.nus.edu.sg

***Abstract*—Protein complexes are key entities in the cell responsible for various cellular mechanisms and biological processes. We propose here a method for predicting protein complexes from a protein-protein interaction (PPI) network, using information on mutually exclusive PPIs. If two interactions are mutually exclusive, they are not allowed to exist simultaneously in the same predicted complex. We introduce a new regularization term which checks whether predicted complexes are connected by mutually exclusive PPIs. This regularization term is added into the scoring function of our earlier protein complex prediction tool, PPSampler2. We show that PPSampler2 with mutually exclusive PPIs outperforms the original one. Furthermore, the performance is superior to well-known representative conventional protein complex prediction methods. Thus, it is is effective to use mutual exclusiveness of PPIs in protein complex prediction.**

I. Introduction

In the cell, many proteins form multi-protein structures, called protein complexes, by non-covalent protein-protein interactions (PPIs). These proteins are allowed to exert their inherent functions in the form of these protein complexes. Thus, protein complexes are necessary basic entities in the cell, with a role in various cellular mechanisms and biological processes.

Various methods for protein complex prediction can be found in survey papers [1], [2], [3]. These methods take PPIs as input because the components of complexes are connected via PPIs. These methods are based on the observation that densely connected subnetworks of a whole PPI network often overlap with known complexes. A commonly recognized problem of these conventional methods is that a given input PPI network is *static*, *i.e.*, the dynamics of proteins and their interactions are not represented in conventional PPI networks, although protein complexes in the cell are dynamical entities [4].

If two proteins are not able to interact with the same protein simultaneously due to physical constraints, those interactions are said to be mutually exclusive. One promising step toward the dynamical protein complex prediction is to exploit the mutual exclusiveness of PPIs [4], [5], [6].

In particular, Jung *et al.* [5] have determined 458 pairs of interactions which are mutually exclusive to each other based on interaction interfaces. Their idea of predicting protein complexes from mutually exclusive interactions as well as a PPI network is to generate PPI subnetworks in which any pair of interactions are not mutually exclusive. The resulting networks are called simultaneous protein interaction networks (SPINs). A disadvantage of this approach is that the possible number of SPINs is 2^n for n pairs of mutually exclusive interactions. Thus, the computational cost of directly applying existing clustering algorithms to every SPIN is prohibitive. The procedure the authors took actually is explained as follows. At first, the whole original PPI network is separated into relatively large PPI subnetworks by a conventional clustering algorithm. From each of the subnetworks, SPINs are generated. Then, the conventional clustering algorithm is again applied to every SPIN generated in the previous step, whose outputs are predicted complexes. They have shown the effectiveness of using mutually exclusive PPIs by applying SPINs to MCODE [7] and LCMA [8], which are conventional clustering algorithms.

In this work, instead of generating many SPINs, we propose another approach to exploit the mutual exclusiveness of interactions. We then formulate a regularization term which determines whether a connected cluster of proteins is connected under the assumption that PPIs that are mutually exclusive to each other are not allowed to exist simultaneously within the cluster. This regularization term is added to the scoring function of our protein complex prediction method, PPSampler2 [9]. An advantage of the method is that we can easily extend the scoring function by adding new terms to the existing scoring function, and optimize the resulting scoring function by the same sampling algorithm. Furthermore, the original version of PPSampler2 checks whether a predicted cluster is connected via PPIs in the optimization process. By introducing the mutual exclusiveness information, the internal PPI structure within a cluster is evaluated more precisely.

We call PPSampler2 with the regularization term based on mutually exclusive interactions PPSampler2-PIME, where PIME stands for "pair of interactions which are mutually exclusive." We show that PPSampler2-PIME outperforms the original version of PPSampler2. Thus we empirically conclude that information of mutually exclusive interactions is effective for predicting complexes more precisely.

II. METHOD

A set of PPIs together with their reliabilities can be represented by an edge-weighted undirected graph, $G = (V, E, w)$, where V is a set of proteins, E is a set of interactions between two proteins, and $w(e)$ represents the weight of an interaction $e \in E$. A partition C of V is defined as

$$C = \left\{ d_1, \ldots, d_n \subseteq V \;\middle|\; \begin{array}{l} 1 \leq i \leq n, d_i \neq \emptyset; \\ \cup_{1 \leq i \leq n} d_i = V; \\ \forall i, j (\neq i), d_i \cap d_j = \emptyset \end{array} \right\}.$$

All elements, d_i, in C of size two or more are considered to be predicted complexes, and called *predicted clusters* of proteins to distinguish them from known complexes. PPSampler2 and PPSampler2-PIME generate a single partition as an output.

A. Regularization of protein clusters by mutually exclusive interactions

An overview of our approach in this work is explained as follows. In the process of predicting complexes, a candidate for them, which is a set of proteins, is checked whether every pair of proteins within the cluster is connected via PPIs under the condition that, for every set of the PPIs that are mutually exclusive to each other, all but at most one PPI are temporarily assumed not to exist. Among all possible combinations of such deletions of these mutually exclusive PPIs, if there is a case where every pair of proteins is still connected, the cluster is allowed to exist.

We here formulate a regularization term based on mutually exclusive interactions which decides whether a predicted cluster is connected without violating mutual exclusion constraints. Suppose that we have two interactions between two proteins, h and p_1, and between h and another protein, p_2, and that p_1 and p_2 can not interact with h simultaneously due to physical constraints. In this case, we say that these two interactions are mutually exclusive, and h is called the host protein of these interactions. We call these interactions a PIME (pair of interactions which are mutually exclusive).

In our previous methods, PPSampler1 and PPSampler2, any predicted clusters of proteins are required to be connected via the internal PPIs, *i.e.*, there should be a path of adjacent PPIs between every pair of proteins within a cluster. This is a biologically reasonable constraint on predicted clusters.

In this work, we design a more biologically plausible connectivity constraint based on PIMEs. Consider the following simple example. Suppose that there is a connected cluster including $n+1$ proteins, $p_1, \ldots, p_n$, and h, and that $p_1, \ldots, p_n$ interact with h and these interactions are mutually exclusive. Namely, every pair of interactions, $\{p_i, h\}$ and $\{p_j, h\}$ is a PIME. Under this context, all but at most one interaction among the n interactions are assumed not to exist within the cluster simultaneously, from the definition of PIME, when determining whether the internal PPI structure of the cluster is still a single connected graph with those PIMEs. If one of the resulting PPI structures is connected, we say that the cluster is connected with the PIMEs, and otherwise it is disconnected.

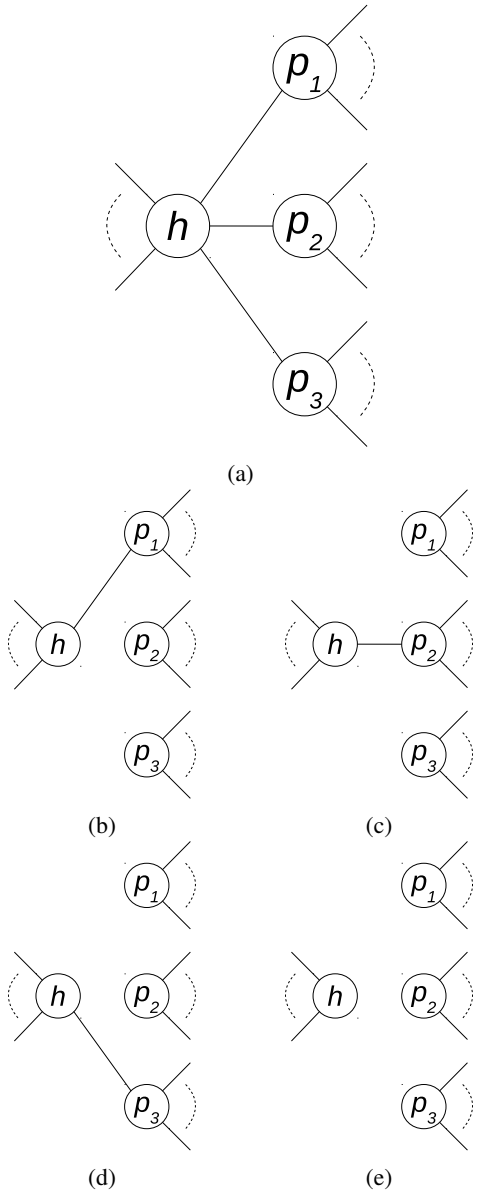


Fig. 1: Example of PIMEs. (a) a cluster including four proteins, p_1, p_2, p_3, and h, where every pair of three interactions between h, and p_1, p_2 and p_3 is a PIME. The four subsequent graphs, (b)-(e), represent different combinations of removed interactions.

Here is an example when $n = 3$ in Fig. 1. Fig. 1a is a given cluster including four proteins, p_1, p_2, p_3, and h. We suppose that every pair of three interactions between h and p_1, p_2 and p_3 is a PIME. In this case, there are four combinations of removed interactions, which correspond to Fig. 1b, 1c, 1d, and 1e.

Consider a set of interactions of a host protein, h, with proteins, $p_1, p_2, \ldots, p_n$. If every pair of two interactions, $\{p_i, h\}$ and $\{p_j, h\}$, is a PIME, we call the set of the interactions a *clique* because the graph where a node is an interaction and an edge represents a PIME is a clique.

We then describe our algorithm for determining whether a connected cluster of proteins is still connected with PIMEs. Recall that the input data sets to our method are a set of PPIs and a set of PIMEs. If an interaction of a PIME is not included in the PPI set, the PIME is discarded.

1) We here consider every PIME whose three proteins are all included in the cluster.
2) For the set of all interactions of PIMEs with the same host protein, all maximal cliques of the set are generated. For example, suppose that the pair of $\{p_1, h\}$ and $\{p_2, h\}$ and that of $\{p_2, h\}$ and $\{p_3, h\}$ are mutually exclusive but that of $\{p_1, h\}$ and $\{p_3, h\}$ is not. In this case, there are two maximal cliques, which are $\{\{p_1, h\}$ and $\{p_2, h\}\}$ and $\{\{p_2, h\}$ and $\{p_3, h\}\}$.
3) At most one interaction in every maximal clique is allowed to exist within the cluster and the remaining interactions are assumed not to exist due to the mutual exclusiveness. We then consider a truth assignment for each maximal clique in which at most one interaction is assumed to be true, *i.e.*, to exist and the others are false, *i.e.*, not to exist. Notice that different maximal cliques can share the same interactions. If an interaction is shared by multiple maximal cliques, the assignment values in different maximal cliques for the interaction should be consistent.
4) For each truth assignment, we check whether the cluster is still connected. If the cluster is disconnected in all truth assignments, the score of this cluster is set to be the infinity. This means that any collection of clusters that includes this cluster is not acceptable. Otherwise, the resulting value of this function is zero, which means no contribution to the current score.

For a partition C of proteins, the function, $h_{\rm pime}(C)$, returns the sum of the output values for the above algorithm for the clusters in C, which is eventually either the infinity or zero.

B. PPSampler2-PIME

Our previous method, PPSampler2, minimizes the scoring function

$$f(C) = h_{\rm den}(C) + h_{\rm sz}(C) + h_{\rm np}(C),$$

for a partition, C, of proteins by a MCMC-based sampling method [9]. The terms of the scoring function, which are $h_{\rm den}(C)$, $h_{\rm sz}(C)$, and $h_{\rm np}(C)$, are formulated as follows.

The first term, $h_{\rm den}(C)$, is defined as the sum of $h_{\rm den}(d)$ over each cluster $d \in C$ where

$$h_{\rm den}(d) = \begin{cases} 0 & \text{if } |d| = 1, \\ \infty & \text{else if } |d| > N \text{ or} \\ & \quad d \text{ is disconnected}, \\ -density(d) & \text{otherwise}, \end{cases}$$

where N is the possible maximum size of clusters in C and $density(d)$ is a weighted density measure defined as

$$density(d) = \frac{1}{\sqrt{|d|}} \sum_{u,v(\neq u) \in d} w(u, v).$$

The second term, $h_{\rm sz}(C)$, regulates the distribution of sizes of predicted clusters in C. It is defined as

$$h_{\rm sz}(C) = \sum_{i=2}^{N} \frac{(\psi_C(i) - \psi(i))^2}{2\sigma_{2,i}^2}$$

where $\psi_C(i)$ is the relative frequency of clusters of size i in C for $i = (2, 3, \ldots, N)$, and $\psi(i)$ is the power-law distribution ranged from 2 to N,

$$\frac{1}{\sum_{j=2}^{N} j^{-\gamma}} \cdot i^{-\gamma}$$

where γ is the power-law parameter. The reason why we add this term is as follows. In major collections of curated complexes of organisms, including the CYC2008 database for yeast [10] and human complexes in the MIPS CORUM database [11], the number of complexes of size i is approximately proportional to the power-law [12].

The function $h_{\rm np}(C)$ regulates the number of proteins within clusters of size two or more in C. It is defined as

$$h_{\rm np}(C) = \frac{(s(C) - \lambda)^2}{2\sigma^2}$$

where $s(C)$ is the number of proteins within clusters of size two or more in C, *i.e.*,

$$s(C) = \sum_{d \in C s.t. |d| \geq 2} |d|.$$

λ is a parameter representing the target value of $s(C)$. σ^2 is the variance parameter in the resulting Gaussian probability density function.

The scoring function of PPSampler2-PIME is obtained by adding $h_{\rm pime}(C)$ to that of PPSampler2. We explain how to minimize these scoring functions in the next section.

C. MCMC algorithm

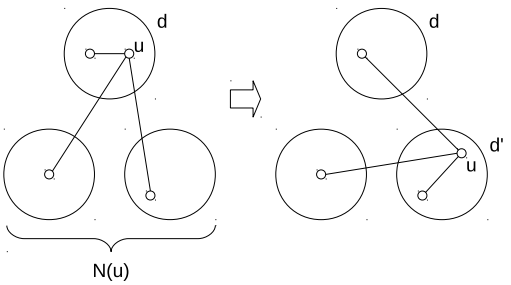


Fig. 2: The proposal function of PPSampler2. A randomly chosen protein, u, will be moved to one of the neighboring clusters to u via PPIs.

The optimization algorithm of the original version of PPSampler2 [9] can be applied to the scoring function, $f(C)$, of PPSampler2-PIME. The search algorithm is a Metropolis-Hastings algorithm. The next candidate partition of proteins is proposed from a current partition by the following algorithm (see Fig. 2).

1) A protein, u, in a current partition, C, is chosen uniformly at random.
2) The cluster including u is denoted by d.
3) With a constant probability, β, u is removed from d and a singleton cluster of u is newly generated.
4) With the remaining probability, $1-\beta$, by the following procedure, u is moved from d to a neighboring cluster to u via PPIs:
   a) All proteins which share an interaction with u are sorted by the interaction weights in descending order. The ith protein in the sorted list is denoted by v_i.
   b) A neighboring cluster, d', is chosen at random with the probability proportional to

$$\sum_{i \text{ s.t. } v_i \in d', \{u, v_i\} \in E} 1/i^2.$$

It should be noted here that one of the advantages of PPSampler2 is that the search algorithm theoretically works for arbitrary scoring functions of a partition of proteins.

D. Matching metrics

In this section, we describe how to evaluate a collection of clusters predicted by a prediction method for protein complexes.

The overlap ratio between a predicted cluster, d, and a known complex, k is measured by the Jaccard index between d and k,

$$\frac{|d \cap k|}{|d \cup k|}.$$

We say that d matches k at matching threshold η if $\frac{|d \cap k|}{|d \cup k|} \geq \eta$.

Based on this matching criterion, we define the precision and recall measures for a set, C, of predicted clusters and a set, K, of known complexes as follows:

$$precision(C, K) = \frac{|N_{pc}(C, K)|}{|C|},$$

where

$$N_{pc}(C, K) = \{d | d \in C, \exists k \in K, d \text{ matches } k\},$$

and

$$recall(C, K) = \frac{|N_{kc}(C, K)|}{|K|},$$

where

$$N_{kc}(C, K) = \{k | k \in K, \exists d \in C, d \text{ matches } k\}.$$

The *F-measure* of C to K is defined as the harmonic mean of the corresponding precision and recall, *i.e.*,

$$F(C, K) = 2 \cdot \frac{precision(C, K) \cdot recall(C, K)}{precision(C, K) + recall(C, K)}.$$

In this work, we use the same very stringent matching criteria as Yong and Wong [4], in which the matching threshold is 1 for the small known complexes of size two and three, and it is 0.75 for the larger complexes.

III. Results

A. Materials

The input data set to PPSampler2-PIME is sets of PPIs and PIMES. The input set of PPIs used in this work is obtained from the WI-PHI [13] database. It includes 49,607 non-self interactions with 5,953 proteins. The degree of a protein is averagely 16.7. The raw weights of the interactions range from 6.6 to 146.6. The normalized weights, which are divided by the maximum value, are used.

Jung *et al.* [5] gave 458 PIMEs found by using the crystal structures recorded in PDB [14]. Among them, 430 PIMEs are those of which both interactions are non-self interactions. Furthermore, among the 430 PIMEs, 304 PIMEs are those of which both interactions are also included in WI-PHI. These 304 PIMEs include 73 host proteins. These 304 PIMEs are used as an input set.

CYC2008 [10] is a protein complex database of *S. cerevisiae*, which contains 408 curated complexes. It is reported in [12] that among the complexes, 172 (42%) and 87 (21%) complexes are of size two and three, respectively.

B. Performance comparison

To evaluate the effectiveness of the regularization term based on mutually exclusive interactions, we have conducted performance comparison of PPSampler2-PIME with various existing tools, MCL [15], MCODE [7], DPClus [16], CMC [17], COACH [18], RRW [19], NWE [20] as well as our previous tool, PPSampler2 [9].

TABLE I: The values of parameters of PPSampler2.

Parameter	Value
λ	2,000
σ^2	10^6
γ	2
$\sigma^2_{2,i}$	$10^3/1.1^i$
Number of iterations	2×10^6
β	0.01

The parameter settings of the seven existing methods are the same as in our previous work [12], which are almost the default settings. However the resulting performance scores reported here are different from those in the previous work [12] because, as mentioned before, the more stringent matching criteria is applied to the predictions of those methods. PPSampler2 and PPSampler2-PIME are executed with the default parameter values, shown in Tab. I, 20 times.

Fig. 3 shows the results on precision, recall, and F-measure calculated from all predicted complexes by PPSampler2 and PPSampler2-PIME. The bar height shows the mean of a performance measure and the error bar shows $\pm$ one standard deviation. As we can see, the precision, recall, and F-measure are slightly but all improved by using PIMEs, from 0.279 to 0.283, from 0.279 to 0.288, and from 0.279 to 0.286, respectively. We have also conducted a two-sample t-test for difference in the mean of the F-measure values of PPSampler2 and that of PPSampler2-PIME. The resulting p-value is 0.0274. This means that the null hypothesis that the two means

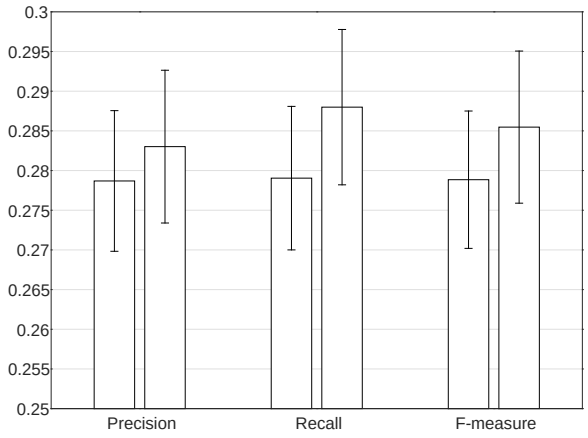


Fig. 3: Precision, recall, and F-measure of PPSampler2, and PPSampler2-PIME. In each group of precision, recall, and F-measure, the left and right bars correspond to PPSampler2 and PPSampler2-PIME, respectively. The height of a bar represents the mean, and the error bar shows $\pm$ one standard deviation.

are the same is rejected with the ordinary significance level of 0.05. This statistical significance is a good evidence supporting that our method with PIMEs can improve the predictability of protein complexes. Thus, we empirically conclude that it is clearly effective to regulate predicted clusters of proteins by adding a regularization term based on mutual exclusiveness PPIs.

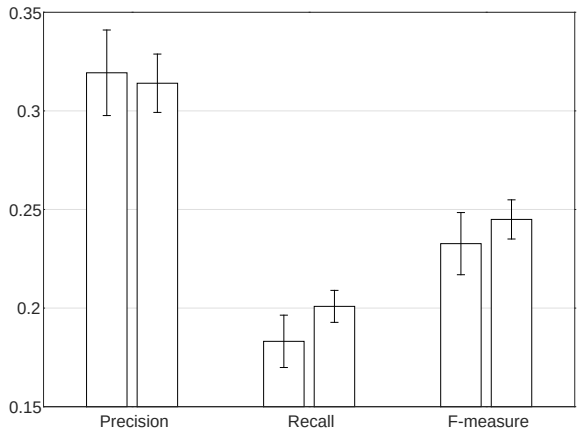
Details of performance of PPSampler2, PPSampler2-PIME, and the other methods are given in Tab. II. We have not included the other methods in Fig. 3 because their F-measure values except that of NWE are considerably lower than those of PPSampler2 and PPSampler2-PIME.

C. Size-specific evaluation

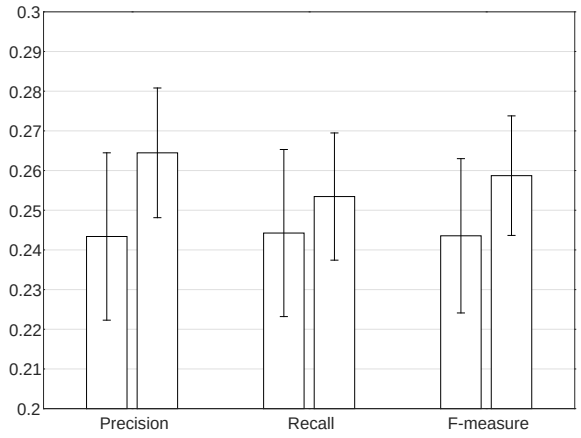
Recall that, among 408 CYC2008 complexes, 172 (42%) and 87 (21%) complexes are of size two and three, respectively. Thus, it is interesting to see how well size-specific predictions perform. In addition, some of existing methods are designed for predicting relatively larger complexes. For example, COACH does not predicted complexes of size two from the WI-PHI PPIs.

We formulate size-specific precision, recall, and F-measure as follows. For a set of clusters of proteins, C, let $C|_i = \{c \in C||c| = i\}$ and $C|_{\geq i} = \{c \in C||c| \geq i\}$. In the same way, we define the notations of $K|_i$ and $K|_{\geq i}$. We then calculate the F-measure values w.r.t. size two, size three, and size four or more from $precision(C|_2, K|_2)$ and $recall(C|_2, K|_2)$, $precision(C|_3, K|_3)$ and $recall(C|_3, K|_3)$, and $precision(C|_{\geq 4}, K|_{\geq 4})$ and $recall(C|_{\geq 4}, K|_{\geq 4})$, respectively.

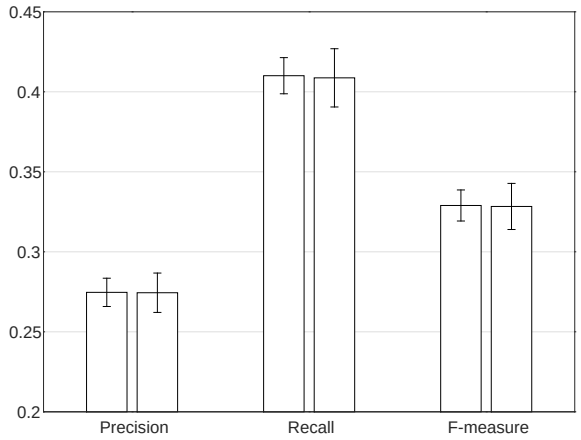
The resulting precision, recall, and F-measure values of PPSampler2 and PPSampler2-PIME are shown in Fig. 4. Fig. 4a shows the result w.r.t. size-two complexes In precision,



(a)



(b)



(c)

Fig. 4: Size-specific performance results. (a), (b), and (c) show the results on size two, size three, and size four or more, respectively. In each of precision, recall, and F-measure, the left and right bars correspond to PPSampler2 and PPSampler2-PIME, respectively.

TABLE II: The matching results of all predicted clusters. #protein shows the number of proteins within predicted clusters. #cluster gives the number of predicted clusters. Avg. size represents the average size of predicted clusters. Additional numbers are standard deviations.

Tool name	#protein	#cluster	Avg. size	N_{pc}	N_{kc}	Precision	Recall	F
MCL	5869	880	6.7	12	13	0.014	0.032	0.019
MCODE	2432	156	15.6	2	2	0.013	0.005	0.007
DPClus	4888	925	6.9	18	19	0.019	0.047	0.027
CMC	4845	3613	4.0	103	100	0.029	0.245	0.051
COACH	4094	1353	13.3	27	22	0.020	0.054	0.029
RRW	4240	1984	2.1	87	88	0.044	0.216	0.073
NWE	1626	720	2.3	117	118	0.163	0.289	0.208
PPSampler2	2010.4±0.9	398.2±6.3	5.1±0.1	111.0±3.6	113.9±3.7	0.279±0.009	0.279±0.009	0.279±0.009
PPSampler2-PIME	2009.9±0.4	405.3±4.1	5.0±0.1	114.7±3.9	117.5±4.0	0.283±0.010	0.288±0.010	0.286±0.010

TABLE III: The matching results of predicted clusters of size two. #protein shows the number of proteins within predicted clusters. #cluster gives the number of predicted clusters. Additional numbers are standard deviations.

Tool name	#protein	#cluster	$N_{pc}(= N_{kc})$	Precision	Recall	F-measure
MCL	462	231	4	0.017	0.023	0.020
MCODE	6	3	0	0	0	0
DPClus	2	1	0	0	0	0
CMC	1534	767	23	0.030	0.134	0.049
COACH	0	0	0	0	0	0
RRW	3648	1824	55	0.030	0.320	0.055
NWE	1264	632	78	0.123	0.453	0.194
PPSampler2	197.4±8.4	98.7±4.2	31.5±2.3	0.319±0.022	0.183±0.013	0.233±0.016
PPSampler2-PIME	220.2±6.6	110.1±3.3	34.6±1.4	0.314±0.015	0.201±0.008	0.245±0.010

TABLE IV: The matching results of predicted clusters of size three. #protein shows the number of proteins within predicted clusters. #cluster gives the number of predicted clusters. Additional numbers are standard deviations.

Tool name	#protein	#cluster	N_{pc}	Precision	Recall	F-measure
MCL	456	152	0	0	0	0
MCODE	162	54	1	0.019	0.011	0.014
DPClus	120	40	1	0.025	0.011	0.016
CMC	2519	1576	20	0.013	0.230	0.024
COACH	60	20	0	0	0	0
RRW	309	103	14	0.136	0.161	0.147
NWE	162	54	19	0.352	0.218	0.270
PPSampler2	262.5±17.6	87.5±5.9	21.3±1.8	0.243±0.021	0.244±0.021	0.244±0.019
PPSampler2-PIME	250.4±11.2	83.5±3.7	22.1±1.4	0.265±0.016	0.253±0.016	0.259±0.015

TABLE V: The matching results of predicted clusters of size four or more. #protein shows the number of proteins within predicted clusters. #cluster gives the number of predicted clusters. Avg. size represents the average size of predicted clusters. Additional numbers are standard deviations.

Tool name	#protein	#cluster	Avg. size	N_{pc}	N_{kc}	Precision	Recall	F-measure
MCL	4951	497	10.0	8	9	0.016	0.060	0.025
MCODE	2264	99	22.9	1	1	0.010	0.007	0.008
DPClus	4799	884	7.1	17	18	0.019	0.121	0.033
CMC	2839	1270	6.4	60	57	0.047	0.383	0.084
COACH	4052	1333	13.4	27	22	0.020	0.148	0.036
RRW	283	57	5.0	18	19	0.316	0.128	0.182
NWE	200	34	5.9	20	29	0.588	0.195	0.292
PPSampler2	1550.5±19.7	212.0±3.5	7.3±0.1	58.2±1.6	61.1±1.7	0.275±0.009	0.410±0.011	0.329±0.010
PPSampler2-PIME	1539.4±13.7	211.8±3.6	7.3±0.1	58.1±2.6	60.9±2.7	0.274±0.012	0.415±0.018	0.330±0.014

the score of PPSampler2-PIME is slightly lower than that of PPSampler2. However, in recall, the score of PPSampler2-PIME is higher than that of PPSampler2. The difference in recall is larger than that in precision. As a result, in F-measure, the score of PPSampler2-PIME is higher than that of PPSampler2. It is 5% improvement. The result that performance is improved in size-two complexes is interesting because the regularization term, h_{pime}, checks the connectivity of clusters of *size three or more* with PIMEs and does not directly regularize clusters of size two. However, the method with PIMEs is superior to the original ones. The reason will be that the direct regularization by h_{pime} to clusters of size three or more indirectly improves the quality of size-two clusters, because the structure of predicted clusters is a partition of proteins and modification of a cluster affects another. Details of the matching results including the other methods can be found in Tab. III.

The result w.r.t. size-three complexes is given in Fig. 4b. As can be seen, in precision, recall, and F-measure, performance is improved by using PIMEs. However, the score of F-measure is still lower than NWE (see Tab. IV), although the score of F-measure of NWE for all complexes is lower than those of PPSampler2 and PPSampler2-PIME.

The result w.r.t. size four or more is given in Fig. 4c. In every measure, the scores of both methods are almost the same, though predicted clusters of those sizes are directly regulated by the regularizer based on mutually exclusive interactions. Details of the matching results can be found in Tab. V.

From the analysis of size-specific evaluation of predicted clusters, we can see that the improvement in F-measure of PPSampler2-PIME from PPSampler2 for all predicted complexes is brought by the improvements w.r.t. size-two and three complexes.

D. Precision-recall graph

Fig. 5 shows a precision-recall graph of PPSampler2 and PPSampler2-PIME. To draw this graph, we used the cohesiveness of a cluster, d, which is defined as

$$\frac{w^{in}(d)}{w^{in}(d) + w^{bound}(d)}$$

where

$$\begin{aligned} w^{in}(d) &= \sum_{u,v(\neq u)\in d} w(u,v), \\ w^{bound}(d) &= \sum_{u\in d, v\notin d} w(u,v). \end{aligned}$$

All predicted clusters of an output are sorted in descending order of cohesiveness, and the precision and recall scores for the top $H\%$ sorted clusters are calculated for $H = 10/3, 20/3, 30/3, \cdots, 100$. Note that the point for all clusters is the rightmost one in the graph. We can see that most of the points of PPSampler2-PIME, except the points in the left region of recall less than 0.1, are located in the right upper region of those of PPSampler2. Though the gaps between the corresponding points are not large, this result is also a good evidence that information on mutually exclusive interaction can contribute to the improvement of performance of protein complex prediction.

IV. Conclusions

We have proposed a regularization term, based on mutually exclusive interactions, which evaluates whether a predicted complex is still connected when mutually exclusive interactions are removed, and presented a prediction method, PPSampler2-PIME, by adding the term to the scoring function of our earlier protein complex prediction method, PPSampler2. PPSampler2-PIME outperforms PPSampler2, and the other conventional methods. Thus we empirically conclude that information of mutually exclusive PPIs is effective for predicting complexes more precisely. In addition, this result implies that if more accurate and genome-wide information on mutually exclusive PPIs are obtained, the resulting prediction will be superior to the current ones.

Furthermore, an interesting future work in protein complex prediction is the modeling of the internal PPI structure of predicted complexes. If we model it in some appropriate way, we will predict complexes more reliably. Information on the dynamics of proteins and PPIs will be a good source for this modeling.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 26330330.

References

[1] S. Brohée and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, p. 488, 2006.

[2] X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Computational approaches for detecting protein complexes from protein interaction networks: a survey," *BMC Genomics*, vol. 11(suppl 1), p. S3, 2010.

[3] S. Srihari and H. W. Leong, "A survey of computational methods for protein complex prediction from protein interaction networks," *Journal of Bioinformatics and Computational Biology*, vol. 11, p. 1230002, 2013.

[4] C. H. Yong and L. Wong, "From the static interactome to dynamic protein complexes: Three challenges," *Journal of Bioinformatics and Computational Biology*, vol. 13, p. 1571001, 2015.

[5] S. H. Jung, B. Hyun, W.-H. Jang, H.-Y. Hur, and D.-S. Han, "Protein complex prediction based on simultaneous protein interaction network," *Bioinformatics*, vol. 26, pp. 385–391, 2010.

[6] Y. Ozawa, R. Saito, S. Fujimori, H. Kashima, M. Ishizaka, H. Yanagawa, E. Miyamoto-Sato, and M. Tomita, "Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions," *BMC Bioinformatics*, vol. 11, p. 350, 2010.

[7] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, p. 2, 2003.

[8] X.-L. Li, S.-H. Tan, C.-S. Foo, and S.-K. Ng, "Interaction graph mining for protein complexes using local clique merging," in *Genome Informatics*, vol. 16, 2005, pp. 260–269.

[9] C. K. Widita and O. Maruyama, "PPSampler2: Predicting protein complexes more accurately and efficiently by sampling," *BMC Systems Biology*, vol. 7(Suppl 6), p. S14, 2013.

[10] S. Pu, J. Wong, B. Turner, E. Cho, and S. Wodak, "Up-to-date catalogues of yeast protein complexes." *Nucleic Acids Res.*, vol. 37, pp. 825–831, 2009.

[11] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H.-W. Mewes, "CORUM: the comprehensive resource of mammalian protein complexes-2009," *Nucleic Acids Res.*, vol. 38, pp. D497–D501, 2010.

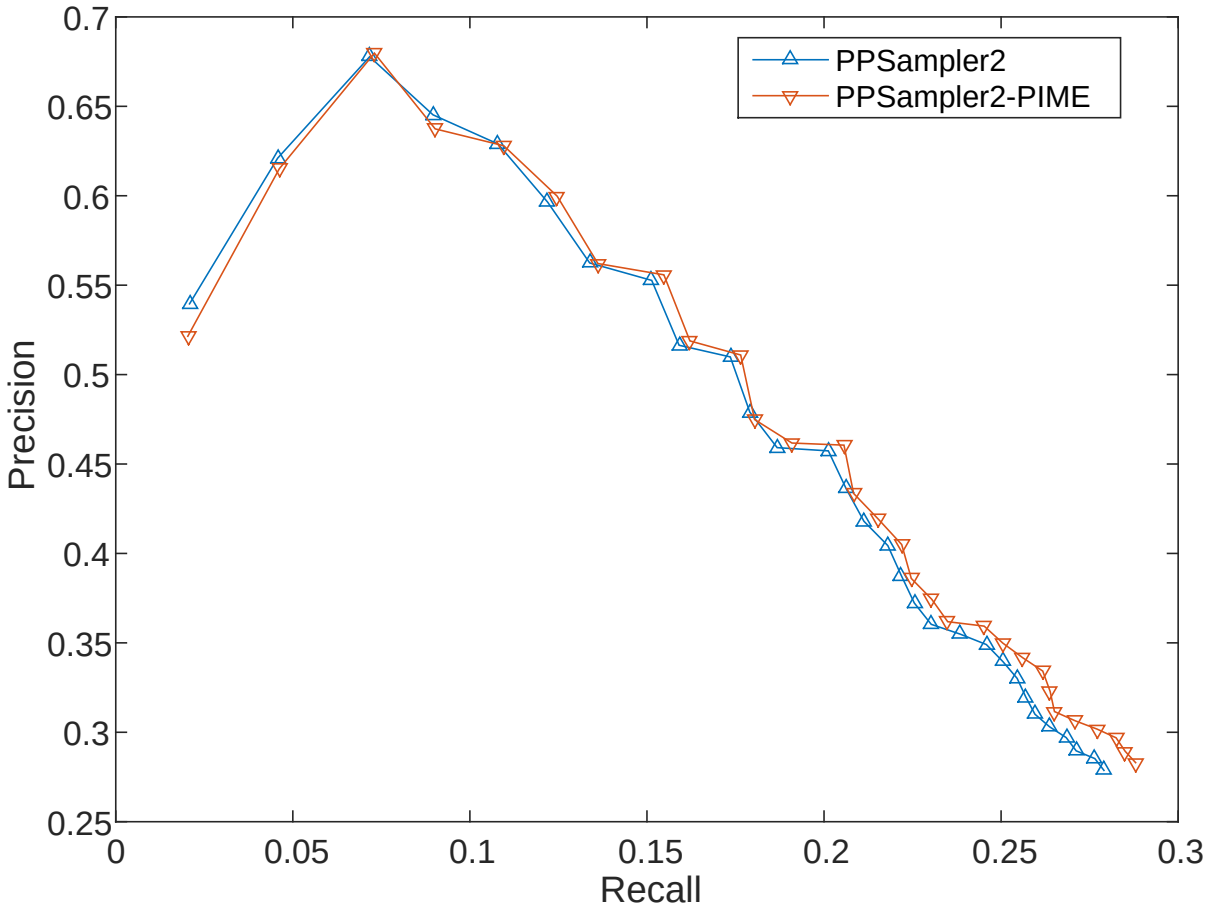


Fig. 5: A precision-recall graph. Each point corresponds to a pair of precision and recall scores for a top $H\%$ clusters sorted by cohesiveness for $H = 10/3, 20/3, 30/3, \cdots, 100$.

[12] D. Tatsuke and O. Maruyama, "Sampling strategy for protein complex prediction using cluster size frequency," *Gene*, vol. 518, pp. 152–158, 2013.

[13] L. Kiemer, S. Costa, M. Ueffing, and G. Cesareni, "WI-PHI: A weighted yeast interactome enriched for direct physical interactions," *Proteomics*, vol. 7, pp. 932–943, 2007.

[14] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[15] A. Enright, S. V. Dongen, and C. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Res.*, vol. 30, pp. 1575–1584, 2002.

[16] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinformatics*, vol. 7, p. 207, 2006.

[17] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics*, vol. 25, pp. 1891–1897, 2009.

[18] M. Wu, X. Li, C. Kwoh, and S. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinformatics*, vol. 10, p. 169, 2009.

[19] K. Macropol, T. Can, and A. Singh, "RRW: Repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinformatics*, vol. 10, p. 283, September 2009.

[20] O. Maruyama and A. Chihara, "NWE: Node-weighted expansion for protein complex prediction using random walk distances," *Proteome Science*, vol. 9(Suppl 1), p. S14, 2011.