BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 3 2010, pages 385–391
doi:10.1093/bioinformatics/btp668

Systems biology

Protein complex prediction based on simultaneous protein interaction network

Suk Hoon Jung[1], Bora Hyun[1], Woo-Hyuk Jang[1], Hee-Young Hur[1] and Dong-Soo Han[2]*

[1]Department of Information & Communications Engineering, Korea Advanced Institute of Science and Technology, 119 Munjiro, Yuseong-gu, Daejeon, 305–714 and [2]Department of Computer Science, Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-gu, Daejeon, 305–701, Korea

Received on February 24, 2009; revised on November 22, 2009; accepted on November 28, 2009

Advance Access publication December 4, 2009

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: The increase in the amount of available protein–protein interaction (PPI) data enables us to develop computational methods for protein complex predictions. A protein complex is a group of proteins that interact with each other at the same time and place. The protein complex generally corresponds to a cluster in PPI network (PPIN). However, clusters correspond not only to protein complexes but also to sets of proteins that interact dynamically with each other. As a result, conventional graph-theoretic clustering methods that disregard interaction dynamics show high false positive rates in protein complex predictions.

Results: In this article, a method of refining PPIN is proposed that uses the structural interface data of protein pairs for protein complex predictions. A simultaneous protein interaction network (SPIN) is introduced to specify mutually exclusive interactions (MEIs) as indicated from the overlapping interfaces and to exclude competition from MEIs that arise during the detection of protein complexes. After constructing SPINs, naive clustering algorithms are applied to the SPINs for protein complex predictions. The evaluation results show that the proposed method outperforms the simple PPIN-based method in terms of removing false positive proteins in the formation of complexes. This shows that excluding competition between MEIs can be effective for improving prediction accuracy in general computational approaches involving protein interactions.

Availability: http://code.google.com/p/simultaneous-pin/

Contact: dshan@kaist.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Recent developments in biotechnology have resulted in an increase in the amount of protein–protein interaction (PPI) data. Modeling a PPI network (PPIN) with simple graphs enables many computational methods for the study of protein functions (Broheè and Helden, 2006; Han *et al*., 2004), one of which is known as the automatic protein complex prediction method. Protein complexes generally correspond to clusters in a PPIN because proteins in a complex are highly interactive with each other (Tong and Drees, 2002). Therefore, computational methods for protein complex predictions, such as MCODE (Molecular Complex Detection; Bader and Hogue, 2003), LCMA (Local Clique Merging Algorithm; Li *et al*., 2005), SPC (Super Para-magnetic Clustering; Blatt *et al*., 1997), RNSC (Restricted Neighborhood Search Clustering; King *et al*., 2004) and DPClus (Altaf-Ul-Amin *et al*., 2006; Li *et al*., 2008), typically focus on the extraction of clusters based on the graph theory.

One specific problem pertaining to conventional methods originates from the fact that with these methods, a PPIN is regarded as a static entity. In reality, a PPIN is not a static but a dynamic entity; the functional state of the network depends on the expression of protein nodes, which is intrinsically controlled by different regulatory mechanisms through time and space (Han *et al*., 2004; Liang and Li, 2007). In a dynamic network, a protein complex is a group of proteins in which individual proteins interact with each other at the same time and place (Spirin and Mirny, 2008). However, a cluster in a PPIN may include proteins that interact dynamically with each other as well. Conventional approaches based on a simple PPIN cannot properly distinguish protein complexes from interactions that may be activated at a different time and place because they disregard interaction dynamics. This leads to false positive results in protein complex detections (Spirin and Mirny, 2008).

A means of tackling this problem is to use the features of proteins additionally as indirect evidence. Some methods use machine learning methods, and some others enrich the protein interaction network by assigning weights based on functional annotations; gene expression data; or biological, chemical and physical properties (Pei and Zhang, 2006; Qi *et al*., 2008; Zhang *et al*., 2006). Given that these features are known to be relevant to protein mechanisms in general, considering them in clustering algorithms may improve the prediction results. However, using indirect evidence is not adequate in itself to pinpoint complexes in PPIN because indirect evidence is not determinative in complex formations.

Unlike previous approaches, this article focuses on competitive interactions in a PPIN, which are considered to provide more direct and determinative evidence in the identification of proteins in the formation of a complex. Interactions must occur simultaneously in a protein complex; consequently, excluding interaction competitions is a necessary condition during the formation of a complex. Thus, competitive interactions in a PPIN should be prudently selected before they are included in a predicted complex.

In addition, many proteins are known to usually have a number of interacting partners, some of which may cooperate or even

*To whom correspondence should be addressed.

compete for the activation of certain functions. The cooperation and competition between partners of a protein are controlled by different regulatory mechanisms (Colley *et al*., 1997), and they determine which function is going to be activated among those the protein may serve (Bryce *et al*., 2001; Pierrat *et al*., 2007; Tabuchi *et al*., 2002). Moreover, some proteins are reported to have alternative interaction partners, which are competitive, but their alternation serves the same or a similar function (Elion *et al*., 1991; Qi and Elion, 2005). Consequently, the cooperation and competition may generate variations in the formations of complexes and functional modules that overlap with each other (Hu *et al*., 2005; Valente *et al*., 2009).

Among a number of interaction partners, detecting the cooperative partners for a certain function is essential for an understanding of functional mechanisms of proteins. However, too few genes have been studied through experiments, which are typically accomplished only with great difficulty. Therefore, in this research, an understanding of the cooperation between a protein and its partners is approached by eliminating instances of interaction competition through computations.

In this article, a network model is developed that incorporates interaction competition information drawn from the structural interface data of protein domains. A framework using the network model for graph-theoretic clustering methods is then proposed for protein complex predictions. The network model, simultaneous protein interaction network (SPIN), captures different sets of non-competitive interactions extracted from the original PPIN. Network clustering on non-competitive interactions excludes superfluous members in the formation of a protein complex.

This research seeks instances of interaction competition based on interaction interfaces. Many competitive interactions are mediated by the same interfacial surface. More than one protein cannot physically bind to the same or an overlapping surface on a protein at the same time; such interactions are identified as mutually exclusive interactions (MEIs; Hu *et al*., 2005; Kim *et al*., 2006)

A SPIN is a simple graph composed of nodes and edges which allows any naive graph-theoretic clustering algorithm to be applied to it to computationally predict protein complexes. In this article, MCODE and LCMA are applied to SPINs, as constructed from *Saccharomyces cerevisiae* (yeast) interactome and are then applied to plain yeast PPIN for comparison. The prediction results are compared with experimentally derived yeast protein complexes recorded in the MIPS complex database (Guldener *et al*., 2006).

According to the result analysis, SPIN-based clustering outperforms simple PPIN-based clustering. Our model results in a significantly improved *F1*-score when compared with PPIN-based methods. Moreover, it detects all of the complexes detected by PPIN-based clustering while also generating additional true positives in all thresholds. This result was possible because only superfluous members for a complex formation were removed by the SPIN-based method apart from a small number of cases.

2 METHOD

2.1 Competition between MEI partners

A close look into the physical interfaces between interacting proteins provides information on mutual exclusiveness among the interacting partners of a protein, and mutual exclusiveness results in interaction competition. If two or more interaction partners can bind to a common or an overlapping interfacial surface of a protein, the surface is considered to be physically available only for one partner at a given moment. Such interactions are mutually exclusive, as the occurrence of one of these interactions automatically excludes the occurrence of the remaining interactions. A target protein whose partners compete for the interaction is termed the *host protein* in this article. In addition, the term MEI is used to denote a pair of interactions that is mutually exclusive for a host protein. A case in which more than two interfacial surfaces are overlapped or partially cascaded is represented by a set of MEIs. Figure 1 depicts an example of the modeling of an MEI.

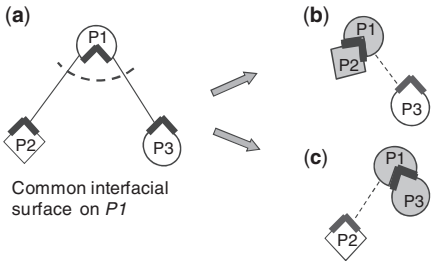

Fig. 1. An example of MEIs: (**a**) Two proteins, *P2* and *P3*, can bind to a common surface on *P1*. (**b** and **c**) Only one of these interactions can occur at a given moment because the surface on protein *P1* is available only for one interaction.

The first step in the detection of MEIs is to identify the interface of each protein interaction, which is represented by a set of interfacial residue pairs. In this research, an interface between a protein pair is examined at the level of the protein domain. The protein domain is an evolutionary conserved unit of the structure and function of the protein; therefore, it is regarded as a subunit that mediates PPIs (Boxem *et al*., 2008).

Figure 2 illustrates the process of MEI extraction using PSIMAP (Gong *et al*., 2005). PSIMAP provides information pertaining to interfacial residue pairs in physical domain–domain interactions (DDI) based on an analysis of the crystal structures of proteins, the protein interacting pairs and the complexes recorded in the PDB (Berman *et al*., 2000). Similar to PSIMAP, this study adopts the SCOP domain definition. For each domain, we compute overlapping interfacial residues for all possible pairs of partners with which the domain interacts.(Fig. 2b and c) In this process, self-pairing of each partner domain should be considered as well because a protein may have several interacting partner proteins mediated by an identical DDI.

Another consideration is that a pair of domains can interact through several different interfaces (Aragues *et al*., 2007; Winter *et al*., 2006). Hence, although two partner domains seem to have an overlapping binding site on a host domain, they could still bind simultaneously by using disjoint alternative binding sites on the host. Therefore, two partners are recognized to be mutually exclusive if and only if they have no other option but to compete for an overlapping interfacial surface on the host domain.

The next step is protein domain assignment by referring Interpro (Hunter *et al*., 2009) that offers integrative protein signature data. A DDI interface is used in identifying the interface of a PPI mediated by the corresponding DDI, and MEIs are inferred by referring to the mutually exclusive DDI data that is obtained (Fig. 2d). In this process, the DDI within a protein is ignored because its interface is considered to be already occupied by an intra-molecular interaction (Gong *et al*., 2005).

It is possible to represent an MEI relationship using a Boolean expression. In conventional network model, an interaction is represented with a static edge regardless of the time and/or conditions. With this conjecture, the interaction list of a protein can be represented as a *conjunction* of all interactions where an interaction has a value of true when it occurs. However, two interactions of a MEI should be connected by *XOR* ($\oplus$) as both cannot occur simultaneously.

Figure 3 illustrates an example of representing MEI information in a simple network. The notation $xInt_{pi}$ is used to represent interactions with

Fig. 2. (**a**) 3D structures of proteins and complexes recorded in PDB. (**b**) PSIMAP detects interfacial residues between domains. (**c**) A DDI map including the information of mutually exclusive interfaces. (**d**) Two PPIs are mutually exclusive when their interaction structures correspond to mutually exclusive DDIs.

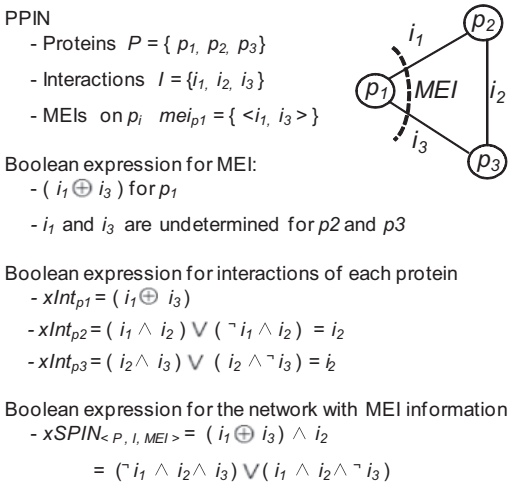

Fig. 3. Boolean expression of MEI information for a simple network.

MEI information for protein p_i using a Boolean expression. In the network, p_1 has an MEI pair, $< i_1, i_3 >$; thus, its interactions are represented as $xInt_{p1} = (i_1 \oplus i_3)$. Another consideration for representing MEI information is that the interaction of an MEI that will occur cannot be determined. Therefore, interactions of the protein p_2 are represented as $xInt_{p2} = (i_1 \wedge i_2) \vee (\neg i_1 \wedge i_2)$, and consequently i_2, which ignores i_1, which participates in the MEI process on counterpart protein p_1.

Using $xInt_{pi}$ annotated to each protein, the Boolean expression $x\text{SPIN}_{<P,I,\text{MEI}>}$ is generated to represent the MEI information in a PPIN, where P is a protein set, I is an interaction set and MEI is a set of MEIs in the network. $x\text{SPIN}_{<P,I,\text{MEI}>}$ is reserved by the *conjunction* of $xInt_{pi}$ for all proteins in the network. Accordingly, it represents all interactions and mutually exclusive relationships in the network. In the *disjunctive normal form* (DNF) of $x\text{SPIN}_{<P,I,\text{MEI}>}$, each *conjunctive clause* represents a set of non-competitive interactions in the PPIN.

2.2 SPIN

The SPIN is a subnetwork of a PPIN. A SPIN is comprised of a set of non-competitive interactions and all of the proteins inherited from the original network. A non-competitive interaction set selectively includes one of the mutually exclusive pairs of each protein in order to achieve mutual exclusion among the interactions. Therefore, its interactions may be activated simultaneously without competition in nature. SPINs from a PPIN can be viewed as snapshots, each of which represents a possible coactive state that the dynamic network may attain.

Based on the $x\text{SPIN}_{<P,I,\text{MEI}>}$ computed from the MEIs, SPINs are extracted from the PPIN based on each conjunctive clause in $x\text{SPIN}_{<P,I,\text{MEI}>}$. In Figure 4, the PPIN has two MEIs $< i_3, i_5 >$ and $< i_3, i_6 >$; therefore, its set of interactions is represented by $(i_1 \wedge i_2 \wedge \neg i_3 \wedge i_4 \wedge i_5 \wedge i_6) \vee (i_1 \wedge i_2 \wedge i_3 \wedge i_4 \wedge \neg i_5 \wedge \neg i_6)$ in the DNF. As each *conjunctive clause* represents a non-competitive interaction set, two SPINs are generated from each clause.

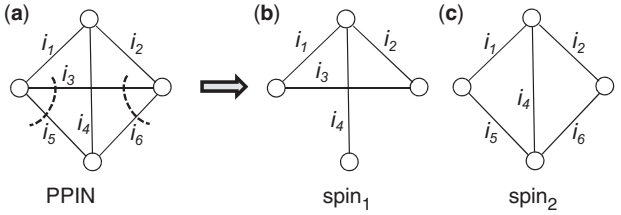

Fig. 4. An example of a SPIN construction: from a PPIN with two MEIs in (**a**), the SPIN construction process generates two SPINs (**b** and **c**).

2.3 SPIN-based framework for protein complex prediction

SPINs are not necessarily generated from the whole interactome because the SPIN is constructed only to find a set of possibly coactivated interactions cooperating for a function. In addition, the computation cost of the SPIN construction process is high because the number of SPINs is at most 2^n with n MEIs based on the two choices of including one interaction or the other for each MEI and the number of nodes in a SPIN is the same as that of the original PPIN. Therefore, it is appropriate to generate SPINs from a subnetwork that is small but nonetheless large enough to include a functional module or a complex. For protein complex predictions, clustering algorithms deal with dense regions in a PPIN, implying that a subnetwork for SPIN construction should cover one of the dense regions.

Figure 5 illustrates SPIN framework for protein complex prediction consisting of three phases. In the proposed framework, subnetwork preparation precedes the SPIN construction process, and a clustering is finally performed on generated SPINs to predict the protein complexes. The subnetwork preparation adopts a naive clustering algorithm which is used in post-clustering as well. Adopting the same clustering algorithm dramatically reduces the computation cost for SPIN construction but does not change the prediction results because the generated subnetworks cover all of the complexes that can be predicted by the post-clustering algorithm. Although a subnetwork is a cluster, a SPIN generated from the network may not be a cluster as it will lose some interactions. Therefore, clustering is performed on generated SPINs in the post-clustering phase.

The proposed framework focuses on the extraction of non-competitive sets of proteins in a PPIN; hence, protein complex detection from extracted sets exploits conventional clustering algorithms. In this research, MCODE and LCMA are adopted from among various conventional graph-theoretic clustering algorithms for the evaluation of the framework.

MCODE (Bader and Hogue, 2003) utilizes connectivity values in a PPIN to detect protein complexes. This algorithm is based on vertex weighting

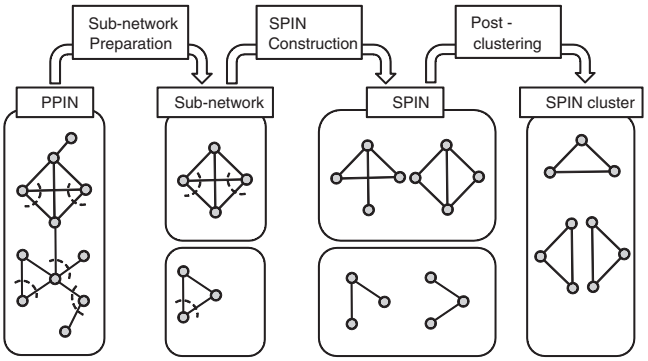

Fig. 5. Outline of the SPIN-based framework.

according to the local neighborhood density and then on an outward traversal from a dense seed protein with a high weighting value to include neighboring vertices recursively whose weight satisfies some given threshold.

LCMA (Li *et al*., 2005; Zhang *et al*., 2006) generates overlapping clusters based on local clique merging. It first locates local cliques for each vertex of the graph and then merges the detected local cliques according to their affinity to form maximal dense subgraphs.

Most clustering algorithms require parameters that serve to influence the accuracy of the prediction results. The MCODE algorithm requires the two parameters known as *WVP* and *Fluff*. In this research, *WVP* was set to 0.1, and *Fluff* was set to 0. These values were expected to extract good results according to the original research of MCODE (Bader and Hogue, 2003). For the same reason, *NA* was set to 0 in the LCMA (Li *et al*., 2005, 2008).

A set of SPINs, which are generated from a PPIN, shares all of the proteins and many of the interactions from the PPIN; hence, extracted clusters may be identical. Therefore, identical results should be trimmed at the end of the post-clustering stage.

When the procedures in Figure 5 are viewed from a different angle, they can collectively be regarded as a filter attachment for conventional clustering methods because the results of subnetwork preparation are identical to those of a conventional clustering method. However, in this study, subnetwork preparation is not considered to be an actual clustering step but is instead considered to be a type of preprocessing for SPIN construction, as the primary purpose of the subnetwork preparation is to reduce the computational costs in this framework.

3 RESULTS

In an effort to evaluate the proposed framework for detecting protein complexes, it was compared with plain PPIN-based clustering methods. Two experiments were performed on a SPIN using the two clustering algorithms MCODE and LCMA. They are considered SPIN-based methods and were termed SPIN_MCODE and SPIN_LCMA, respectively. The same algorithms were also applied to a plain PPIN for comparison, and they are termed PPIN_MCODE and PPIN_LCMA for convenience sake. For fair comparisons, identical parameters for the PPIN- and SPIN-based methods were used in a clustering algorithm.

Additionally, we performed experiments based on random SPIN in a comparison to determine whether or not our improvement stems from using structural MEI information. In these experiments, the same procedure used with the SPIN framework was utilized; however, after a subnetwork preparation step, each prepared subnetwork was assigned with randomly generated MEIs with the same number as its structural MEIs.

The following subsections present the results of the experiments, explain the relationship between the clusters in the SPINs and a plain PPIN, compare the prediction results with known complexes, and discuss the effect of the SPIN-based framework.

3.1 Reference sets

Experiments were performed on the *S.cerevisiae* (yeast) interactome downloaded from the MIPS MPact database (Guldener *et al*., 2006). After removing all the self-interactions, the final network contained 15 524 interactions among 4579 yeast proteins. Two clustering algorithms on two base networks, PPIN and SPIN, generated four predicted cluster sets, and these prediction results were compared with known protein complexes recorded in the MIPS yeast complex database (Guldener *et al*., 2006). There were 267 manually annotated complexes that were considered as gold standard data.

3.2 MEI extraction

From 14 594 multi-domain PDB entities (release date December 5, 2008), PSIMAP extracted 4948 DDIs with 64 985 examples of interface evidence among 2527 domains. Among them, 1842 domains were revealed to have at least one pair of partner domains that were mutually exclusive, and the number of mutually exclusive pairs was 6174 in total. Supplementary Table 1 lists the mutually exclusive DDIs, overlapping residue indexes and PDB evidence.

In PPI network, it was found that 100 proteins had at least one MEI competing for the interaction with them, and there were 458 MEIs in the network. Supplementary Table 2 lists the MEIs on each host protein along with the mutually exclusive DDI pair to which the MEIs refer.

As discussed an actual complex should not have MEIs within it, we investigated the occurrence of MEIs in MIPS complexes. There were 14 MEIs in six out of 267 MIPS complex data. We hypothesized that there might be incomplete interface data. Specifically, host proteins of the 14 MEIs might have an unknown alternate binding site which allow for the MEIs to occur simultaneously.

3.3 The relationship between the clusters in PPIN and SPIN

A SPIN is constructed by refining a PPIN, and proteins in the refined network cannot be more interactive compared with those in the PPIN. Additionally, SPIN-based methods use the same clustering algorithm as comparison methods with the same parameters. Therefore, a SPIN cluster must be a subgraph of the corresponding PPIN cluster.

Table 1 shows a summary of the prediction results of the four methods. LCMAs extract a much larger number of clusters compared with MCODEs because, unlike MCODE, LCMA finds loosely connected clusters that may be overlapped.

Applying the SPIN concept increases the number of predicted clusters in both MCODE and LCMA. However, the number of distinct proteins in SPIN clusters is fewer than that in PPIN clusters. This indicates that the clustering on the SPINs results in the removal of proteins in the original clusters. As many interactions and all of the proteins appear in common in the SPIN and the PPIN, many SPIN clusters are identical to PPIN clusters. On the other hand, the occurrence of MEIs creates a difference between the results of the SPIN- and PPIN-based methods. PPIN_MCODE generates nine unique clusters that have MEIs. When the SPIN is constructed,

Table 1. Summary of the prediction results from the four methods

	MCODE		LCMA	
	PPIN_MCODE	SPIN_MCODE	PPIN_LCMA	SPIN_LCMA
(a) Predicted clusters	140	171	1696	2073
(b) Proteins in clusters	620	543	3731	2967
(c) Identical clusters	131		1554	
(d) Unique clusters	9	40	142	519
(e) MEIs included	147	0	1274	0

(**a**) the number of predicted clusters. (**b**) The number of distinct proteins of the predicted clusters. (**c**) The number of clusters that are predicted by the naive- and SPIN-based method in common. (**d**) The number of unique clusters (a–c). (**e**) The number of MEIs included in clusters without distinction; all unique PPIN clusters have MEIs.

enforcing mutual exclusion eliminates some of the interactions in the network regions where nine PPIN clusters are located. As a result, SPIN_MCODE generates 40 unique clusters that were also subgraphs of the corresponding PPIN clusters. Likewise, 142 clusters of PPIN_LCMA were redefined into 519 clusters by the SPIN-based method.

3.4 Comparison with known complexes

The results were assessed using an evaluation metric used in earlier studies (Altaf-Ul-Amin *et al*., 2006; Bader and Hogue, 2003; Li *et al*., 2005, 2008) to determine how effectively a predicted cluster matches a known complex, and vice versa. Equation 1 calculates the overlapping score OS(*p, m*) between a predicted cluster $p \in P$ and a known complex $m \in M$, where P is the set of predicted clusters and M is the set of known complexes as recorded in MIPS.

$$\mathrm{OS}(p,m) = \frac{|V_p \cap V_m|^2}{|V_p| \times |V_m|} \quad (1)$$

In Equation (1), $| V_p \cap V_m |$ is the size of the intersection protein set of the predicted cluster and the known complex, $|V_p|$ is the number of proteins in the predicted cluster and $|V_m|$ is the number of proteins in the known complex. A known complex and a predicted cluster are considered as a match if their overlapping score is equal to or larger than a specific threshold. Conventionally, a predicted cluster and a known complex are considered to a match if OS(*p, m*) ≥ 0.2 (Altaf-Ul-Amin *et al*., 2006; Bader and Hogue, 2003; Li *et al*., 2005, 2008).

After all known complexes and predicted clusters have their best match calculated according to their OS scores, three evaluation criteria are applied to quantify the quality of the protein complex detection methods:

- Precision (p): measures the fraction of the predicted clusters that match the positive complexes among all predicted clusters.
- Recall (r): measures the fraction of known complexes matched by predicted clusters, divided by the total number of known complexes.
- *F1*: the *F1* score combines the precision and recall scores. It is defined as $2pr/(p+r)$.

Recall quantifies the extent to which a prediction set captures the known complexes. *Precision* measures the exactness or fidelity of the prediction set. The *F1* measure provides a reasonable combination of both precision and recall. All three values range from 0 to 1, with 1 being the best score. These three criterions are frequently used in many computational areas including protein complex detection (Qi *et al*., 2008). Here, because our reference set MIPS is incomplete, some predicted clusters which are most likely true complexes will be regarded as false positives if they do not match the current MIPS complexes well. As such, the F-measure of the algorithms should not be taken at their absolute values but only as comparative measures.

Table 2. Performance comparison between the methods based on PPIN, SPIN and random SPIN (*Ran_SPIN*)

Algorithm	Network	Recall	Precision	*F1*
MCODE	PPIN	0.213	0.314	0.254
	SPIN	0.243	0.441	0.314
	Ran_SPIN	0.199	0.358	0.255
LCMA	PPIN	0.401	0.098	0.158
	SPIN	0.528	0.128	0.207
	Ran_SPIN	0.438	0.094	0.155

The performance comparison is presented in Table 2. For each method, we report the *precision*, *recall* and *F1*, with the threshold OS $\geq$ 0.2. As can be seen, our methods based on SPIN dominate PPIN-based methods in all measures. In terms of the *F1* measures, SPIN_MCODE achieved a 23% higher value compared with the PPIN_MCODE value. When using the LCMA algorithm, SPIN_LCMA achieved a 31% higher *F1*-value compared with the PPIN_LCMA result.

In contrast with SPIN-based methods, the experiments based on random SPIN showed minor changes compared with the PPIN-based results in all three measures. This indicates that the improvements of the SPIN-based methods stem from the use of structural MEI information.

As the proposed framework aims to exclude superfluous proteins in the formation of complexes, the overlapping score of a known complex with a SPIN cluster is expected to be equal to or greater than the score of the corresponding PPIN cluster. Table 3 shows the number of known complexes matched by the clusters extracted by MCODE and LCMA from the PPIN, the SPIN and the random SPIN with respect to different thresholds. The word *loss* in parentheses refers the number of complexes that are matched by PPIN clusters

Table 3. The number of known complexes matched by predicted clusters from PPIN, SPIN and random SPIN with respect to different thresholds

	MCODE			LCMA		
	PPIN	SPIN (gain, loss)	Ran_SPIN (gain, loss)	PPIN	SPIN (gain, loss)	Ran_SPIN (gain, loss)
OS>0.0	133	133 (0, 0)	128 (0, 5)	261	261 (0, 0)	247 (0, 14)
OS≥0.1	88	91 (3, 0)	80 (6, 14)	180	216 (36, 0)	176 (18, 22)
OS≥0.2	57	65 (8, 0)	53 (4, 8)	107	141 (34, 0)	117 (27, 17)
OS≥0.3	43	51 (8, 0)	40 (1, 4)	67	96 (29, 0)	64 (6, 9)
OS≥0.4	33	38 (5, 0)	30 (0, 3)	27	47 (20, 0)	29 (6, 4)
OS≥0.5	28	32 (4, 0)	28 (0, 0)	14	30 (16, 0)	14 (0, 0)
OS≥0.6	17	23 (6, 0)	17 (0, 0)	6	18 (12, 0)	6 (0, 0)
OS≥0.7	11	16 (5, 0)	11 (0, 0)	1	8 (7, 0)	1 (0, 0)
OS≥0.8	10	14 (4, 0)	10 (0, 0)	0	5 (5, 0)	0 (0, 0)
OS≥0.9	7	9 (2, 0)	7 (0, 0)	0	3 (3, 0)	0 (0, 0)
OS=1.0	7	9 (2, 0)	7 (0, 0)	0	3 (3, 0)	0 (0, 0)

The word '*loss*' in parentheses refers the number of complexes that are matched by PPIN clusters but missed after our modification, and '*gain*' denotes the number of true positives found in addition to the result of the PPIN-based method.

but missed after our modification, and *gain* denotes the number of true positives found in addition to the result of the PPIN-based method.

The table discards the number of known complexes in the case that OS = 0, which are matched by no predicted cluster. OS > 0 indicates that the known complex has a matching predicted cluster in that it shares at least one protein. As a SPIN cluster is a subgraph of a PPIN cluster, the number of complexes matched by SPIN clusters cannot exceed that matched by PPIN clusters at the threshold of OS > 0. However, for the remaining thresholds, SPIN-based methods show better results than PPIN-based approaches.

The values of *loss* were all zero for the SPIN-based methods. This finding indicates that the results of the SPIN-based methods perfectly covered all the true positive matches from the PPIN-based methods with the thresholds listed in the table while also generating additional true positives. Unlike SPIN constructed using structural MEIs, alternating with random MEIs results in some loss of known complexes as well as additional gains. This result was possible because randomizing the MEI information may remove true and false positive proteins all together as results.

Our model may incorrectly remove true positive protein members, although it generates no loss of matched complexes. In this experiment, SPIN_MCODE removed only false positive proteins, whereas SPIN_LCMA showed two cases of protein loss as it removed two true positive members for matching with known complexes. (See MIPS complexes 410.20 and 160 in Supplementary Table 4.) However, in these cases, the known complexes had a higher OS with the SPIN cluster compared with those that used the PPIN cluster, as the SPIN framework removed many superfluous proteins. This result indicates that the proposed network model using structural MEI information can be successfully applied to graph-theoretic clustering methods for complex predictions with few faults.

3.5 The effect of the SPIN construction

The proposed network model refines a PPIN by excluding interaction competitions and it generates several subnetworks that represent possible coactive states in a process of interaction dynamics. This refinement consequently removes superfluous proteins and identifies overlapping complexes in a network clustering.

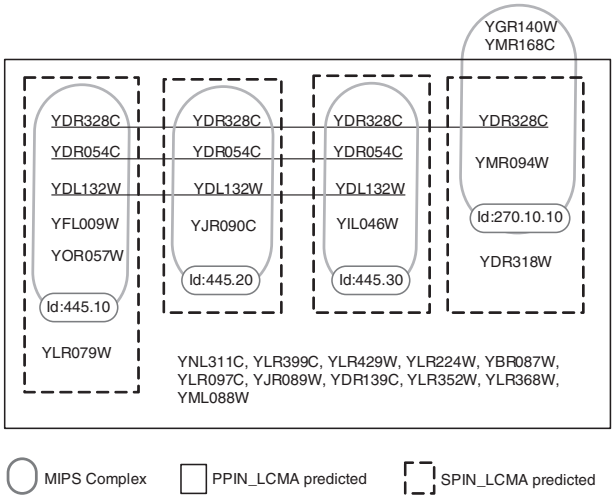

Fig. 6. Comparisons among the known complexes and clusters predicted by LCMAs based on PPIN and SPIN. The gray ovals represent known complexes from MIPS, the quadrangle is a PPIN cluster, and the dotted quadrangles are SPIN clusters. A protein that appears in several complexes is underlined.

Figure 6 is a example illustrating a refinement effect by contrasting the known complexes and clusters predicted by LCMA based on the PPIN and the SPIN. The gray ovals represent known complexes from MIPS, the quadrangle is a PPIN cluster and the dotted quadrangles are SPIN clusters. A protein that appears in several known complexes is underlined. The three complexes shown in the figure, 445.10, 445.20, 445.30, are Skp1-Cdc53-F-box protein (SCF) complexes that appear to be E3 ubiquitin-protein ligases that target a number of important regulatory proteins for ubiquitin-dependent proteolysis (Skowyra *et al*., 1997). They share three proteins YDL132W, YDR054C and YDR328C, which serve as a core, and have exchangeable adaptor subunits that specifically recruit various substrates to the core. In this research, PPIN_LCMA

could not differentiate these overlapping complexes but predicted a massive cluster matched by these three and another complex 270.10.10 with several superfluous proteins. On the other hand, given the structural interface data, our SPIN construction process specified competitions among exchangeable proteins YFL009w, YJR090c and YIL046w for the interaction with the core protein YDR328c. Consequently, for the network region in which the PPIN cluster was found, SPIN_LCMA redefined four smaller clusters that correspond to known complexes with a higher overlapping score, differentiating the varieties of SCF complexes. Like the above example, SPIN-based methods not only removed superfluous proteins but also identified variations in complex formations when additional proteins share an interface.

Supplementary Table 3 lists MIPS complexes matched by MCODEs based on PPIN and SPIN along with their overlapping scores and matched proteins, and Supplementary Table 4 lists those for LCMAs.

4 CONCLUSIONS

This study introduces a network refinement model based on the structural interface data of protein pairs for protein complex predictions. A simple PPIN, which is represented as a static entity, includes competitive interactions that cannot participate in complex formations together. In the proposed framework, a SPIN construction reserves sets of non-competitive interactions by considering mutual exclusions among the interactions in a network. This allows network-clustering algorithms to identify stable clusters that may possibly be matched by to actual protein complexes.

An evaluation of the proposed framework involved the testing of two graph-theoretic clustering algorithms on SPIN and on a simple PPIN for comparison. The comparison showed that the SPIN-based framework outperforms the plain PPIN-based method. It found all of the complexes that the PPIN-based method found as well as additional true positives by removing superfluous proteins for complex formations.

From the evaluation, it is concluded that considering MEIs is worthwhile for complex predictions. Information on mutual exclusiveness is drawn from structural interface data, which remains insufficient. This indicates that SPIN-based methods will become more useful as the additional interface data becomes available.

The authors are planning to extend the concept of SPIN so that it can represent the dynamics of complex formations and functional modules of a PPIN. Modeling the dynamics of an interaction network will lead to a better understanding of protein mechanisms.

Funding: Korea government (MEST) (Korea Science and Engineering Foundation No. 2008-0061123).

Conflict of Interest: none declared.

REFERENCES

Altaf-Ul-Amin,M. *et al.* (2006) Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, **7**, 207.

Aragues,R. *et al.* (2007) Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput. Biol.*, **3**, e178.

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

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acid Res.*, 28, 235–242.

Blatt,M. *et al.* (1997) Superparamagnetic clustering of data. *Phys. Rev. Lett.*, **176**, 3251–3254.

Boxem,M. *et al.* (2008) A protein domain-based interactome network for C. elegans early embryogenesis. *Cell*, **134**, 534–545.

Broheè,S. and Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Bryce,R.A. *et al.* (2001) Carbohydrate-protein recognition: molecular dynamics simulations and free energy analysis of oligosaccharide binding to concanavalin A. *Biophys. J.*, **81**, 1373–1388.

Colley,W.C. *et al.* (1997) Phospholipase D2, a distinct phospholipase D isoform with novel regulatory properties that provokes cytoskeletal reorganization. *Curr. Biol.*, **7**, 191–201.

Elion,E.A. *et al.* (1991) FUS3 represses CLN1 and CLN2 and in concert with KSS1 promotes signal transduction. *Proc. Natl Acad. Sci. USA*, **88**, 9392–9396.

Gong,S. *et al.* (2005) PSIbase: a database of Protein Structural Interactome Map (PSIMAP). *Bioinformatics*, **21**, 2541–2543.

Guldener,U. *et al.* (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.

Han,D. *et al.* (2004) PreSPI: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res.*, **32**, 6312–6320.

Han,J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hu,C.D. *et al.* (2005) Visualization of protein interactions in living cells using bimolecular fluorescence complementation (BiFC) analysis. *Curr. Protoc. Cell Biol.*, **26**, 21.3.1–21.3.18.

Hunter,S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D224–D228.

Kim,P.M. *et al.* (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.

King,A. *et al.* (2004) Protein complex prediction via costbased clustering. *Bioinformatics*, **20**, 3013–3020.

Li,M. *et al.* (2008) Modifying the DPClus algorithm for identifying protein complexes based on new topology structures. *BMC Bioinformatics*, **9**, 398.

Li,X. *et al.* (2005) Interaction graph mining for protein complexes using local clique merging. *Genome Inform.*, **16**, 260–269.

Liang,H. and Li,H. (2007) MicroRNA regulation of human protein-protein interaction network. *RNA*, **13**, 1402–1408.

Pei,P. and Zhang,A. (2006) Towards detecting protein complexes from protein interaction data. *LNCS*, **3992**, 734–741.

Pierrat,O.A. *et al.* (2007) Control of protein translation by phosphorylation of the mRNA 5-cap-binding complex. *Biochem. Soc. Trans.*, **35**, 1634–1637.

Qi,M. and Elion,E.A. (2005) MAP kinase pathways. *J. Cell Sci.*, **118**, 3569–3572.

Qi,Y. *et al.* (2008) Protein complex identification by supervised graph local clustering. *Bioinformatics*, **24**, i250–i268.

Skowyra,D. *et al.* (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell*, **91**, 149–151.

Spirin,V. and Mirny,L. (2008) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

Tabuchi,K. *et al.* (2002) CASK Participates in Alternative Tripartite Complexes in which Mint 1 Competes for Binding with Caskin 1, a Novel CASK-Binding Protein. *J. Neurosci.*, **22**, 4264–4273.

Tong,A. and Drees,B. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

Valente,A.X.C.N. *et al.* (2009) Functional organization of the yeast proteome by a yeast interactome map. *Proc. Natl Acad. Sci. USA*, **106**, 1490–1495.

Winter,C. *et al.* (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.

Zhang,S.H. *et al.* (2006) Prediction of protein complexes based on protein iteraction data and functional annotation data using kernel methods. *LNBI*, **4115**, 514–524.