USING INDIRECT PROTEIN-PROTEIN INTERACTIONS FOR PROTEIN COMPLEX PREDICTION

Hon Nian Chua[1], Kang Ning[2], Wing-Kin Sung[2], Hon Wai Leong[2] and Limsoon Wong[2]

[1]*Graduate School of Integrated Sciences and* [2]*Department of Computer Science, National University of Singapore*

g0306417@nus.edu.sg, {ningkang, ksung, leonghw, wongls}@comp.nus.edu.sg

Protein complexes are fundamental for understanding principles of cellular organizations. Accurate and fast protein complex prediction from the PPI networks of increasing sizes can serve as a guide for biological experiments to discover novel protein complexes. However, protein complex prediction from PPI networks is a hard problem, especially in situations where the PPI network is noisy. We know from previous work that proteins that do not interact, but share interaction partners (level-2 neighbors) often share biological functions. The strength of functional association can be estimated using a topological weight, FS-Weight. Here we study the use of indirect interactions between level-2 neighbors (level-2 interactions) for protein complex prediction. All direct and indirect interactions are first weighted using topological weight (FS-Weight). Interactions with low weight are removed from the network, while level-2 interactions with high weight are introduced into the interaction network. Existing clustering algorithms can then be applied on this modified network. We also propose a novel algorithm that searches for cliques in the modified network, and merge cliques to form clusters using a "partial clique merging" method. In this paper, we show that 1) the use of indirect interactions and topological weight to augment protein-protein interactions can be used to improve the precision of clusters predicted by various existing clustering algorithms; 2) our complex finding algorithm performs very well on interaction networks modified in this way. Since no any other information except the original PPI network is used, our approach would be very useful for protein complex prediction, especially for prediction of novel protein complexes.

Keywords: protein-protein interaction, protein complex prediction, level-2 interaction, partial clique merging

1 INTRODUCTION

Identification of functional modules in protein interactions network is a first step in understanding the organization and dynamics of cell functions. Protein-protein interaction networks (PPIs) are rapidly becoming larger and more complete as research on proteomics and systems biology proliferates [1]. As a result, more protein complexes are been identified [2]. A protein complex is a group of two or more associated proteins. Protein complex is a form of quaternary structure. Similar to phosphorylation, complex formation often serves to activate or inhibit one or more of the associated proteins. Many protein complexes are established, particularly in the model organism *Saccharomyces cerevisiae* (Bakers' yeast). With a wealth of and constantly increasing size of PPI datasets, efficient and accurate intelligent tools for identification of protein complexes are of great importance. In this paper, we have focused on predicting protein complexes from PPI data.

Currently, there are several approaches to the protein complex prediction problem [3-8]. Spirin et. al. [3] proposed using clique finding and super-paramagnetic clustering with Monte Carlo optimization to find clusters of proteins. They found a significant number of protein complexes that overlap with experimentally derived ones. While clique finding [3] imposes stringent search criterion, and generally results in greater precision, recall is limited because: 1) protein interaction networks are incomplete; and 2) protein complexes may not necessary be complete subgraphs. Another approach, such as MCODE [5], are clustering based. MCODE makes use of local graph density to find protein complex. PPI networks are transformed to weighted graphs in which vertices are proteins and edges represent protein interactions. The algorithm operates in three stages: vertex weighting, complex prediction and optimal post-processing. Each stage involves several parameters that can be fine-tuned to get better predictions. However, clustering approaches [5, 8] yield good recall but sacrifice precision. To make clustering based approaches more viable, [4, 7] show that it is possible to identify high precision subsets of clusters from clustering results by post-processing based on functional homogeneity, cluster size and interaction density. While post processing significantly improves precision, recall is drastically reduced. Moreover, the approach makes use of functional information, which limits its applicability in less studied genomes such as *Homo sapiens*, *Mus muculus* and *Arabidopsis thialiana*. Recently, a popular clustering algorithm, Markov

clustering algorithm (MCL) [9], has also been shown to perform well in an evaluation of algorithms for protein clustering in PPI networks [6]. MCL partitions the graph by discriminating strong and weak flow in the graph, which is shown to be very robust against graph alternations. Table 1 gives the main features of the algorithms that we have used for comparison in this paper.

Table 1. Main features of protein complex prediction algorithms.

	RNSC	**MCODE**	**MCL**
Type	Local search cost based	Local neighbourhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

We know from [10] that many proteins that do not interact, but share common interaction partners, share functions and participate in similar pathways. The interactions between these proteins are referred to as "level-2 neighbors". [10] also proposed a topological weight, FS-Weight for estimating functional association between direct and indirect interactions, which is shown to work well. In this paper, we propose using these indirect interactions with FS-Weight to modify the existing PPI as a preprocessing step to complex prediction. The original PPI network is expanded by including indirect interactions (relationship between pairs of proteins that do not interact but share common interactors). A topological weight, *FS-Weight* (functional similarity weight), is then computed for both direct and indirect interactions. Interactions with weights below a threshold are removed. We also propose a new algorithm that incorporates FS-Weight for complex prediction. The algorithm employs clique finding on a modified PPI network, retaining the benefits of clique based approaches while improving recall. The algorithm first searches for cliques in the modified network, and iteratively merges them by "partial clique merging" to form larger clusters. For the rest of this paper, we refer to predicted protein clusters as *clusters*, and known protein complexes as *complexes*.

2 INTRODUCTION OF INDIRECT NEIGHBORS

The PPI network is transformed into a graph G=(V, E). Each vertex $v_k \in V$ represents a protein, while each edge $\{v_i, v_j\} \in E$ represents an interaction between the proteins v_i and v_j. For the rest of this section, we consider PPI networks in this graph-based representation. We refer to level-1 interactions as the original interactions in the PPI network, and level-2 interaction as an indirect interaction between two proteins which do not interact, but share common interaction interactors.

Members in a real complex may not have physical interactions with all other members; hence conventional methods (clique-based, density-based) may miss the detection of many members. By introducing level-2 interactions, which represent strong functional relations (from [10]), we will be able to capture members with less physical involvement in the complex.

[10] showed that a topological weight, the FS-Weight, can identify both level-1 and level-2 interactions that are likely to share common functions within the local (level-1 and level-2) PPI interaction neighborhood. Since proteins within a complex interact to perform a common function, it makes sense to identify protein complexes using FS-weight. Through topological weighting, we can identify interactions reasonably with a good likelihood of indicating functional relationship, and use these for complex prediction. This will also reduce the impact of noise and make predictions more robust.

Topological Weighting

All level-1 and level-2 interactions in the PPI network are given a weight using the topological weight, FS-Weight, defined as follows:

$$S_{FS}(u,v) = \frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_u} r_{u,w} + \sum_{w\in(N_u\cap N_v)} r_{u,w}\left(1-r_{v,w}\right)\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}+\lambda_{u,v}} \times \frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_v} r_{v,w} + \sum_{w\in(N_u\cap N_v)} r_{v,w}\left(1-r_{u,w}\right)\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}+\lambda_{v,u}} \quad (1)$$

N_p refers to the set that contains protein p and its level-1 neighbors; $r_{u,w}$ refers to the estimated reliability of the interaction between u and w. In [10], $r_{u,w}$ is estimated based on annotated proteins in the training set during cross validation. To avoid possible bias that may be caused by using additional information (functional annotation), we exclude reliability estimation of interactions and set all $r_{u,w}$ to 1. $\lambda_{u,v}$ is a pseudo-count included in the computation to penalize similarity weights between protein pairs when proteins has very few level-1 neighbors, and is defined as:

$$\lambda_{u,v} = \max(0, n_{avg} - (|N_u - N_v| + |N_u \cap N_v|)) \qquad (2)$$

in which n_{avg} is the average number of neighbors per protein in the PPI network.

Using FS-Weight, we modify an existing protein-protein interaction network in the following manner: 1) Level-1 interactions in the network that have low FS-Weights (weight below a certain threshold, FS-$Weight_{min}$) are removed from the PPI network. 2) Level-2 interactions that have high FS-Weights (above or equal to FS-$Weight_{min}$) are added into the PPI network. FS-$Weight_{min}$ is a value that is determined empirically.

3 PCP ALGORITHM

After we have generated a modified PPI network, existing protein complex prediction algorithms can be applied on it for more reliable protein complex prediction. However, we have also designed a novel algorithm, ProteinComplexPrediction (PCP), for complex prediction using "partial clique merging". This method differs from existing approaches in the following ways: 1) it uses the FS-Weight information during the merging of cliques (clusters); 2) merging based on cliques is a clear and rigid method in graph theory and it is more viable based on reliable PPI networks. PCP attempts to achieve the high precision of clique-finding algorithms whilst providing greater recall and computational tractability, without using any external information. Results show that this method performs well and is robust against noises.

Maximal Clique Finding

We first find all maximal cliques within the modified PPI. To do this, we implement the maximal clique finding algorithm described in [11]. This algorithm has been shown to be very efficient on sparse graphs. All cliques of at least size 2 is reported. To make sure that there is no overlap among cliques, any overlap between cliques can only be assign to one clique. There can be many ways to do this. Since FS-Weight is an estimate for the likelihood of sharing functions, a cluster with a larger average FS-Weight would more likely represent a subset of a real complex. We define the Average FS-Weight of a subgraph S with edges E_s is defined as:

$$FS_{avg}(S) = \frac{\sum_{(u,v)\in E_s} FS(u,v)}{|E_s|} \qquad (3)$$

Ideally, we want to find the best way to remove overlaps so that the total average FS_{avg} of all the final non-overlapping cliques is maximized. However, since this is a NP-hard problem, we turn to heuristics. All cliques are first sorted by decreasing FS_{avg}. The clique with the highest FS_{avg} is selected and compared with the rest of the cliques. Whenever an overlap is found with another clique, the overlapping nodes are assigned to one of the two cliques such that the two cliques have a higher average FS_{avg}. An example is given in Fig 1 (b).

InterClusterDensity

A protein complex is likely to consist of proteins forming a dense network of interactions, but may not necessarily form a complete clique. Due to the stringent definition of a clique, the resulting maximal cliques from the clique finding step are relatively small and are likely to be partial representations of real complexes. To reconcile these smaller protein clusters into larger clusters that form fuller representation of real complexes, we previously tried to merge overlapping clusters based on the amount of overlapping vertices between them. However, the corresponding prediction results are not good, since each merge considers only overlapping vertices between two clusters, but overlooks the density of interactions between them. Hence we define Inter-Cluster Density (ICD), which is a measure of interconnectedness between two subgraphs, as a criterion for merging clusters. The ICD essentially computes the FS-Weight density of inter-cluster interactions between the non-overlapping proteins of two clusters. High ICD indicates that the two clusters are highly connected. Using ICD to impose criteria for

merging ensures that merged clusters retain a certain degree of interconnectedness between its members. The Inter-Cluster Density (ICD) between subgraphs S_a and S_b is defined as:

$$ICD(S_a,S_b)=\frac{\sum S_{FS}(i,j)\mid i\in(V_a-V_b),j\in(V_b-V_a),(i,j)\in E}{|V_a-V_b|\cdot|V_b-V_a|} \quad (4)$$

where V_x is the set of vertices of subgraph S_x. An example of ICD computation is given in Fig 1 (a).

Partial Clique Merging

To merge cliques found in the PPI network, we define the term "partial cliques" as strongly connected subgraphs formed from the amalgamation of one or more cliques. Trivially, all cliques in the PPI network G are partial cliques. We begin with an initial graph G_p^0 in which each vertex represents a partial clique, and add an edge (*u*, *v*) between any pair of partial cliques u and v in G_p^0 if $ICD(u,v)\geq ICD_{thres}$. From G_p^0, we can again find maximal cliques among the vertices. Each clique in G_p^0 is therefore a cluster of partial cliques from G, where all pairs of partial cliques in the cluster fulfils a minimum level of interconnectedness defined by ICD. In other words, the vertices in each clique from G_p^0 can be merged to form a larger *partial clique*.

This process is then repeated to form bigger partial cliques. In each iteration *i*, a graph G_p^i is formed from PC^{i-1}, the partial cliques from the previous iteration, i.e. $G_p^i = (PC^{i-1}, \{(u,v) \mid ICD(u,v)\geq ICD_{thres}, u,v\in PC^{i-1}\})$. From G_p^i, we can again find maximal cliques among the vertices (partial cliques in G_p^{i-1}) and merge the proteins in these cliques to form bigger partial cliques. This is done until no further merge can be made. In order for the more connected partial cliques to merge first, we first perform the merge using $ICD_{thres} = 1$. The merging process is then repeatedly reinitiated while reducing ICD_{thres} by 0.1 until $ICD_{thres} \leq ICD_{min}$. ICD_{min} is a threshold to be determined empirically. A smaller ICD_{min} will yield bigger clusters and vice versa. We refer to this merging method as "partial clique merging".

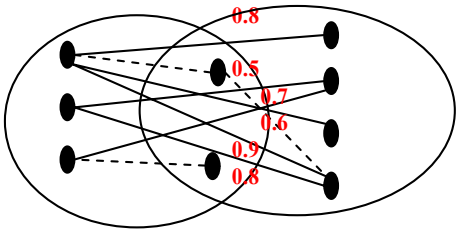
4 EXPERIMENTS

Experiment Settings and Datasets

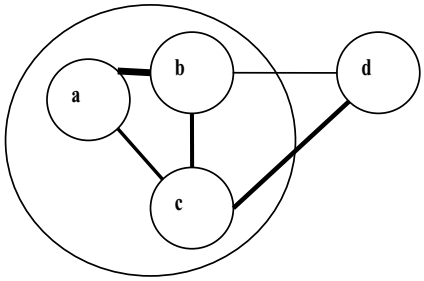
The PCP algorithm is implemented in C++ and Perl. We compare PCP with state-of-the-art algorithms: RNSC [4], MCODE [5] and MCL [6] algorithms. The experiments are performed on a PC with 3.0 GHz CPU and 1.0 GB RAM, running a Linux system.

- **PPI datasets**

We use two high-throughput datasets obtained from different sources for analysis of these algorithms. The first dataset is obtained from the GRID database [12]. This dataset is a combination of six protein interaction networks from the *Saccharomyces cerevisiae* (Bakers'



ICD(S_a, S_b)
=(0.8+0.5+0.7+0.6+0.9+0.8)/(3*4)=0.36
(a)



$$FS_{Avg}(\{a,b,c\})+FS_{Avg}(\{d\})>FS_{Avg}(\{a\})+FS_{Avg}(\{b,c,d\})$$

Merge({a,b,c},{b,c,d}) = {a,b,c},{d}
(b)

Fig 1. (a) Example of ICD computation. There are two clusters, and solid lines are used for ICD calculation. (b) Example of resolving overlapping cliques. Edge thickness represents the FS-Weight of the edge.

Yeast) genome. These includes interactions characterized by mass spectrometry technique from Ho *et al.*[13], Gavin *et al.*[14], Gavin *et al.* [15] and Krogan *et al.* [16], as well as two-hybrid interactions from Uetz *et al.* [1] and Ito *et al.* [17]. We shall refer to this dataset as PPI[Combined]. The second dataset is taken from a current release of the BioGRID database [18]. We only consider interactions derived from mass spectrometry and two-hybrid experiments since these represents physical interactions. We shall refer to this dataset as PPI[BioGRID]. Table 3 presents the features of the two datasets, as well as some characteristics of the clusters predicted by different algorithms.

- **Protein Complex datasets**

As a yardstick for prediction performance, we use protein complex data from the MIPS database [2]. These protein complexes are treated as a golden standard for analysis.

To examine whether false positives in predictions may turn out to be undiscovered annotations, we use two releases of the MIPS complex datasets - a dataset released on 03/30/2004 and a newer dataset released on 05/18/2006. We refer to two protein complex datasets as PC_{2004} and PC_{2006}, respectively. During validation, proteins that cannot be found in the input interaction network are removed from the complex data.

- **Cluster Scoring**

Density of a graph G = (V,E) is defined as $D_G = |E|/|E|_{max}$, where for a graph with loops and $|E|_{max} = |V|\ (|V|+1)/2$ and for a graph with no loops, $|E|_{max} = |V|\ (|V|-1)/2$. So, D_G is a real number ranging from 0.0 to 1.0. Resulting cluster S = (V,E) from the algorithm are scored and ranked by *cluster score*, which is defined as the product of the density and the number of vertices in S, ($D_C \times |V|$). This ranks larger more dense clusters higher in the results.

- **Validation Criterion**

In order to study the relative performance of PCP against existing algorithms, we need to define the criterion that determines whether a predicted protein cluster matches a true protein complex. [5] defined a matching criterion using the overlap between a protein cluster S and a true protein complex C:

$$Overlap(S,C) = \frac{|V_S \cap V_C|^2}{|V_S| \cdot |V_C|} \qquad (5)$$

V_s are the vertices of the subgraph defined by S; and V_c are the vertices of the subgraph defined by C.

In [5], an overlap threshold of 0.2 is used to determine a match. [4] used a modified version of the overlap which is more stringent but involves many empirically derived parameters which may not be applicable across different datasets. To simplify comparison, we used an overlap threshold of 0.25 to determine a match for all experiments in this work. Predicted protein clusters that match one or more true protein complexes with overlap score above this threshold are identified as "matched predicted complexes", and the corresponding complexes are identified as "matched known complexes". Note that the number of "matched clusters", $matched_{cluster}$, may differ from the number of "matched complex", $macthed_{complex}$ because one known complex can match one or more predicted clusters.

To measure the accuracies of prediction, the analysis on the Precision and Recall, of different algorithms are computed. Precision and Recall are defined as

$$\text{Precision} = \frac{matched_{clusters}}{predicted_{clusters}} \qquad (6)$$

$$\text{Recall} = \frac{matched_{complexes}}{known_{complexes}} \qquad (7)$$

where $predicted_{clusters}$ and $known_{complexes}$ are the number of predicted clusters and the number of known (real) complexes, respectively.

The recall measure in our validation is determined by matched complexes instead of predicted clusters, and is hence not prone to bias. Moreover, the precision measure uses the number of predicted clusters as a denominator. Hence there should not be any significant bias in these validation measures. We only consider clusters and complexes of size 4 and above, since matches between clusters and complexes of smaller sizes have relatively high probabilities of occurring by chance [4]. Note that unlike the validation measures used in [6], we do not seek to evaluate the clustering properties of each algorithm. Rather, we are concerned about the actual usefulness of the algorithms in detecting clusters that match real complexes reasonably well.

To avoid bias that may arise from large variations in the size of predicted complexes, we also introduce another precision-recall analysis based on protein membership assignment. For this analysis, we defined two terms: protein-cluster pair (*PCl*) and protein-complex pair (*PCo*). Each *PCl* represents an unique protein-cluster relationship. For example, given two predicted clusters $Cl(A) = \{P_1, P_2\}$ and $Cl(B) = \{P_1, P_3\}$, we have four *PCl*s, namely $(Cl(A), P_1)$, $(Cl(A), P_2)$, $(Cl(B), P_1)$ and $(Cl(B), P_3)$. Similarly, each *PCo* represents an unique protein-complex relationship.

$\text{Precision}_{\text{protein}}$: A *PCl* is considered to be *matched* if its protein belongs to some complex that matches its cluster. The definition of a match between a predicted cluster and a complex is described earlier in this section. $\text{Precision}_{\text{protein}}$ is defined as:

$$\text{Precision}_{\text{protein}} = \frac{|\, matched_{PCl} \,|}{|\, predicted_{PCl} \,|} \qquad (8)$$

$\text{Recall}_{\text{protein}}$: A *PCo* is considered to be *matched* if its protein belongs to some cluster that matches its complex. $\text{Recall}_{\text{protein}}$ is defined as:

$$\text{Recall}_{\text{protein}} = \frac{|\, matched_{PCo} \,|}{|\, known_{PCo} \,|} \qquad (9)$$

Results

- **Parameters determination**

The optimal parameters for RNSC, MCODE and MCL algorithms are given by [6] (Table 2).

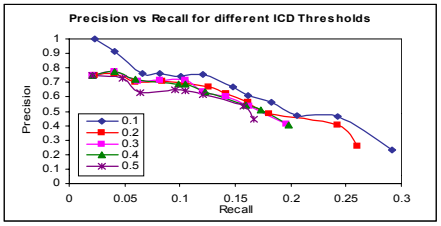
Table 2. Optimal parameters for RNSC, MCODE and MCL algorithms.

Algorithm	Parameter	Optimal value
RNSC	No. of experiments	3
	Tabu length	50
	Scaled stopping tolerance	15
MCODE	Depth	100
	Node score %	0
	Haircut	True
	Fluff	False
	% of complex fluffing	0.2
MCL	Inflation	1.8

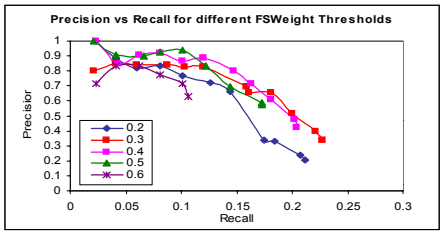
There are two tunable parameters in our experiments: $\text{FS-Weight}_{\text{min}}$, and ICD_{min}. $\text{FS-Weight}_{\text{min}}$ determines the FS-Weight (1) threshold for filtering out level-1 and level-2 interactions. ICD_{min} determines the Inter-Cluster Density (4) threshold for which two clusters are allowed to merge during clustering for the PCP algorithm. Based on PPI[Combined] and PC_{2004}, we use level-1 interactions (without any filtering) to determine ICD threshold. FS-Weight threshold is determined on the same dataset using PCP algorithm.

Inter-Cluster Density Threshold: We first vary ICD_{min}, the Inter-Cluster Density threshold for merging clusters between 0.1 and 0.5 and perform the predictions. The corresponding precision and recall of the predictions are shown in Fig 2 (a). Lower ICD_{min} results in more clusters being merged and vice versa. We find that ICD_{min}=0.1 yields the best precision against recall and use this for the rest of our experiments.

FS-Weight Threshold: [10] showed that filtering level-1 and level-2 interactions with a FS-Weight threshold of 0.2 resulted in interactions that have a significantly higher likelihood of sharing functions. Here we perform protein complex prediction using the PCP algorithm with a range of $\text{FS-Weight}_{\text{min}}$ to determine which value can yield the best prediction performance. The ICD_{min} is set to 0.1. The corresponding precision and recall of the predictions are shown in Fig 2 (b). We find



(a)



(b)

Fig 2. Effect of (a) ICD threshold and (b) FS-Weight threshold on Precision and Recall values for PPI[Combined] dataset.

that FS-Weight$_{min}$=0.4 yields the best precision against recall, and use this for the rest of our experiments.

- **Introduction of indirect neighbors**

The introduction of indirect neighbors is the key part of our analysis in this paper. To evaluate the performance this process, we transform the original PPI network in three different ways: (1) All level-1 interactions; (2) All level-1 and level-2 interactions; (3) All level-1 interactions, and level-2 interactions with FS-Weight ≥ FS-Weight$_{min}$; and (4) level-1 and level-2 interactions with FS-Weight ≥ FS-Weight$_{min}$. For (2), Due to the large number of level-2 interactions, results can only be obtained for MCL and RNSC. For example, on PPI[combined], there are 20,461 level-1interactions. With the introduction of level-2 interactions, the number of interactions increased to 404,511. After filtering level-2 interactions based on FS-Weight, we have 23,356 interactions. Finally, upon filtering both level-1 and level-2 interactions, we are left with only 7303 interactions.

If two proteins in an interaction belong to some common known complex, we defined the interaction as an intra-complex interaction. To justify our intuition for using level-2 interactions and FS-Weight for complex prediction, we compute the fraction of interactions in the 4 transformed networks that are intra-complex interactions. Since proteins are clustered based on interactions, a higher fraction of intra-complex interactions will naturally yield more accurately predicted clusters. In Fig 3, we present the corresponding fractions for two PPI networks, PPI[Combined] and PPI[BioGRID] using the known protein complexes in PC$_{2004}$.. We observe that the fraction of intra-complex interactions did not change significantly after adding filtered level-2 interactions into the network. However, if both level-1 and level-2 interactions are filtered, the fraction of intra-complex interactions become significantly higher. Without any filtering, level-2 interactions will contain too many false positives to be useful, as reflected by the very small fraction of intra-complex interactions. This is consistent with the findings for function similarity in [10]. From the observations, we believe that using a PPI network with filtered level-1 and level-2 interactions would yield the best results for protein complex prediction.

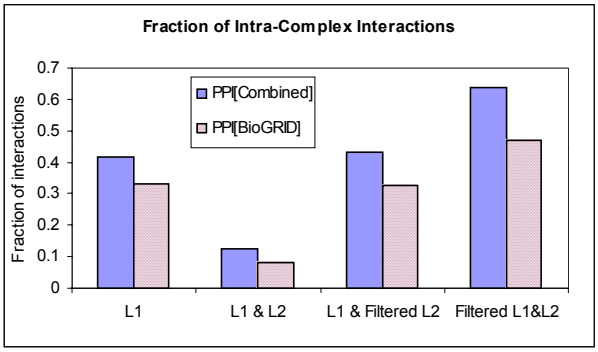


Fig 3. Fraction of intra-complex interactions with nodes sharing some complex membership for different PPI networks.

- **Comparison with existing approaches**

We compared clusters predicted using four clustering algorithms: MCL, RNSC, MCODE and PCP on the two datasets PPI[Combined] and PPI[BioGRID]. PC$_{2004}$ is used to represent real protein complex against which the results from these algorithms are validated.

Table 3 summarizes some general characteristics of clusters predicted by four clustering algorithms. The PPI[BioGRID] dataset is larger than PPI[Combined]. We observe that upon the introduction of filtered level-2 interactions, the number of predicted clusters generally decrease while average cluster sizes increase. This is due to greater connectivity in the graph since more edges are added among the same number of nodes. We also observe that the average cluster sizes of clusters predicted by the MCODE and MCL algorithms are larger than those predicted by the RNSC and PCP algorithms. After filtering both level-1 and level-2 interactions using FS-Weight, all algorithms produced less clusters. With the exception of MCODE, the average cluster sizes of clusters predicted by the various algorithms are also larger.

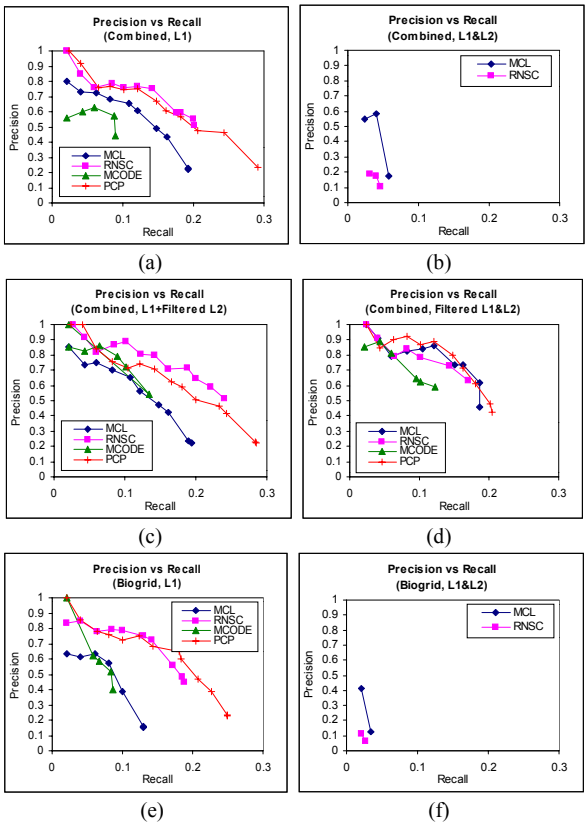
We have also studied the average density of the clusters predicted by the four different algorithms using the different networks. Generally, all algorithms predicted clusters with the highest density using only level-1 interactions, followed by using level-1 and filtered level-2 interactions. Using filtered level-1 and level-2 interactions resulted in clusters of lower density. When level-1 and level-2 interactions without filtering are used, the clusters found have the lowest density. RNSC yielded clusters

Table 3. The features of the datasets, and the features of the clusters that are predicted by different algorithms.

Datasets	Nodes	Edges	No. Complex	Avg. Complex Size	Setting	No. of Clusters				Avg. Cluster Size			
						RNSC	MCODE	MCL	PCP	RNSC	MCODE	MCL	PCP
PPI[Combined]	4672	20461	**815**	*8.80*	1)	**2332**	**121**	**936**	**1537**	*2.00*	*5.75*	*4.99*	*3.04*
					2)	**874**	**-**	**209**	**-**	*5.34*	*-*	*22.35*	*-*
					3)	**2233**	**120**	**720**	**1499**	*2.09*	*6.48*	*6.49*	*3.12*
					4)	**699**	**92**	**259**	**417**	*2.44*	*5.83*	*6.59*	*4.09*
PPI[BioGRID]	5036	27560	**815**	*8.82*	1)	**2404**	**152**	**830**	**1764**	*2.20*	*3.98*	*6.38*	*2.85*
					2)	**811**	**-**	**159**	**-**	*6.21*	*-*	*31.67*	*-*
					3)	**2331**	**142**	**681**	**1557**	*2.16*	*5.69*	*7.40*	*3.23*
					4)	**901**	**121**	**285**	**555**	*2.36*	*5.51*	*7.46*	*3.83*

with the highest density, followed by MCODE, PCP and MCL. Interestingly, we found that the average density of real protein complexes is quite low, around 0.55, which suggests that the density of predicted clusters do not correlate with prediction accuracy.

Fig 4 presents the precision-recall analysis of the predictions made by the four algorithms. By varying a threshold on cluster score, we can obtain a range of recall and precision for the predictions from each algorithm.



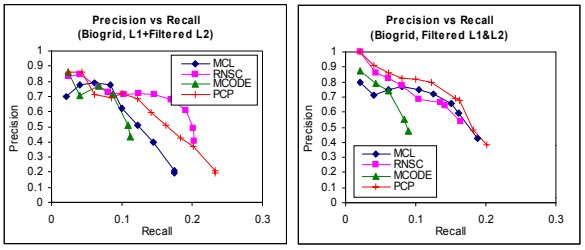


Fig 4. The precisions and recalls of RNSC, MCODE, MCL and PCP algorithms on PPI[Combined] with (a) original level-1 interactions, (b) level-1 and level-2 interactions, (c) original level-1 and filtered level-2 interactions, and (d) filtered level-1 and level-2 interactions; PPI[BioGRID] with (e) original level-1 interactions, (f) level-1 and level-2 interactions, (g) original level-1 and filtered level-2 interactions, and (h) filtered level-1 and level-2 interactions. Results are based on comparison with PC_{2004} protein complex dataset.

From Fig 4 (a)-(d) on the PPI[Combined] dataset, we observed that RNSC performs the best in precision and recall on the original network (level-1 interactions). With the introduction of lelvel-2 interactions, the precision and recall decreased. When these level-2 interactions are filtered, precision and recall are improved in MCODE and RNSC, while PCP and MCL remain almost unchanged. However, when filtered level-1 and level-2 interactions are used, all methods show significant improvement in precision except RNSC. In all the combinations, PCP with filtered level-1 and level-2 interactions performs the best (Fig 4 (d)). A similar trend is observed in the bigger PPI[BioGRID] dataset (Fig 4 (e)-(h)). Precision is improved in most algorithms with the introduction of filtered level-2 neighbors, and further improvement is achieved when level-1 interactions are also filtered based on FS-Weight. In particular, the performance of MCODE and MCL improved substantially with the introduction of level-2 interactions and FS-Weight filtering. Again, PCP

with filtered level-1 and level-2 interactions performs the best (Fig 4 (h)).

To illustrate the contribution of PCP to complex prediction, we compare predictions made by each algorithm natively (i.e. RNSC, MCODE, MCL on original level-1 interactions against PCP on filtered level-1 and level-2 interactions) in Fig 5. We observe that PCP outperforms the other algorithms significantly (Fig 5 (a) and (b)). We arrived at similar conclusions using precision-recall analysis based on protein membership assignment (Fig 5 (c) and (d)).

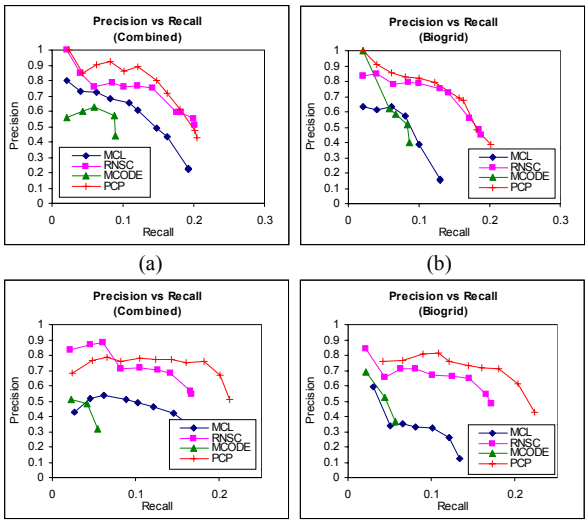


Fig 5. Precisions-recall analysis of RNSC, MCODE, MCL and PCP algorithms on (a) PPI[Combined] and (b) PPI[BioGRID] using native settings (RNSC, MCODE, MCL on original level-1 interactions, and PCP on filtered level-1 and level-2 interactions); Precision-recall analysis based on protein membership assignment on the same predictions on (c) PPI[Combined] and (d) PPI[BioGRID]. Results are based on comparison with PC_{2004} protein complex dataset.

Examples of predicted complexes: We have proposed two new concepts in this paper: the introduction of indirect interactions as a preprocessing step, and the PCP clustering algorithm. To illustrate how these concepts can help to predict protein clusters that better match real complexes, we examine some examples of protein clusters predicted by the PCP based on the modified network, as well as RNSC and MCL algorithms based on the original network, and how they correspond to real protein complexes in the PC_{2004} dataset. Fig 6 shows two examples where PCP can predict protein clusters that match a real complex more precisely than other algorithms. In the first example (Fig 6 (a)), PCP predicted a cluster that matches a 4-member protein complex completely, while RNSC's 3-member cluster has only one member, "YDR121W", that matches the same complex.

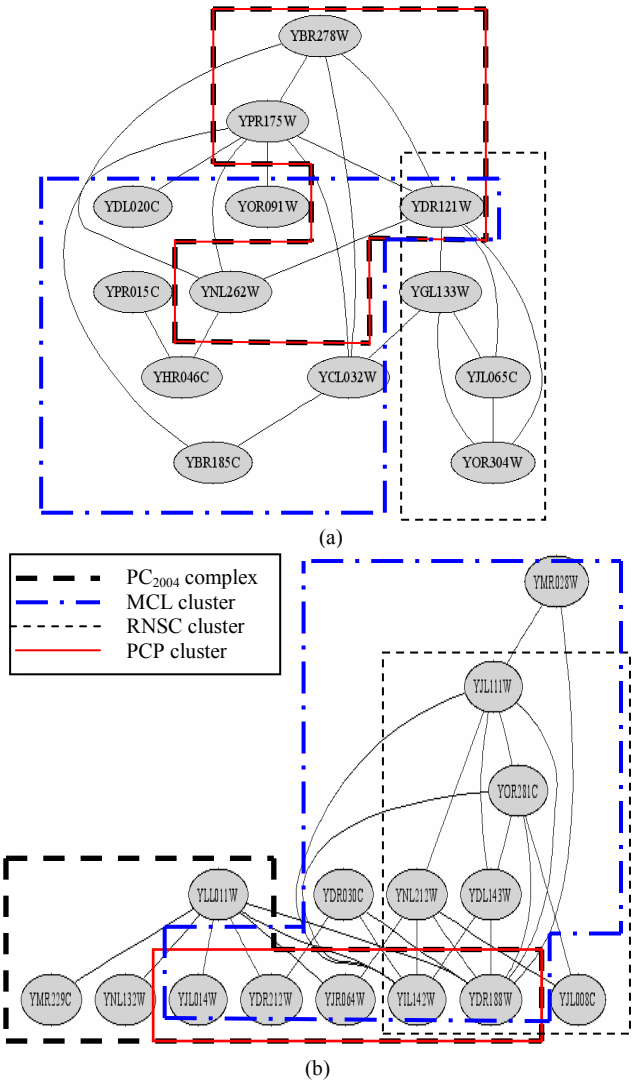
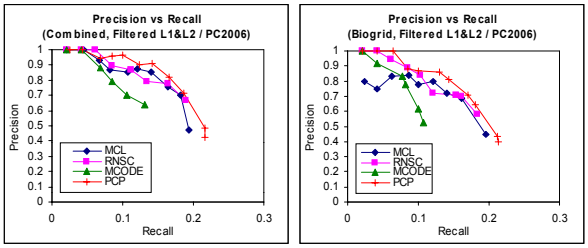


Fig 6. Example of predicted and matched complexes. Complexes in PC_{2004}, the predicted clusters by MCL, RNSC and PCP are shown in different boxes. (a) A complex in PC_{2004} of size 4, PCP's cluster matched it perfectly, while MCL and RNSC's clusters matched 1 and 2 of the proteins in the complex, respectively. (b) In this complex in PC_{2004} of size 8, RNSC's predicted cluster matched only 2 proteins, while PCP's predicted cluster matched 5 proteins, MCL also matched 5 proteins, but predicted 6 proteins that are not in the complex.

This is probably due the fact that members in RNSC's cluster are well connected by level-1 interaction. But by including level-2 interactions and filtering unreliable interactions, their connections are shown not to be strong enough to be in one cluster. Therefore PCP is able to identify the correct complex. Similarly, the cluster predicted by MCL only overlaps with two members of the complex, while the other 6 members of the cluster do not belong to the real complex. The second example (Fig 6 (b)) shows a 5-member protein cluster predicted by PCP, which is a subset of a 8-member protein complex. The best match with the same complex from RNSC is a 7-member cluster, in which only 2 belongs to a subset of the real complex. Though PCP's predicted cluster matched 5 proteins and MCL also matched 5 proteins, but the latter predicted 6 proteins that are not in the complex. A closer look will reveal that PCP's cluster member do not have any interactions among them, and this subset of the real protein complex can only be identified by level-2 interactions with the rest of the complex members. PCP is unable to discover the rest of the complex as their connectivity with the other members is very weak or unknown. The protein "YLL011W" is missed by PCP because its local topology resulted in a low FS-Weight score. This may be due to the reason that "hub proteins" like "YLL011W" are automatically penalized by the FS-Weight score.

- **Validation on newer protein complex data**

A comparison of prediction performance validated against an old protein complex dataset and a newer, more updated standard protein complex dataset can reveal the parameter-independent identification power of the different algorithms. We have previously assessed the RNSC, MCODE, MCL and PCP algorithms with PC_{2004}. Here, we validate the predicted clusters of PCP and other algorithms against a more recent and more updated protein complex dataset, PC_{2006}. We have used modified PPI networks (PPI[Combined] and PPI[BioGRID]) with filtered level-1 and level-2 interactions which have the shown earlier (Fig 4) to yield the best performance for most algorithms studied. The corresponding precision-versus-recall graphs are shown in Fig 7. Comparing Fig 4 against Fig 7, we find that against the same recall range, the precision of all algorithms studied has increased substantially when validating against PC_{2006} for both PPI network datasets. A significant number of clusters which are predicted by PCP, but have been treated as false positives because they cannot be matched against any known complex in PC_{2004}, are now found to match against known complexes in PC_{2006}. This indicates that PCP has a good potential for finding novel protein complexes.



(a) (b)

Fig 7. The precisions and recalls of different algorithms on (a) PPI[Combined] and (b) PPI[BioGRID] with filtered level-1 and level-2 interactions. Results are based on comparison with PC_{2006} protein complex dataset.

We also present two illustrative examples in Fig 8 which show that PCP predicted novel members to some complexes, which are later verified in the newer complex dataset. In the first example (Fig 8 (a)), PCP predicted a cluster of 4 proteins. The cluster is found to match well with a real 4-member complex from PC_{2004} that contains all but 1 of the proteins in the predicted cluster. A comparison with PC_{2006}, however, reveals that the predicted cluster matched a real complex in the dataset that contains all the 4 proteins. The protein "YFL008W" in PC_{2006} has level-1 interactions with the other 3 proteins, but since the FS-Weight of these interactions are low, PCP did not predict it to be in the same cluster. It is also interesting that in Fig 8 (b), PCP has predicted "YHR033W" to be in the same cluster as the other 5 proteins, and this is consistent with PC_{2006} but not PC_{2004}. However, the other 5 proteins in the new complex are not predicted by PCP, since they do not have any level-1 interaction with other proteins. We think that more accurate prediction of this protein complex may be achieved by incorporating additional information such as function annotations. Moreover, while "YJR072C" protein is predicted by PCP, it is not in new protein complex. Since the interactions of this protein with "YDR212W" and "YJR064W" are present in quite a few other protein

complexes [8], we believe that even though this protein is not in the same complex with other proteins, it should be in the same "function unit" [3] with these proteins. Discriminating "function unit" with protein complex may need additional information such as function annotations.

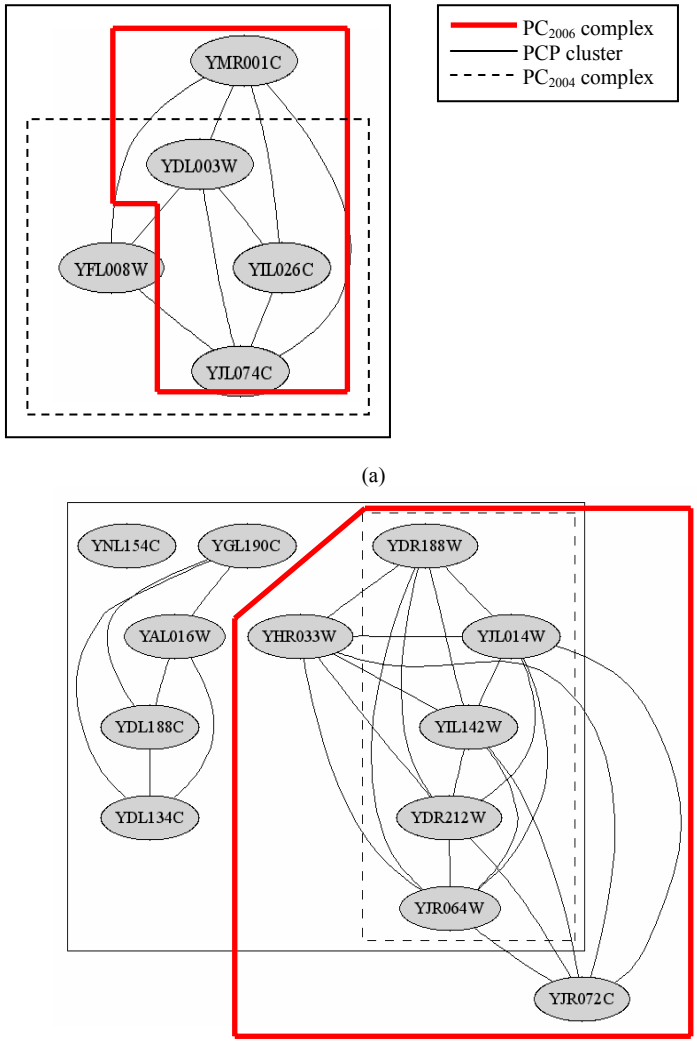


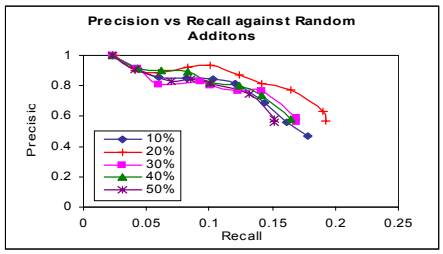
Fig 8. Examples of predicted and matched complexes based on old and new PPI networks. Complexes in PC_{2004}, PC_{2006} and the predicted PCP clusters are shown in different boxes for comparison. (a) The complex in PC_{2004} is of size 4, while in PC_{2006}, its size is 5. PCP predicted 4 proteins in this complex correctly. (b) This complex is of size 5 in PC_{2004}, for which PCP predicted all 5 protein correctly. In PC_{2006}, its size is 11, while PCP algorithm predicted 6 of them correctly.

- **Robustness against noise in interaction data**

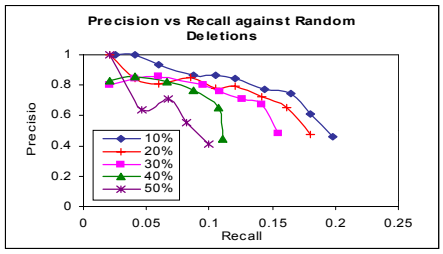
To assess the robustness of the algorithm, we have computed the precision and recall of predictions by PCP when noise of different types and amount is randomly added into the reliable PPI[Combined].

In robustness experiments, noises are usually introduced by swapping edges, or randomize the node labels. However, these methods, which are used in estimating p-values and uniqueness of PPI motifs, are not a good model for our purpose. We are considering errors produced by high-throughput PPI experiments. In this type of experiments, the errors should be closer to edges missing (not detected) or sticky proteins, which are modeled by random noises. Hence, to simulate such noise, we randomly add, delete and reroute (delete and add) 10% to 50% of "pseudo" interactions in the network. The precision and recall of the predicted clusters on the various perturbed datasets are shown in Fig 9.

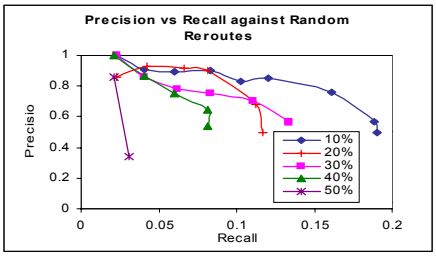
We can see from Fig 9 (a) that the precision against recall of the clusters predicted by PCP remains fairly consistent even with random additions of interactions up to 50% of the original interactions in PPI[Combined]. This is a clear indication that PCP algorithm is robust against spurious interactions. The filtering of the PPI network based on FS-Weight removes most of these random additions, and retains only confident interactions for clustering. Random deletion of interactions has a greater impact on clustering performance, as can be seen in Fig 9 (b). This is analogous to a lack of information, leading a reduction in recall. As FS-Weight is a local topology measure, it becomes less effective when the interaction network become very sparse, since there will be insufficient interactions in the local neighborhood to give a confident score. The formulation of the measure will assign low weights in these cases, which will cause many interactions to be filtered. Nonetheless, precision remains high for clusters that can be discovered. A combination of random addition and deletions results in a simultaneous reduction in precision and recall.



(a)



(b)



(c)

Fig 9. The precision and recall of predictions made by the PCP algorithm when different types and amount of noise are introduced into the reliable PPI network. Three ways of perturbing the network are studied: (a) Random addition (b) Random deletion (c) Random deletion and addition (reroute).

5 DISCUSSIONS AND CONCLUSIONS

Since protein complexes plays an important role in cells, identification of protein complex from PPI networks is an interesting and challenging problem in systems biology. However, current PPI networks are incomplete and contain many errors.

In this paper, we proposed a preprocessing step on PPI networks before complex prediction: 1) introduce level-2 interactions; 2) weigh level-1 and level-2 interactions using FS-Weight; and 3) remove interactions with weight lower than a certain threshold. From our experiments, we have shown that existing clustering algorithms are able to produce clusters that match protein complexes with significantly higher precision and recall using PPI networks processed in this way.

Based on modified PPI network, we have also proposed the PCP clustering algorithm in which, cliques are identified in the network, and merged progressively using the "partial clique merging" method. We have compared PCP with RNSC, MCODE and MCL algorithms and showed that PCP has superior precision and recall in complex prediction. By validating against newer MIPS complex data, we find that PCP can discover novel members of complexes which are only found in the newer complex dataset. Through comprehensive noise analysis, we also showed that PCP maintains high precision even when used on significantly noisier datasets.

Nonetheless, one limitation still plague previous and our current approach: complexes which has subsets of proteins that are not tightly connected to the rest of the complex members cannot be identified, as illustrated in Fig 8 (b). This is inevitable since clustering methods are highly dependent on interaction density. We are currently studying the possibility of using other biological information to represent a more reliable and complete network of relationships between proteins for complex prediction.

Acknowledgements

We would like to thank Igor Jurisica for kindly provide us the source codes of RNSC algorithm. We would also like to thank Sylvian Brohée for providing us with the source codes of the MCL and MCODE algorithms.

References

1. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al*: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000, 403(6770):623-627.

2. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D: MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* 1999, 27(1):44-48.

3. Spirin V, Mriny LA: Protein complexes and functional modules in molecular networks. *PNAS* 2003, 100(21):12123-12128.

4. King AD, Pržulj N, Jurisica I: Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, 20(17):3013-3020.

5. Bader GD, Hogue CW: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, 4(2):27.
6. Brohee S, Helden Jv: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, 7:488.
7. Pržulj N, Wigle DA, Jurisica I: Functional topology in a network of protein interactions. *Bioinformatics* 2003, 20(3):340 - 348.
8. Asthana A, King OD, Gibbons FD, Roth FP: Predicting Protein Complex Membership Using Probabilistic Network Reliability. *Genome Research* 2004, 14(6):1170-1175.
9. Dongen Sv: Graph Clustering by Flow Simulation. 2000(PhD thesis, University of Utrecht).
10. Chua HN, Sung WK, Wong L: Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics* 2006, 22(13):1623-1630.
11. Tomita E, Tanaka A, Takahashi H: The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science* 2006(363):28-42.
12. Breitkreutz BJ, Stark C, Tyers M: The GRID: the General Repository for Interaction Datasets. *Genome Biol* 2003, 4(3):R23.
13. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, Millar A, Taylor P, Bennett K, Boutilier K *et al*: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002, 415:180 - 183.
14. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al*: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, 415(6868):141-147.
15. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al*: Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, 440(7084):631-636.
16. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP *et al*: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 2006, 440(7084):637-643.
17. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 2001, 98(8):4569-4574.
18. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006, 34(Database issue):D535-539.