B. Comp. Dissertation

Protein Complex Inference enhanced by Text Mining

By

Lee Yu Ling Joanne

Department of Computer Science

School of Computing

National University of Singapore

2009/2010

B. Comp. Dissertation

Protein Complex Inference enhanced by Text Mining

By

Lee Yu Ling Joanne

Department of Computer Science

School of Computing

National University of Singapore

2009/2010

Project No: H114200

Advisor: Prof Wong Limsoon

Deliverables:

Report 1 Volume

Abstract

Protein complexes play a vital role in living organisms as they regulate and execute biological processes. As experimental methods of extracting protein complexes are fraught with difficulties, scientists look towards protein complex prediction. However, protein-protein interaction (PPI) data which are used to predict protein complexes are often noisy and incomplete. As published literature may hold a wealth of PPI data which goes unnoticed, this paper aims to enhance the prediction of protein complexes by text mining PPI data from literature abstracts. In this paper, we explore various rule-based methods of extracting PPI data from MEDLINE abstracts. Additionally, we show that the removal of non-hub proteins can reduce the impact of noisy PPI data on protein complex prediction and retrieve smaller and more accurate protein complexes which would otherwise be discarded by CMC. Moreover, we show that the selection of abstracts for augmentation is worth doing to overcome the incompleteness of PPI data.

Subject Descriptors:

J.3 Life and Medical Sciences

H3.3 Information Search and Retrieval

Keywords:

Computational biology, Information retrieval, Information extraction, Protein complex prediction, Protein interaction prediction.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Professor Wong Limsoon for his patience and guidance during the course of this project. I would also like to thank Dr Liu Guimei for providing me with the source codes for CMC and evaluation using GO, Dr Rajesh Chowdhary for providing me the source codes for BN and his guidance in using the program, Su Zhan for his advice in the downloading of MEDLINE abstracts and Dr Upali Kohomban for his help in tagging the abstracts.

Table of Contents

1. Introduction

Proteins are the functional molecules of the cell. Subunits of proteins interact in varying combinations to form protein complexes which regulate and execute biological processes. Despite the importance of protein complexes, there are numerous bottlenecks in the discovery and prediction of protein complexes. While large protein complexes are experimentally difficult to capture, the incompleteness and noisiness of protein-protein interaction (PPI) data hinders the prediction of protein complexes. This is caused by missing edges and the presence of incorrect edges in the PPI graph. Therefore, existing methods of protein complex prediction could not achieve high recall because interaction data is absent or incorrect. This paper hypothesizes that missing information might be reported in literature databases such as MEDLINE which stores close to 11 million records. Hence, the goal of this project is to improve the prediction of protein complexes through text mining of PPI data from MEDLINE abstracts.

The hypothesis was investigated by combining the PPI network built from experimental PPI data with PPIs that are mined from MEDLINE abstracts. Figure 1 gives an outline of the experimental procedure.

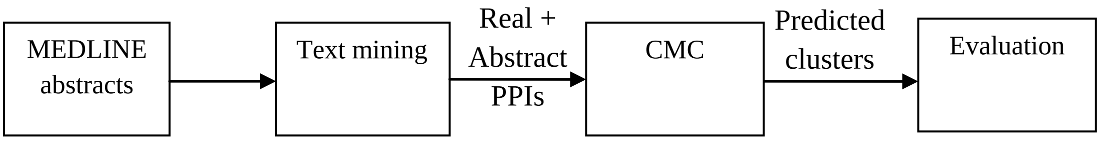


Figure 1: Diagram of experimental procedure.

Three different methods of extracting PPI data from text were attempted. They are (i) co-occurrences of two proteins in a sentence (Co), (ii) co-occurrences of two proteins in a sentence together with a verb from a dictionary of four interaction words (Dict) and (iii) using a trained Bayesian network (BN) (Chowdhary, Zhang and Liu, 2009) that extracts PPI information from abstracts. Separate PPI networks are built from real PPIs, text mined PPIs and a combination of real and text mined PPIs. These PPI networks were used to predict protein complexes and their results were compared.

In further detail, the combined network of real and abstract PPIs fared better in protein complex prediction as compared to separate PPI networks. However, the combined network still contained noisy and incomplete PPI data. Incompleteness of PPI data means that the network has missing PPIs or edges that are part of a real protein complex and noisy PPI data means that incorrect PPIs or edges are present in the network which misleads the prediction of protein complexes. We attempt to overcome the incompleteness of PPI data through augmentation. This was done by text mining the possible missing edges from a new set of MEDLINE abstracts, which is referred to as the augmenting set, that mutually excludes our original set of MEDLINE abstracts and augmenting the new edges to the combined PPI network. Additionally, we attempt to overcome the noisiness of PPI data through the removal of incorrect edges so that smaller and more accurate protein complexes can be predicted. This was done by removing non-hub proteins of predicted complexes from the PPI network and using the edited network to predict new complexes. All PPI networks were weighted and protein complexes were predicted using Clustering based on Maximal Cliques(CMC) (Liu, Wong and Chua, 2009).

The results showed that the hypothesis of this paper is valid. The combined network of real PPIs and PPIs that are mined from MEDLINE abstracts lead to an improvement in the prediction of real protein complexes. Out of the three methods used to extract PPIs from text, Dict produced a network which fared better in predicting protein complexes. Moreover, the combined network of real PPIs and PPIs from Dict performed the best in predicting protein complexes. While only a small augmenting set of abstracts were used to augment the network, some complexes benefitted greatly from the augmentation. In addition, the removal of non-hub proteins allowed the protein prediction program used in this project to capture smaller and more accurate complexes which were not predicted by the original network.

1.1 Report outline

In the introduction, we identified the incompleteness and noisiness of PPI data to be a bottleneck in protein complex prediction and stated the hypothesis that missing PPI data might be found in MEDLINE abstracts. In addition, we mentioned that the goal of the project was to improve protein complex prediction with the use of PPI data from MEDLINE abstracts. This was followed by a summary of the experiments that were carried out in this project. Furthermore, we briefly described the corresponding results and these results supported our hypothesis.

The rest of the report will be organised in the following manner. Firstly, background information which is important to understanding this project will be presented. This includes PPI network, techniques for extracting PPI from literature and predicting protein complexes given a PPI network. This will be followed by a description of the experiments that were carried out. Next, the methods that were used to evaluate the quality of the predicted complexes will be discussed. Subsequently, the results of the experiments will be presented and analyzed. This will be followed by a chapter on future work and the conclusion of this paper.

2. Background information

This chapter presents some basic information about PPI network and the experimental techniques and non-experimental techniques used to derive PPI data. In addition, techniques which are used for extracting PPIs from text will be presented. Lastly, this chapter will also discuss the algorithms that are used to predict protein complexes given a PPI network.

2.1 Discussing PPI network

A PPI network summarizes PPI data into an undirected graph G = (V, E), where distinct proteins are represented by vertices and the interaction between any two proteins by edges. PPI data are generated on a large scale using high-throughput experimental techniques such as yeast-two hybrid (Y2H), affinity purification-mass spectrometry (AP-MS) and protein microarray. The experimental techniques of producing PPIs and their limitations will be briefly described followed by non-experimental methods of getting PPI data such as using databases and natural language processing.

2.1.1 Yeast-two hybrid (Y2H) method

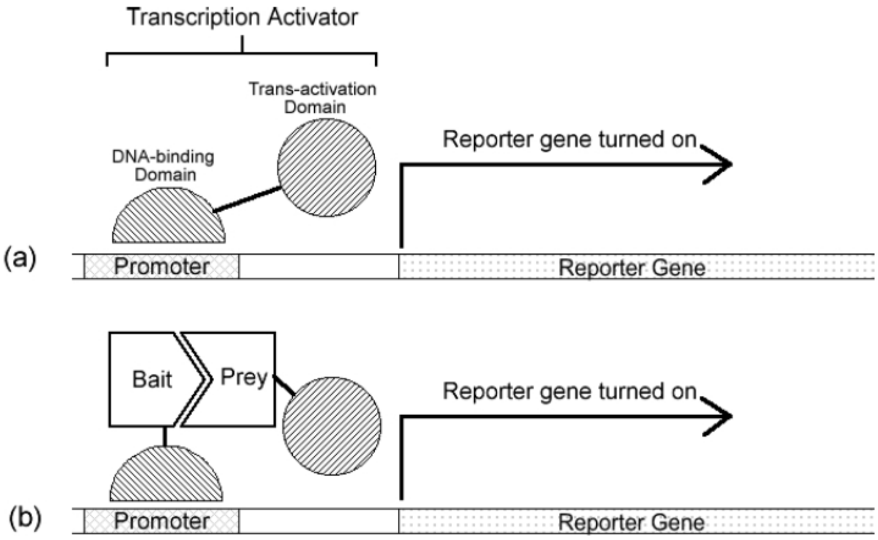


Figure 2: Interaction detection by Y2H. (a) Activation of reporter gene by transcriptional activator. (b) Activation of reporter gene by reconstituted transcriptional activator (Ng and Tan, 2004).

In Y2H, scientists detect PPI by making use of the naturally occurring transcriptional activator in yeast. The proteins of interest which are called “bait” and “prey” are attached to separate domains of the transcriptional activator before introducing them into the yeast cell. An interaction between the bait and prey will turn on the reporter gene, such as the green fluorescent protein gene. This will indicate the presence of an interaction.

There are two limitations to detecting PPI by Y2H. They are low coverage of proteins and high error rates with some experiments reporting 50% false positives (Ng, 2004).

2.1.2 Affinity purification-mass spectrometry (AP-MS) method

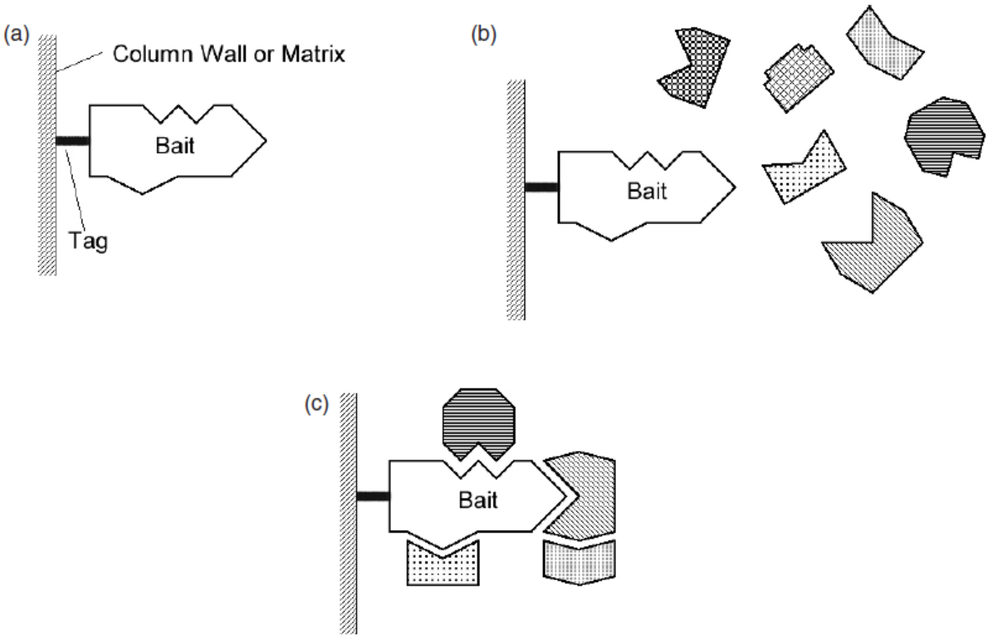


Figure 3: Interaction detection by AP-MS (Ng, 2004).

In AP-MS, groups of interacting proteins can be detected. Similar to Y2H, a bait protein is used and it is immobilized on a column wall while mixtures of candidate proteins are passed through the column. Interacting proteins will be captured by the bait while non-interacting proteins will be eluted away. The interacting proteins are identified using mass spectrometry.

Similar to Y2H, the limitations of AP-MS include low coverage of proteins and high error rates in terms of false positives and false negatives (Ng, 2004).

2.1.3 Protein microarray method

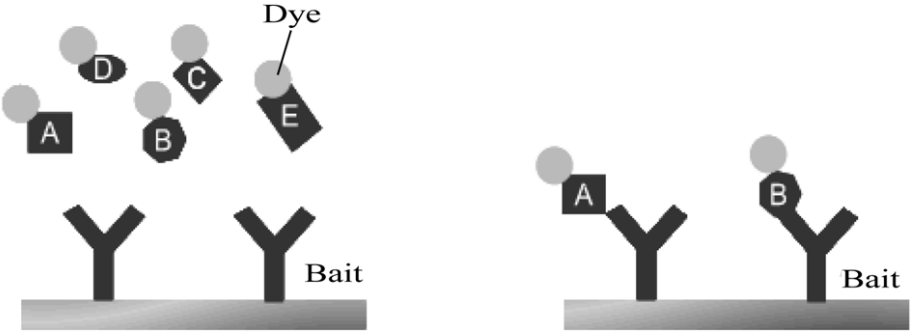


Figure 4: Interaction detection by protein microarray (Vancouver prostate centre, 2005).

In protein microarray, the detection of groups of interacting proteins is multiplexed on a microarray chip. Thousands of bait proteins are spotted at a unique location on the chip. Candidate proteins which are pre-attached to a suitable dye are allowed to pass through the chip. Interacting proteins will be captured by the bait while non-interacting proteins will be eluted away. The attached dye allows the interacting proteins to be detected.

The limitation of protein microarray is in the synthesis of bait proteins and the maintenance of its structure and function while it is on the chip.

2.1.4 PPI database

As the experimental techniques discussed tend to produce noisy and incomplete data, PPI databases such as the Database of Interacting Proteins (DIP) were created to provide comprehensive and accurate PPI data sources. These databases are hand curated from experimentally determined protein interactions and biological literature to ensure valid entries (Shoemaker and Pancheko, 2007). Thus, evaluating literature becomes the rate limiting step in the growth of the database (Marcotte, Xenarios and Eisenberg, 2000). As such, people turn to natural language processing to automatically discover protein interactions.

2.1.5 Natural language processing (NLP)

NLP can be used to discover PPIs by parsing sentences in biology papers into grammatical units. This can be broadly divided into three categories: rule-based, shallow parsing and deep parsing (Zhou, He and Koh, 2006). Rule-based methods uses a set of predefined patterns to extract PPIs, shallow parsing extract local dependencies among phrases without reconstructing the structure of the entire sentence to determine PPIs and deep parsing takes into account the structure of the entire sentence. Table 1 shows the performance reported so far in terms of recall and precision for the various methods.

Category	Performance	
	Recall (%)	Precision (%)
Rule-based	86	94
Shallow parsing	62	89
Deep parsing	48	80

Table 1: Performance of mining PPI from literature (Zhou et al, 2006).

Although no benchmark datasets have been used to provide a fair comparison of the methods, rule-based methods appear to achieve the best performance so far. In addition, rule-based methods are also popular in extracting PPIs as recent studies which extract PPIs from PubMed abstracts tend to use rule-based methods (Chowdhary et al, 2009). For example, it is used in Protein Interaction Information Extraction System (PIE), a good PPI extraction web service that extract PPIs from literature (Sun et al, 2008).

2.2 Techniques for extracting PPI from literature

The problem of PPI extraction consists of two components. The first deals with protein name recognition or tagging of protein names and the second deals with PPI extraction. (Chowdhary et al, 2009) This section discusses the techniques of extracting PPIs given correct tagged protein names. As it was previously mentioned that rule-based methods are popular and fare well in PPI extraction, this paper will focus on using rule-based methods in mining PPIs from text.

Rule-based methods include using co-occurrences of two proteins (Co), manually specified rules using a dictionary of four interaction verbs (interact, bind, complex, associate) (Dict), concept profile-based relation (Herman et al, 2009) and machine learning such as using a trained Bayesian network (BN). (Chowdhary et al, 2009) The rule-based methods will be discussed in greater detail.

2.2.1 Co-occurrences of two proteins (Co)

This method of PPI extraction extracts two proteins based on their appearance in the same sentence. It assumes that two proteins tend to interact if it occurs together in the same sentence. This is the simplest rule-based method and it represents how biologists might search for information (Herman et al, 2009). However, it tends to produce a large number of false positives (Chowdhary et al, 2009).

2.2.2 Manually specified rules using a dictionary of four interaction verbs (Dict)

This method of PPI extraction extracts two proteins based on their appearance in the same sentence as well as the occurrence of a dictionary word in that sentence. The dictionary consists of four interaction verbs: interact, bind, complex, associate. This method of extraction achieved an average recall of 83.6% and precision of 93.2% by Ono et al (Chowdhary et al, 2009). However, an analysis of Ono et al's dataset by Chowdhary et al revealed that a high proportion of true samples, comprising of 72.7%, was present in the dataset. This shows that a simple method of extracting PPIs can be used on literature abstracts if the set of abstracts tend to contain protein interactions.

2.2.3 Concept profile-based relation

This method of PPI extraction extracts two proteins based on their appearance in the same abstract as well as similarity in concept profiles of the two proteins. A protein concept profile is a summary of the context in which the protein appears in the literature. The attributes of the concept profile include concepts such as diseases, symptoms, tissues and biological processes. The successful use of concept profile-based relation to extract PPIs show that PPIs can be predicted without having their interaction explicitly described in literature (Herman et al, 2009).

2.2.4 Bayesian network (BN)

This method of PPI extraction first extracts two proteins based the presence of a "PPI triplet" and outputs the PPI if the likelihood of their interaction being true from the trained BN is greater than 50%. A PPI triplet is defined as two proteins which co-occur in the same sentence together with the presence of an interacting word. A BN summarizes a set of language features into a directed acyclic graph with probabilistic relationship. Each vertex represents a language feature and the edge that connect the features make up the rules with some conditional probability attached to it. Unlike previous methods of PPI extraction, this method uses machine learning to learn the structure of the BN and its parameters.

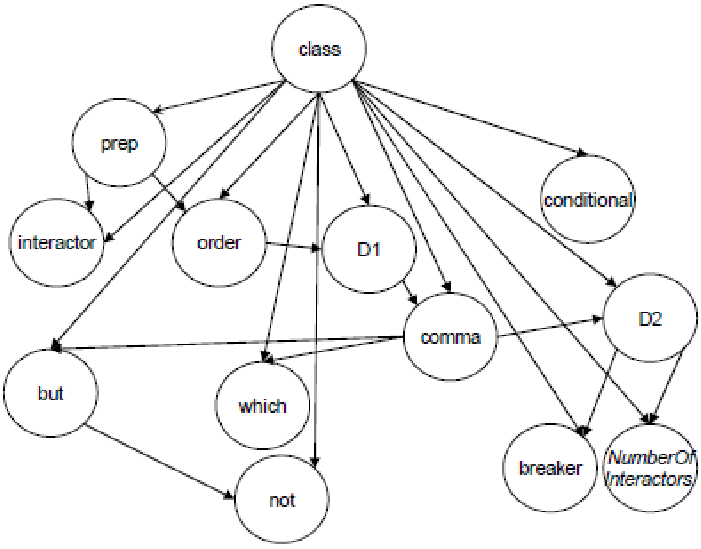


Figure 5: BN structure learned from cross-validation (Chowdhary et al, 2009).

The BN comprises of 12 manually selected language features and each feature is constrained to a maximum of two parent node. When a PPI triplet is found, the language feature of the sentence that the triplet is in will be calculated and the PPI triplet is classified using the network structure and parameters in figure 5 by calculating the posterior probability of the class C being true using Bayes' theorem.

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

Figure 6: Bayes' theorem.

Class C comprises of two categories: true or false. The PPI triplet being evaluated will be classified into one of the two categories depending on whether the BN evaluates the interaction to be a true or false interaction. The feature vector E represents the language features in the learnt BN. From the equation in figure 6, P(C) is the prior probability of C, P(E) is the marginal probability of the feature vector E and P(E|C) is the conditional probability of observing the feature vector E in a given class C. The calculation for P(E) is simplified by assuming conditional independence.

The advantage of BN is this method extract PPIs based on rules which are representative of those used in unstructured texts such as word order, co-occurrences and distance between two protein names. In addition, it is scalable for large PPI extraction as it uses Dirichlet parameter priors in the BN framework which helps to resolve missing data and ambiguity (Chowdhary et al, 2009).

2.3 Protein complex prediction given PPI network

Protein complexes can be inferred from a PPI network as the combination of PPI data can accurately define members of a protein complex (Krycer, Pang and Wilkins, 2008). Algorithms that have been proposed to predict protein complexes from PPI networks usually search for dense sub-graphs as it has been observed that dense regions in PPI networks correspond to protein complexes (Wu, Li and Kwoh, 2004). This is a clustering problem and proposed algorithms adopt different strategies to find clusters. The three algorithms that will be discussed are Markov Clustering (MCL), Molecular Complex Detection (MCODE) and Clustering based on Maximal Cliques (CMC) (Liu et al, 2009).

2.3.1 Markov Clustering (MCL)

MCL find clusters by simulating many random walks in a graph. It is based on the assumption that many random walks in a graph can capture dense region of the graph since there is a higher probability of starting and ending in the same dense region. The MCL algorithm iterates two steps: the expansion and contraction step. The expansion step visits more neighbours while the contraction step rank neighbours which are more favourable (lim, 2009). MCL looks for clusters globally and generate only non-overlapping clusters (Wu et al, 2004).

2.3.2 Molecular Complex Detection (MCODE)

MCODE find clusters using dense regions of the graph as seeds and expand outwards to form clusters. There are three steps to the algorithm. Firstly, every vertex in the graph is weighted by their local neighbourhood densities. Secondly, vertices with high weights are selected as seeds or initial clusters and neighbouring vertices are augmented if their weight is above a given threshold. The algorithm continues to augment neighbouring vertices until no more vertices can be added. Lastly, the generated clusters will be post-processed. The post-processing stage removes clusters that do not contain vertices with minimum degree of two and adds some vertices, called “fluff”, to the remaining clusters. The addition of fluff allows the generation of overlapping clusters. Thus, MCODE looks for clusters locally and may generate overlapping clusters (Bader and Hogue, 2003).

2.3.3 Clustering based on Maximal Cliques (CMC)

CMC uses clique finding and merging strategy on a weighted PPI network to find protein complexes. A clique is defined as a maximal complete sub-graph. Figure 7 shows the schematic diagram of the algorithm.

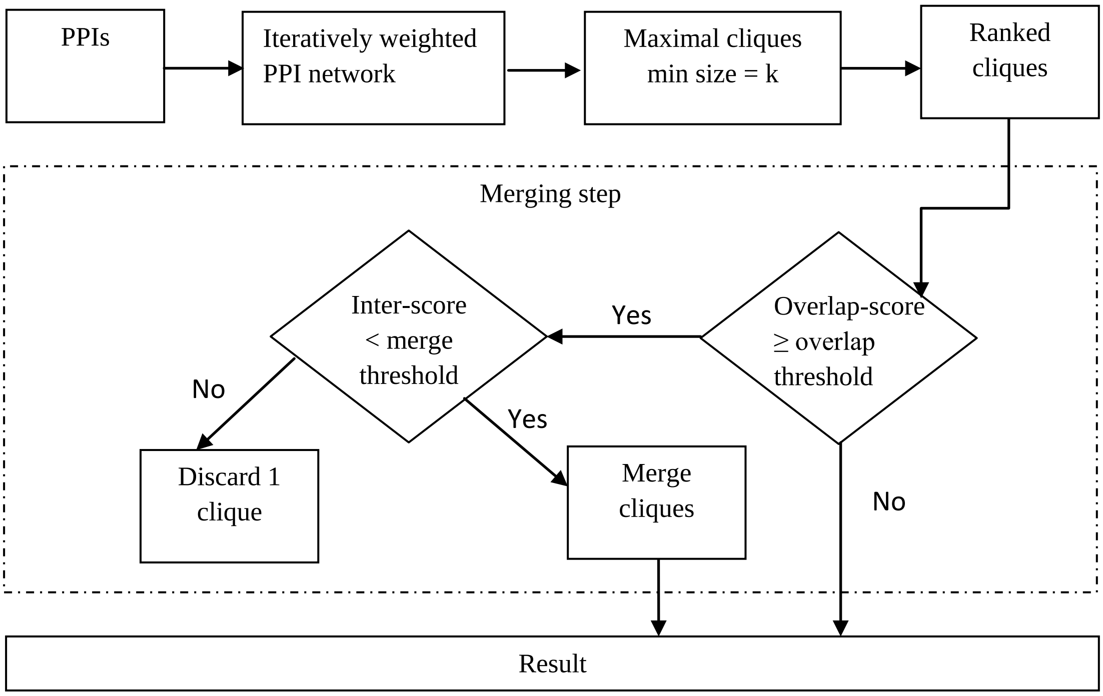

PPIs
Iteratively weighted PPI network
Maximal cliques min size = k
Ranked cliques
Merging step
Inter-score < merge threshold
Overlap-score ≥ overlap threshold
Yes
No
Yes
No
Discard 1 clique
Merge cliques
Result

Figure 7: Diagram of CMC algorithm.

A PPI network is first built from the set of PPIs. The edges of the PPI network are then iteratively weighted using the AdjustCD weighting algorithm which is based on the neighbourhood densities of the vertices that are connected by the edge. The intention behind weighting the edges is to provide a measure of reliability of the PPIs so as to reduce the impact of noisy data on clique finding. After weighting, the PPI network is used to enumerate all maximal cliques greater than or equal to size k. The enumerated cliques are scored according to their weighted density and ranked in decreasing order of their score before going to the merging step. In the merging step, CMC searches for the existence of another clique which satisfies the conditions in the merging step and merges, discards or accepts the cliques accordingly. Discarding highly overlapped cliques serves to reduce result size while merging of two cliques with high overlap generates a bigger clique which was not predicted due to the incompleteness of the PPI data (Liu et al, 2009).

2.3.3.1 AdjustCD weighting algorithm

As mentioned, AdjustCD is used to iteratively weigh the edges of the PPI network. The AdjustCD equation will be discussed first followed by the iterated AdjustCD equation.

Given a pair of proteins (u, v) in a PPI network G = (V, E),

$$AdjustCD(u,v) = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

$$\lambda_w = \max\{0, \frac{\sum_{x\epsilon V} |N_x|}{|V|} - |N_w|\}$$

Where Nu: set of neighbours of u in G and Nv: set of neighbours of v in G

Figure 8: The AdjustCD equation (Liu et al, 2009).

The AdjustCD equation in figure 8 scores the edges based on the number of neighbours of the two proteins and penalizes proteins with few interactions. The intuition behind the formula can be illustrated with an example with two cases. In the first case, u has 1 neighbour ($|N_u|=1$), v has 1 neighbour ($|N_v|=1$) and the neighbour of u and v are themselves ($|N_u \cap N_v|=1$). In the second case, u have 10 neighbours ($|N_u|=10$), v have 10 neighbours ($|N_v|=10$) and the neighbours of u and v are the same ($|N_u \cap N_v|=10$). If the average degree of the network is 4, AdjustCD(u,v) = 0.25 for the first case and AdjustCD(u,v)=1 for the second case. The values reflect the reliability of the interaction because if u has 1 incorrect neighbour in the first case, the interaction between u and v is incorrect. On the other hand, even if u and v have 4 incorrect neighbours in the second case, their interaction may still be correct. Therefore, figure 8 allows the weight of the edge to reflect the reliability of the PPIs in complex prediction (Liu et al, 2009) unlike edges that are represented in binary where 1 indicates the presence of an edge and 0 indicates the absence of an edge.

Given a pair of proteins (u, v) in a PPI network G = (V, E),

$$w^k(u,v) = \frac{\sum_{x\epsilon Nu \cap N_v}(w^{k-1}(x,u) + w^{k-1}(x,v))}{\sum_{x\epsilon N_u} w^{k-1}(x,u) + \lambda_u^k + \sum_{x\epsilon N_v} w^{k-1}(x,v) + \lambda_v^k}$$

$$\lambda_y^k = \max\{0, \frac{\sum_{x\epsilon V} \sum_{z\epsilon N_x} w^{k-1}(x,z)}{|V|} - \sum_{x\epsilon N_y} w(x,y)\}$$

Where $w^{k-1}(x,u)$ is the score of (x,u) in the (k-1)th iteration

Figure 9: The iterative AdjustCD equation (Liu et al, 2009).

The difference between AdjustCD and iterative AdjustCD is the latter uses the score of the edge from the previous iteration to calculate the score for the next iteration. Iteratively weighting the network attempts to get a value for the edge which is a better estimate of the true value of the interaction. The authors of CMC have shown that iterative AdjustCD can improve functional homogeneity and localization coherence of top ranked interactions.

2.3.4 Comparing clustering algorithms

Previously, three clustering algorithms for protein complex prediction were discussed. They are MCL, MCODE and CMC. A comparative study of Markov Clustering (MCL), Molecular Complex Detection (MCODE) and two other algorithms have reported MCL to perform better in predicting real complexes and MCODE to be better in predicting high quality complexes (Wu et al, 2004). The result of the study is shown in figure 10.

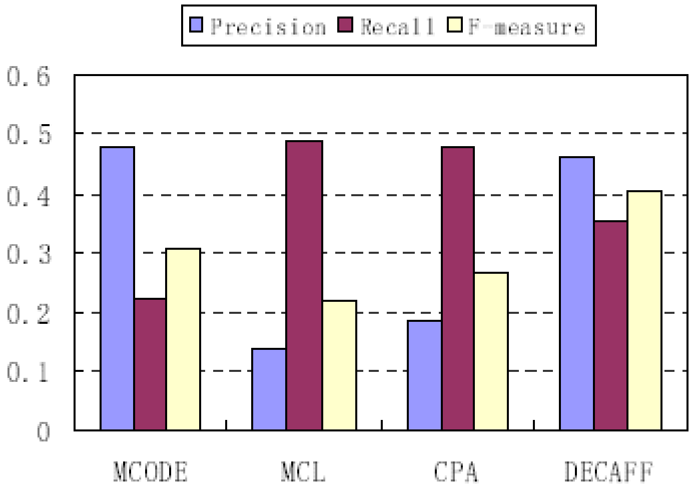
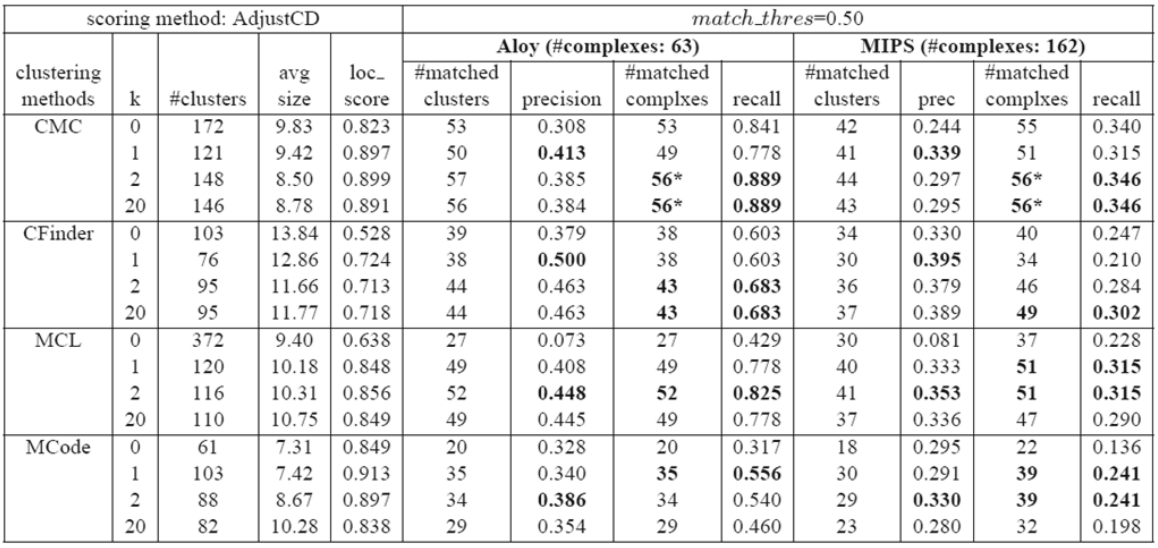


Figure 10: Comparison of four clustering algorithms using MIPS dataset. (Wu et al, 2004)

The study also suggests that future algorithms should have high recall like MCL and better precision like MCODE. Although DECAFF seems to perform well as it has a relatively high recall and precision as well as the highest F-measure, the algorithm generated complexes with high redundancy (Wu et al, 2004). CMC is an algorithm that achieves higher recall than MCL and higher precision than MCODE (Liu et al, 2009).



scoring method: AdjustCD					*match_thres*=0.50							
					Aloy (#complexes: 63)				**MIPS (#complexes: 162)**			
clustering methods	k	#clusters	avg size	loc_ score	#matched clusters	precision	#matched complxes	recall	#matched clusters	prec	#matched complxes	recall
CMC	0	172	9.83	0.823	53	0.308	53	0.841	42	0.244	55	0.340
	1	121	9.42	0.897	50	**0.413**	49	0.778	41	**0.339**	51	0.315
	2	148	8.50	0.899	57	0.385	**56***	**0.889**	44	0.297	**56***	**0.346**
	20	146	8.78	0.891	56	0.384	**56***	**0.889**	43	0.295	**56***	**0.346**
CFinder	0	103	13.84	0.528	39	0.379	38	0.603	34	0.330	40	0.247
	1	76	12.86	0.724	38	**0.500**	38	0.603	30	**0.395**	34	0.210
	2	95	11.66	0.713	44	0.463	**43**	**0.683**	36	0.379	46	0.284
	20	95	11.77	0.718	44	0.463	**43**	**0.683**	37	0.389	**49**	**0.302**
MCL	0	372	9.40	0.638	27	0.073	27	0.429	30	0.081	37	0.228
	1	120	10.18	0.848	49	0.408	49	0.778	40	0.333	**51**	**0.315**
	2	116	10.31	0.856	52	**0.448**	**52**	**0.825**	41	**0.353**	**51**	**0.315**
	20	110	10.75	0.849	49	0.445	49	0.778	37	0.336	47	0.290
MCode	0	61	7.31	0.849	20	0.328	20	0.317	18	0.295	22	0.136
	1	103	7.42	0.913	35	0.340	**35**	**0.556**	30	0.291	**39**	**0.241**
	2	88	8.67	0.897	34	**0.386**	34	0.540	29	**0.330**	**39**	**0.241**
	20	82	10.28	0.838	29	0.354	29	0.460	23	0.280	32	0.198

Table 2: Comparison of four clustering algorithms using MIPS and Aloy dataset (Liu et al, 2009).

Despite the good results achieved by CMC, there are three limitations to the algorithm. The first two limitations to the CMC algorithm were identified by Lim (Lim, 2009). Firstly, important clusters may be discarded during the merging step. Secondly, initial cliques that have a good representation of a particular protein complex may produce a false negative when merged. A third limitation of CMC was identified during this project and it is the use of cliques being a stringent condition in predicting protein complexes. This is because most PPI network graphs that correspond to protein complexes were not as dense as what previous complex finding algorithms have theorized (Gallagher and Goldberg, 2009). Hence, we propose and investigate if the removal of non-hub proteins in the network will lead to the prediction of smaller and more accurate complexes using CMC.

3. Experiments

The goal of this project is to attempt various experiments to improve the prediction of protein complexes with the use of PPI data from MEDLINE abstracts. This is because experimental data tend to be noisy and incomplete. According to the previous chapter, rule-based methods are commonly used to extract PPI data from text. Moreover, rule-based methods tend to produce better results in the extraction of PPIs. The three rule-based methods that will be used are Co, Dict and BN. For the prediction of protein complexes, CMC will be used as it was shown that CMC achieves better recall and precision than other existing algorithms such as MCL and MCODE.

A total of five experiments were conducted and all experiments include abstract PPI data. The purpose of experiment 1 and 2 is to establish the parameters used in CMC for subsequent experiments. Experiment 1 determines the number of iteration to be used for weighing the PPI network using iterative AdjustCD (Figure 9). Experiment 2 determines the optimal merge threshold and overlap threshold for the merging step in CMC (Figure 7). In the subsequent experiments, various methods were attempted to improve the prediction of protein complexes. Experiment 3 uses the two different methods: Dict and BN to extract PPIs from abstracts. The goal of experiment 3 is to determine the best method for extracting PPIs from text. Next, experiment 4 attempts to remove noisy PPI edges by removing non-hub proteins from the PPI network. Lastly, experiment 5 carried out augmentation of the PPI network in an attempt to fill in missing edges of the network. The new set of abstracts which is named the augmenting set mutually excludes the initial set of abstracts.

For all experiments, only clusters of size 4 and above were enumerated since larger sized protein complexes are biologically difficult to capture. The dataset used will be presented followed by details of the five different experiments.

3.1 Dataset

This section gives details about the real PPIs, abstracts used for extracting PPIs and the reference complexes used for evaluating the predicted complexes. The abstracts used comprises of the initial set of abstracts (Li, 2008) and the augmented set that is retrieved for experiment 5. The initial set of abstracts was used in all experiments.

3.1.1 Real PPIs

The real PPIs are from Liu et al. The dataset contain yeast protein interactions generated by six different individual experiments. The dataset contain 3295 proteins and 15900 interactions, among which 10458 interactions have common neighbours.

3.1.2 Initial set of abstracts

The initial set of abstracts was retrieved from the Pubmed database. A querying program looks for abstracts containing the names and synonyms of the proteins from Saccharomyces Genome Database (SGD). The abstracts which are included were limited to the first 1000 abstracts and an added constraint to limit the search to title and abstracts with the Pubmed filter option (Li, 2008).

The initial set of abstracts comprises of a total of 192082 abstracts. Of which, 186798 were non-empty abstracts and they contain a total of 659598 sentences.

3.1.3 Augmenting set of abstracts

The augmenting set of abstracts is the new abstracts that were retrieved from MEDLINE database. A querying program looks for abstracts containing the names and synonyms of proteins from SGD. The abstracts which are included were limited to the first 100 abstracts for each protein. This set of abstracts mutually excludes the initial set of abstracts. This is because only proteins which are found in reference complexes but not in predicted complexes from the combined network of real and Dict PPIs in experiment 3 are used to query the database and the initial set of abstracts would only contain abstracts with proteins that are found in the predicted complexes. The abstracts were tagged for their protein names using the Name Entity Recognition (NER) software (Zhou and Su, 2004) at the Institute for Infocomm Research (I^2R).

The augmenting set of abstracts comprises of a total of 43521 abstracts. Of which, 43516 were non-empty abstracts and they contain a total of 432971 sentences. The augmenting set was only used in experiment 5.

3.1.4 Reference complexes

Two reference sets of protein complexes are used. The first set is hand-curated complexes from MIPS and the second set is from Aloy. For both sets, only complexes with size 4 and above are kept (Liu et al, 2009). A total of 164 complexes are present in MIPS and a total of 62 complexes are present in Aloy. These two sets were used in experiment 1.

It was noted that some complexes in Aloy overlapped with MIPS. Hence, the union of both reference sets called AloyMIPS were obtained. A total of 213 complexes are present in AloyMIPS. Of which, 164 complexes were from MIPS and 49 complexes were from Aloy. AloyMIPS was used as the reference set for experiments 2 to 5 while Aloy and MIPS were used separately in experiment 1.

3.2 Experiment 1 – Number of iteration for AdjustCD

In this experiment, co-occurrences of two proteins in the same sentence (Co) was used to extract the PPIs from the initial set of MEDLINE abstracts. No real PPIs were used in this experiment and abstract PPIs are from the initial set of abstracts. The network was weighted with varying number of iterations of iterative AdjustCD. The overlap threshold of 0.5 and merge threshold of 0.25 that were suggested by Liu et al for real PPIs was used in CMC. The threshold for evaluating the clusters using recall and precision was set to 0.5. The threshold of 0.5 indicates that the predicted complexes must have at least 50% overlap with the reference complexes in order to be counted in recall and precision. Recall measures the completeness of the predicted complexes while precision measures the fidelity of the predicted complexes. Further details are given in section 4.1.

	Aloy		MIPS	
Number of iteration	Recall	Precision	Recall	Precision
5	0.403	0.015	0.294	0.026
10	0.403	0.016	0.281	0.025
20	0.403	0.016	0.281	0.025
30	0.403	0.016	0.281	0.025

Table 3: Recall and precision for Co with different number of iteration.

Table 3 shows stability of recall and precision after 10 iterations. To be on the safe side, the number of iterations was set to 20 for future experiments.

3.3 Experiment 2 – Optimal merge threshold and overlap threshold

In experiment 2, co-occurrences of the two proteins in the same sentence with a dictionary verb of “interact”, “bind”, “complex” or “associate” (Dict) were used to extract the PPIs from the initial set of abstracts. No real PPIs were used in this experiment and abstract PPIs are from the initial set of abstracts. Protein complexes are predicted from the weighted Dict network using different values of overlap threshold and merge threshold.

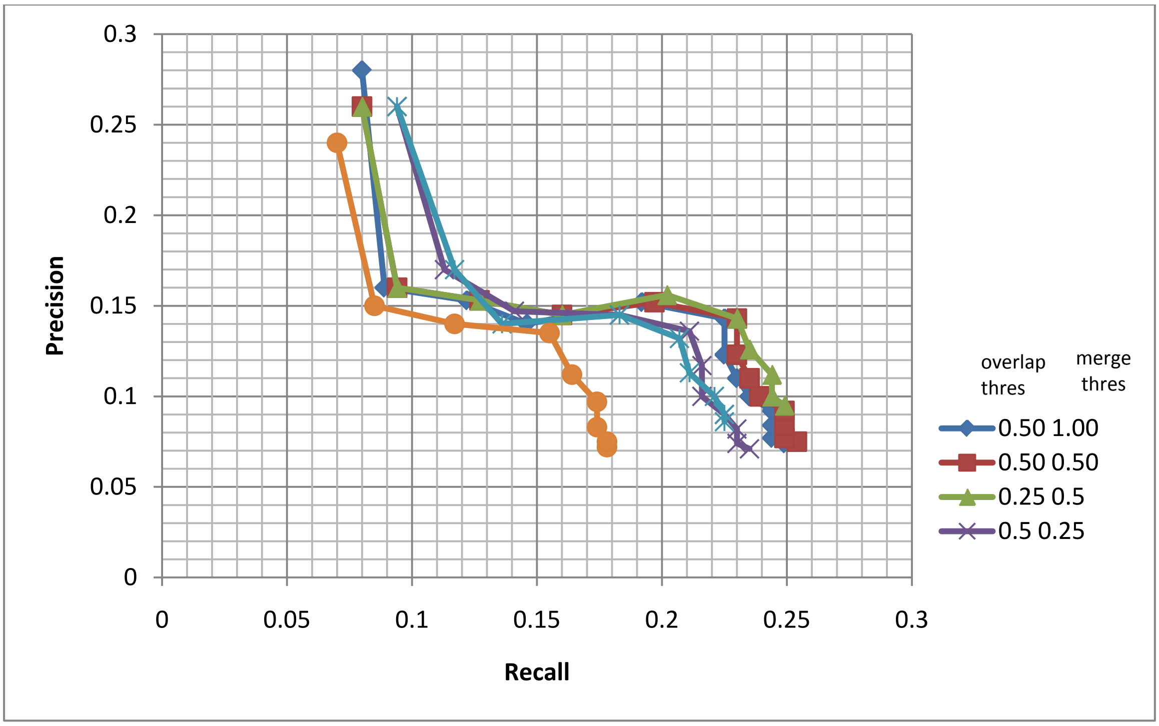


Figure 11: Precision-recall of PPI network from Dict under different parameter settings.

The result showed that an overlap threshold of 0.25 and merge threshold of 0.5 is optimal as it has the largest area under the precision-recall curve in the graph. Hence, these threshold values are used for future experiments that contain abstracts PPIs.

3.4 Experiment 3 – Using Dict and BN to extract PPI from text

After determining the necessary parameters such as number of iteration, overlap threshold and merge threshold. The effectiveness of the different methods of extracting PPIs from text could be compared. In experiment 3, four different PPI networks were built for protein complex prediction and the results were evaluated. The four different PPI networks are based on real PPI, Dict, trained BN by Chowdhary et al and the combined network of real PPI and Dict. The network which comprises of only real PPIs used an overlap threshold of 0.5 and merge threshold of 0.25 suggested by Liu et al for real PPIs. The other three networks used an overlap threshold of 0.25 and merge threshold of 0.5 which is optimal for abstract PPIs from experiment 2. As the combined network of real PPI and Dict gave the best results for prediction of complexes, this method will be used in experiment 4 and 5.

3.5 Experiment 4 – Iterative removal of non-hub proteins

As PPI data tend to be noisy, experiment 4 attempts to improve the prediction of protein complexes through the removal of noisy data. Although the effect of noisy edges on the PPI network are dampened due to the weighting of the network, noisy edges may still affect the prediction of protein complexes due to the clique finding and merging strategy in CMC. This is because in the merging step, some highly overlapping clusters are discarded. Noisy edges may contribute to the overlap. Besides noisy edges, real complexes can overlap one another and a real complex may be discarded by CMC. Hence, we attempt to remove non-hub proteins from the PPI data to reduce the overlap between clusters and attempt to retrieve clusters that are discarded. Non-hub proteins are proteins that participate in very little complexes.

At iteration 0, there were no changes to the procedure and the experiment followed figure 1. The predicted clusters from iteration 0 form the set of unique clusters. In iteration 1, proteins that occurred in n=1 unique clusters were removed from the PPI data and fed into CMC for prediction. Predicted clusters that were not in the unique clusters from iteration 0 were added to the set of unique clusters. This will be followed by iteration 2 where proteins that occurred in n=2 unique clusters will be removed from the PPI data. This process continues until iteration 5. The intuition for iterating the removal of non-hub proteins is to prevent the removal of hub-proteins which will affect clique generation in CMC. Proteins that occur less frequently in the predicted clusters were removed first to try and reduce the overlap between cliques generated by CMC so that we can retrieve clusters that would have been discarded in the previous iteration. The modification that was described is illustrated in figure 12.

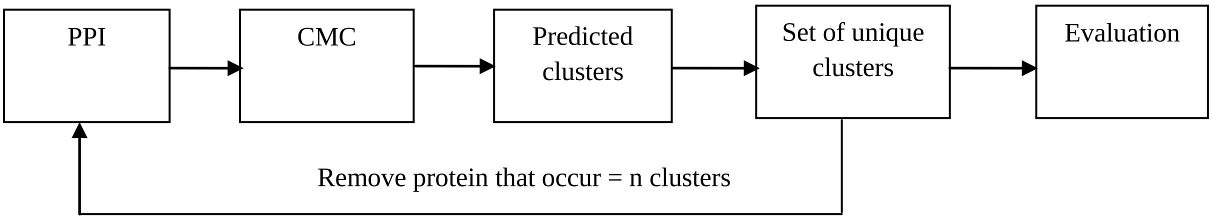


Figure 12: Procedure to remove non-hub proteins.

The PPI data that was used at iteration 0 is the combined set of real PPIs and Dict. The iteration stopped at n=5 as the network became small and the recall stabilized (Table 5).

3.6 Experiment 5 – Augmenting the PPI network

The purpose of augmentation is to add in correct edges into the network to determine if more accurate protein complexes would be predicted. First, a list of proteins which are found in AloyMIPS but not in the predicted complexes from the combined network of real and Dict PPIs in experiment 3 were generated. As there are many different names for the same protein, their synonyms were added to the list. This list was used to query the database of MEDLINE abstracts and the augmenting set of abstracts was retrieved and tagged. Details of retrieval and tagging were given in section 3.1.3. Lastly, PPIs were extracted from the new abstracts using Dict and the extracted PPIs were augmented to the original set of real and Dict PPIs. This augmented set of PPIs was passed to CMC for protein complex prediction. The results of the experiment will be discussed in section 5.3.

4. Evaluation methods

The quality of the predicted clusters are evaluated using 3 methods: Jaccard index, subset evaluation and Gene ontology (GO). This chapter will present and explain the evaluation methods.

4.1 Jaccard index

Jaccard index was used to calculate the match score of the predicted cluster against the reference complex.

Given a predicted cluster S and reference complex C,

$$match\_score(S, C) \ = \frac{|V_S \cap V_C|}{|V_S \cup V_C|}$$

Where V_s: set of proteins contained in S and V_c: set of proteins contained in C

Figure 13: Equation for Jaccard index (Liu et al, 2009).

The equation in figure 13 calculates the proportion of overlap between a predicted cluster and a reference complex. The predicted cluster is said to match the reference complex if there is more than 50% overlap between the two complexes. Hence, a match threshold of 0.5 was used in this project and there is a match when match_score(S, C) ≥ match threshold. This definition of match is used in the calculation of recall and precision in figure 14.

Given a set of reference complexes C = {C_1, C_2, ... , C_n} and a set of prediction clusters P = {S_1, S_2, ... , S_n}

$$Recall = \frac{|\{C_i | C_i \in C \land \exists S_j \in P, S_j \ \ matches\ C_i\}|}{|C|}$$

$$Precision = \frac{|\{S_j | S_j \in P \land \exists C_i \epsilon\ C, C_i\ matches\ S_j\}|}{|P|}$$

Figure 14: Equation for calculating recall and precision (Liu et al, 2009).

Recall is a measure of completeness or the ratio of the number of predicted clusters that match the reference complexes against the total number of reference complexes. A high recall indicates that the experiment retrieved more genuine complexes.

Precision is a measure of exactness or the ratio of the number of reference complexes that match the predicted clusters against the total number of predicted clusters. A high precision indicates that the experiment retrieved better quality complexes.

4.2 Subset evaluation

In addition to recall and precision, subset evaluation is another method to evaluate the quality of complexes and uses subset score.

Given a set of reference complexes C = {C_1, C_2, ... , C_n} and a set of prediction clusters P = {S_1, S_2, ... , S_n}

$$subset\_score(S_i, C) = \max_{\mathrm{C_i \epsilon C}} \frac{|S_i \cap C_i|}{|S_i|}$$

$$subset\_score(C_i, S) = \max_{\mathrm{S_i \epsilon S}} \frac{|C_i \cap S_i|}{|C_i|}$$

Figure 15: Equation for calculating subset score.

Subset_score(S_i, C) is calculated for all predicated clusters. A high score for subset_score(S_i, C) indicates that a large part of the predicted cluster is a subset of the reference complexes. Similarly, subset_score(C_i, S) is calculated for all reference complexes and a high score for subset_score(C_i, S) indicates that a large part of the real complex is a subset of the predicted complexes. Subset_score(Si, C) is similar to precision and subset_score(Ci, S) is similar to recall. Hence, good quality complexes will have a high score for subset_score(S_i, C) and high score for subset_score(C_i, S). To give an overall view of the quality of the predicted clusters, the frequencies of occurrences of the scores are tabulated and a graph of frequency versus subset score is drawn.

4.3 Gene ontology (GO)

GO is a systematic way to describe gene and protein function. It comprises of three ontologies: molecular function, biological process and cellular component. In addition to Jaccard index and subset evaluation, the predicted clusters are also evaluated using GO terms for cellular component. This measures the localization coherence (lc) of the clusters which indicate the percentage of clusters which show some minimum percentage of proteins in the cluster that occur together in a cellular component. The intuition for using this is proteins that form complexes will seldom be in different cellular components. Thus, lc is also a measure of the quality of the predicted complexes (Liu et al, 2009).

4.4 Other evaluation method

In addition to evaluating predicted clusters using Jaccard index, subset evaluation and GO for cellular component, it will be useful to also evaluate predicted clusters based on pathway coherence. This is because proteins which form complexes tend to be in the same pathway and predicted clusters that are not found in the same cellular component may be investigated for pathway coherence. However, there was insufficient time to prepare comprehensive pathway information for yeast. While GO for biological process may be used to replace the evaluation method for pathway coherence, it is not as precise as pathway coherence.

5. Results and discussion

As shown in chapter 3, Experiment 1 and 2 determined the best parameters to use for CMC. This chapter presents and discusses the results for experiment 3 to 5. The result from experiment 3 proved that using PPIs from abstracts to enhance protein complex prediction is valid. The results from experiment 4 showed that noisy edges can be pruned by removing non-hub proteins. Lastly, the results from experiment 5 showed that augmentation of the PPI network is beneficial.

5.1 Results of experiment 3

In experiment 3, four different PPI networks were built and passed to CMC. Table 4 displays the result for experiment 3 under the match threshold of 0.5.

Method	Network size	Avg node degree	Number of clusters	Recall	Precision	Localization coherence (lc)
PPI network from real PPIs	1836	3.86	186	0.474	0.333	At least 69% of clusters show 86% lc
PPI network from abstracts (Dict)	2594	3.02	482	0.249	0.095	At least 66% of clusters show 78% lc
Combined network of real PPIs and Dict	3225	4.02	617	0.549	0.154	At least 66% of clusters show 84% lc
PPI network from abstracts (BN)	1283	1.53	138	0.061	0.065	At least 60% of clusters show 80% lc

Table 4: Recall, precision and localization coherence of protein complexes from 4 different PPI networks.

It can be observed that recall and precision of protein clusters which are predicted from PPIs from BN is extremely low. However, the localization coherence result shows that protein clusters that are predicted are relevant since at least 60% of the clusters have 80% of the proteins in the same cluster also occurring in the same cellular component. Hence, low precision is not a concern as it is most likely due to the incompleteness of protein complex data in the reference complexes AloyMIPS. The low recall is due to the PPI network being small and sparse as compared to the other networks. As the same set of abstracts is used for PPI extraction in BN and Dict, it is most likely that the BN method generated many false-negative PPIs.

Additionally, table 4 provides evidence that using a combined network to predict protein clusters produces better quality prediction than separate networks of real PPIs and abstract PPIs. This is because the average node degree for the combined network of real PPIs and Dict is higher than the individual networks. Moreover, the combined network shows an increase in recall. This means that PPIs from abstracts help to fill in the missing edges of the real PPI network to predict more protein clusters that match the reference complexes. We illustrate this with a visualization of a real complex which was only predicted after abstract PPIs are added into the network (Figure 16).

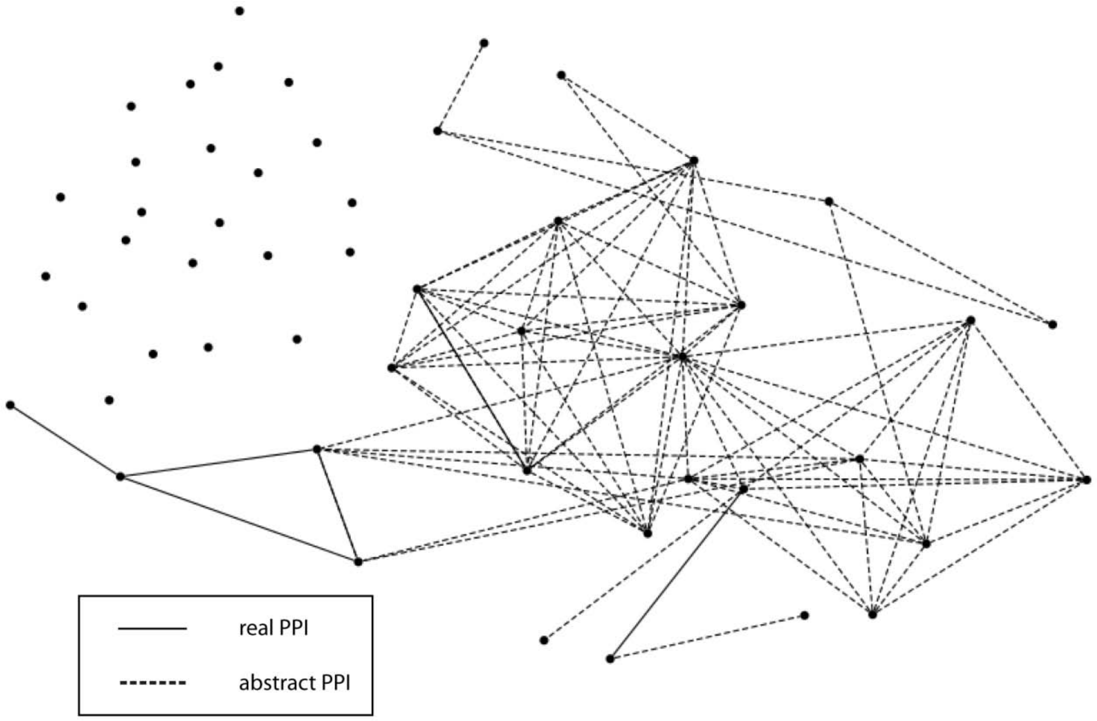


Figure 16: Graph of real complex 420.

Furthermore, the large increase in recall for the combined network of real PPIs and Dict suggests that the recall of the network from abstracts is likely to be limited by the number of abstracts that are analyzed. However, the combined network shows a decrease in precision as compared to the network derived from real PPIs. A possible reason for the decrease in precision is that edges that are extracted from abstracts contain a higher level of noise than those from real PPIs. Lower precision should not be a cause for alarm since the localization coherence results for combined network and network from real PPIs does not differ much even though there are 3 times more clusters that are predicted in the combined network.

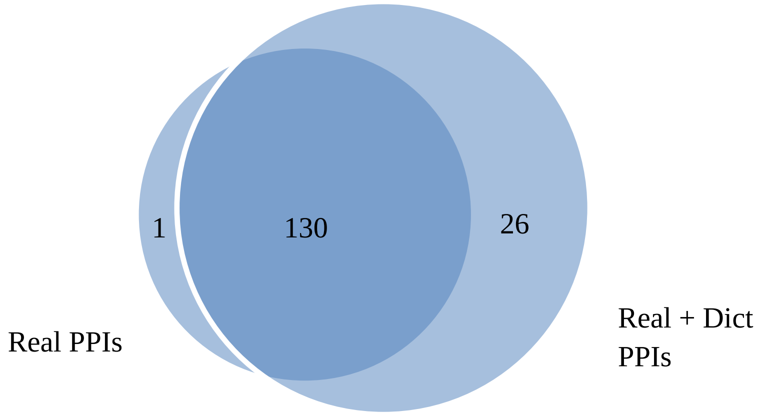


Figure 17: Venn diagram of correctly predicted clusters from 2 different PPI networks.

An analysis of the clusters showed that 130 clusters are predicted by both the combined network and network from real PPIs, 26 clusters are predicted by the combined network but not by the network from real PPIs, and 1 cluster was predicted by the network from real PPIs but not by the combined network. This suggest that the combined network is performing reasonably well (20% more correctly predicted clusters) as the predicted clusters are of comparable quality to those predicted based on real PPI only.

Further analysis of the PPIs that were extracted is done in the next section to confirm that using PPIs from abstracts to enhance the prediction of protein complexes is valid.

5.1.1 Comparison between real PPIs and PPIs from abstracts (Dict)

From the initial set of abstracts in section 3.1.2, 17413 abstracts contained at least a dictionary word and 25838 sentences contained at least a dictionary word. A total of 32497 PPIs were extracted from the abstracts. By comparing with 15900 real PPIs, 32493 abstract PPIs are not found in real PPIs while 15896 real PPIs are not found in abstract PPIs. These numbers suggest two things. Firstly, abstracts can potentially fill in PPIs which are missing from PPI databases and high-throughput experiments. This is provided that the PPIs that are extracted from abstracts are generally correct. Secondly, as many real PPIs in the databases used are not found among PPIs which are extracted from abstracts, it is likely that we have considered too few abstracts or the methods used for extracting PPIs from abstracts also needs improvement.

As mentioned, most abstract PPIs are not found in real PPIs and vice versa. Therefore, 161 randomly chosen abstract PPIs were manually checked with their abstracts. Of which, 21 PPIs were found to definitely not interact, the interactions of 45 PPIs were unsure and 95

PPIs were found to definitely interact. This implies that for the PPIs whose interaction status is clear, the odds that an edge derived from abstracts is a real PPI is better than 4:1. This provides good evidence that using and combining PPIs from abstracts to predict protein complexes is valid.

5.2 Results of experiment 4

The purpose of experiment 4 is to remove non-hub proteins which could potentially be noisy data which prevents the discovery of smaller and more accurate complexes.

Iteration	Network size	Avg node degree	Number of clusters	Recall	Precision	Localization coherence (lc)
0	3225	4.02	617	0.549	0.154	At least 66% show 84% lc
1	1514	3.34	617+163= 780	0.559	0.145	At least 69% show 84% lc
2	1339	3.42	780+29= 809	0.559	0.142	At least 69% show 84% lc
3	999	2.89	809+77= 886	0.563	0.132	At least 70% show 83% lc
4	901	2.88	886+30= 916	0.563	0.13	At least 71% show 84% lc
5	783	2.65	916+41= 957	0.563	0.126	At least 71% show 84% lc

Table 5: Recall, precision and localization coherence after different iterations of non-hub protein removal.

The result of the experiment shows that the removal of non-hub proteins leads to an increase in recall and localization coherence. This means that the experiment is generating a greater number of protein clusters which match the reference complexes and these clusters tend to be real complexes since they have high localization coherence. It is most likely that the iterated removal of non-hub proteins is causing clusters which were initially discarded to be found. The quality of the predicted clusters are visualised on a subset evaluation graph (Figure 18).

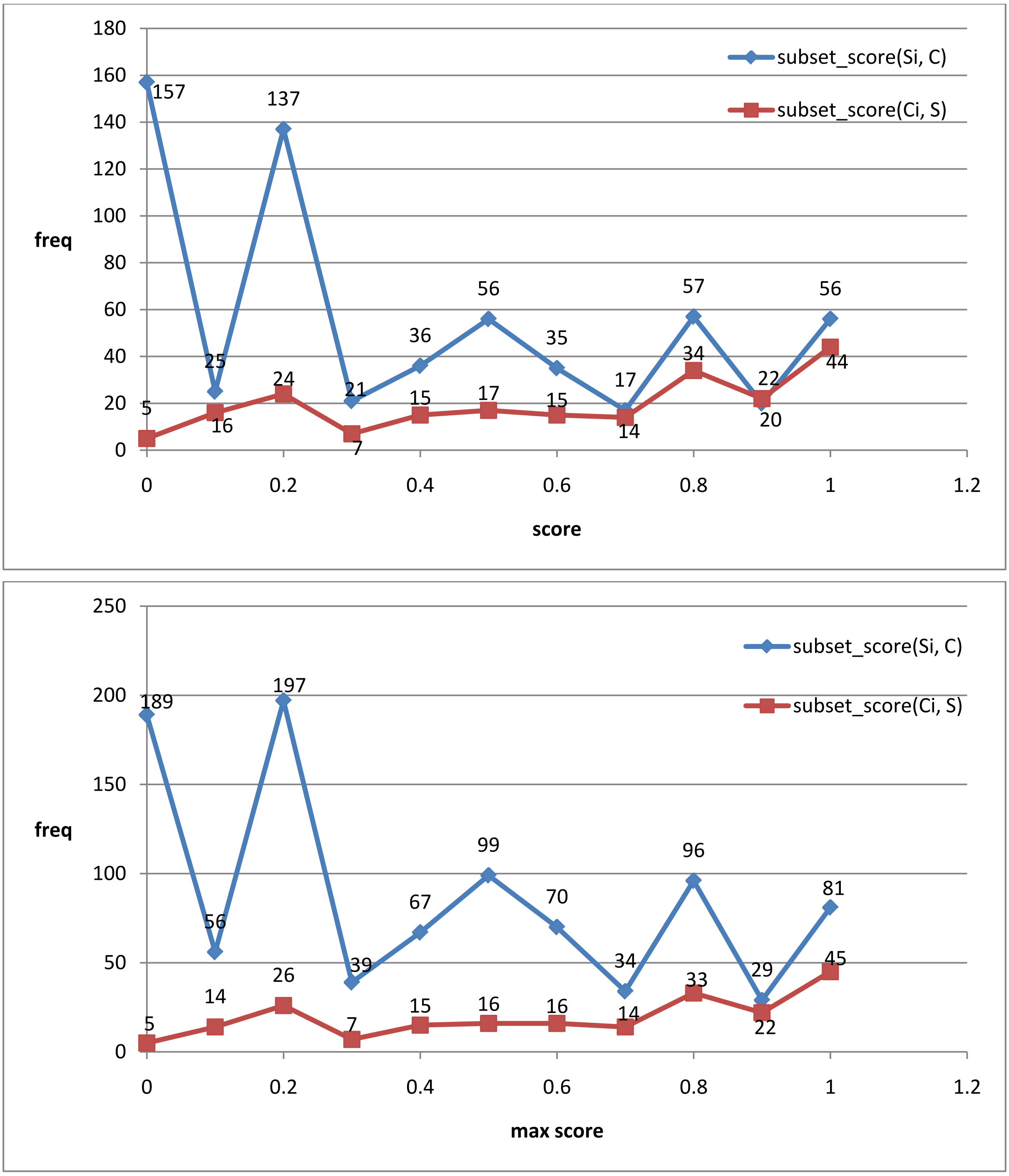


Figure 18: (top) subset evaluation graph of protein complexes before iterated removal of non-hub proteins. (bottom) Subset evaluation graph of protein complexes after 5 iterated removals of non-hub proteins.

By comparing both graphs in figure 18, it can be seen that there is a large increase in the number of complexes in subset_score(S_i, C) after iterated removal of non-hub proteins. A large increase in subset_score(S_i, C) means that there are more predicted clusters that have higher overlap with real complexes. By similarly comparing both graphs for subset_score(C_i, S), it can be observed that two complexes that have a maximum score of 0.1 before iterated removal improved to a maximum score of 0.2, one complex that has a maximum score of 0.5

before iterated removal improved to a maximum score of 0.6 and one complex that has a maximum score of 0.8 before iterated removal improved to a maximum score of 1. In short, a total of 4 complexes have improved their score while none have decreased. Therefore, the overall quality of the predicted complexes has improved.

Another observation is both graphs show that a large number of predicted clusters have a low score. For instance, in figure 18 (bottom), there are 189, 56 and 197 clusters that have a low score of 0, 0.1 and 0.2 respectively. The size of the clusters which produces the scores after modification was investigated by plotting a 3D graph of frequency versus score versus size of the clusters. The red lines refer to real complexes while the blue lines refer to predicted clusters.

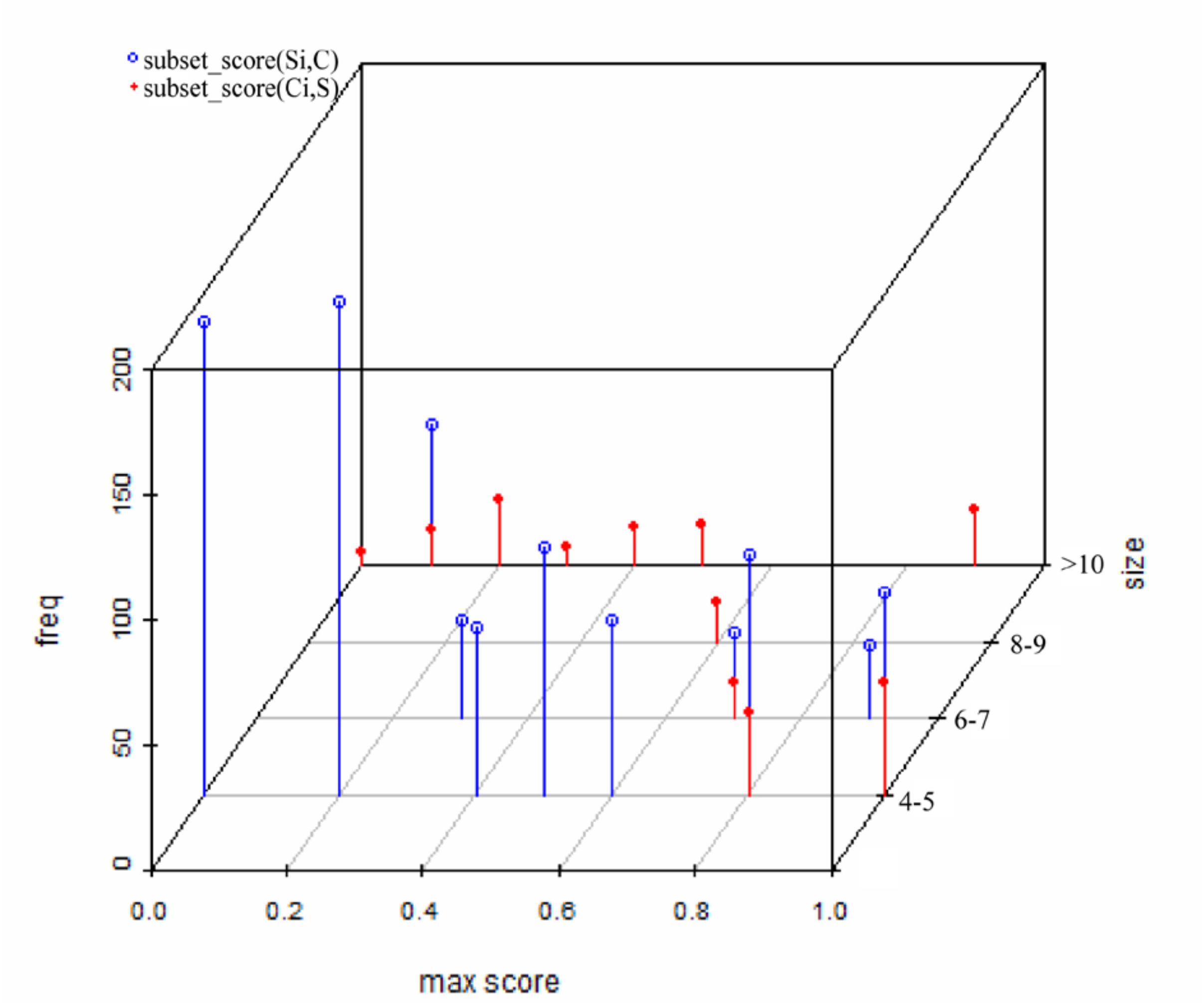


Figure 19: 3D graph of protein complexes after iterated removal of non-hub proteins.

The 3D graph for subset_score(S_i, C) shows that many of the small predicted clusters do not overlap real complexes. For example, for the complexes of size 4-5, 48% of the predicted clusters may have less than 20% overlap with real complexes. Nevertheless, for the 189 predicted clusters with a score of 0, a localization coherence check shows that at least 52% of the complexes show 83% localization coherence. This means that approximately half of the clusters are in the same cellular component and have a good chance of being real complexes even though they are not found in the reference set AloyMIPS. This may be due to the fact that the set of benchmark complexes are very incomplete.

The 3D graph for subset_score(C_i, S) shows that the smaller real complexes are generally well captured within large clusters but larger complexes with size greater than 10 are missed as they have a score of less than 0.5. A total of 5 complexes with a score of 0 were missed. The complexes were visualised and their edges were filled in based on the real PPI data mentioned in section 3.1.1. Two of complexes are presented for illustration.

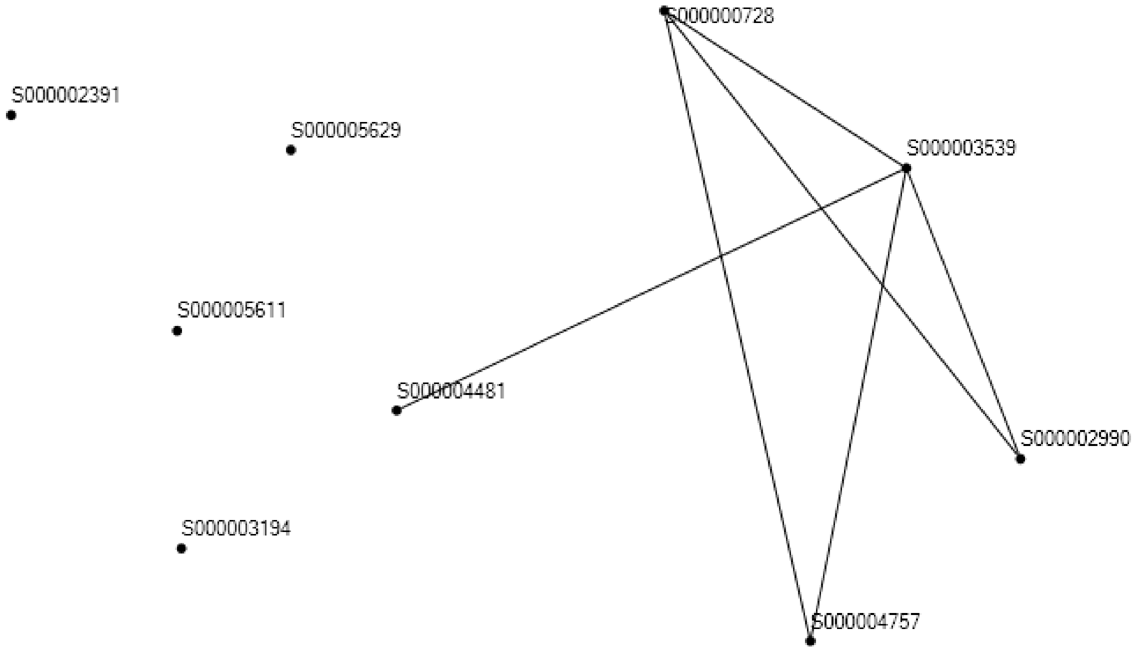


Figure 20: Graph of real complex 520.20.

Figure 20 show that complex 520.20 was not captured by CMC because it is not a clique. Moreover, the high number of isolated nodes suggests that the PPI data is very incomplete for the complex. This supports the limitation identified in section 2.3.4 that some complexes are not discovered because (i) using cliques as a basis to predict protein complexes is a stringent condition and (ii) many PPIs in the complex are probably missing in the PPI network since proteins in a real complex should not be disconnected from the complex.

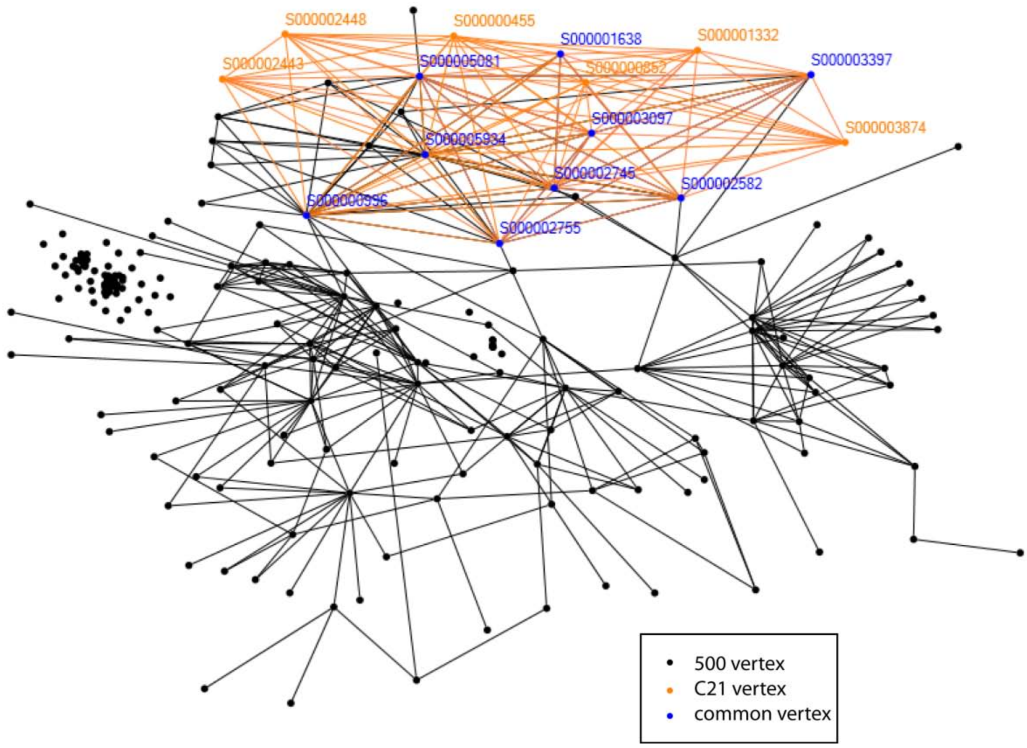


Figure 21: Graph of real complex 500 and predicted cluster C21.

In general, the structure of real complex 500 seems to comprise of 3 dense regions. While CMC is able to predict a part of the protein complex, the full complex remains undetected since the 3 parts are separated by sparse regions. The other 2 parts of the complex was probably not detected by CMC as they do not fulfil the clique criteria.

The visualisation of complex 500 also shows that information about many edges in the protein is not found. If information about these edges is known, CMC may be able to predict the complex.

5.3 Results from experiment 5

Experiment 5 augments the combined network of real and Dict PPIs with edges of proteins that are found in real complexes but not in the combined network. In order to evaluate if augmentation benefited protein complex prediction, the subset evaluation of the reference complexes before augmentation and after augmentation is done. The subset score equation in figure 15 was used to calculate subset_score(C_i, U) and subset_score(C_i, A), where Ci is the reference complex AloyMIPS, U is the set of clusters predicted from the combined network of real and Dict PPIs without augmentation and A is the set of clusters predicted from the augmented network. A graph of subset_score(C_i, U) versus subset_score(C_i, A) was plotted.

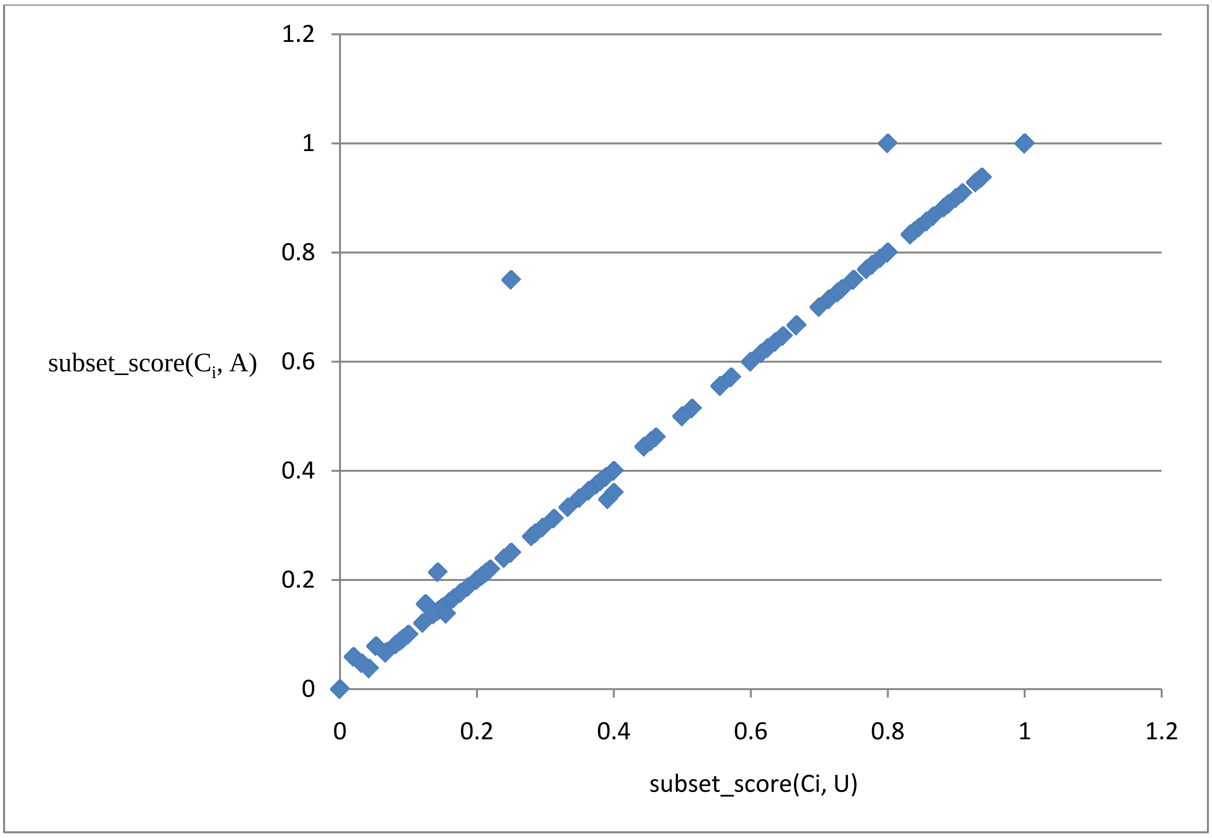


Figure 22: Graph of subset_score(C_i, U) versus subset_score(C_i, A).

From the graph, most of the points are aligned along the diagonal. This means that those complexes were unaffected by the augmentation. Data points that are above the diagonal indicate that the complexes benefitted from the augmentation. It can be seen that there are some complexes which benefitted a lot from augmentation. Hence, augmentation of the network is worth doing.

Due to the time constraints of this project, the augmenting set of abstracts is small. In addition, the querying program which searches for the abstracts which contain the proteins of interest retrieved significant false-positives abstracts which were only detected after the abstracts were tagged. In order to obtain abstracts that are more likely to contain relevant PPIs, we suggest using Bayesian inference based on the frequency of discriminating words (Marcotte et al, 2000) to determine whether a given paper discusses PPIs before retrieving the paper based on the occurrence of the proteins of interest. This method is currently used to help in the expansion of DIP.

6. Future work

This section suggests some future work for this project which an interested researcher may work on. Based on this project, there are three things worth implementing. They are the development of an evaluation method for pathway coherence, the prediction of complexes based on the largest k-connected sub-graphs and better selection of abstracts for augmentation.

6.1 Evaluation by pathway coherence

It was mentioned in section 4.4 that we were unable to evaluate the predicted clusters based on pathway coherence as we did not have sufficient time to collect and prepare comprehensive yeast pathway information. Such an evaluation method can be used to determine if clusters that are not found in the same cellular component is found in the same pathway. This is beneficial as the GO annotation is incomplete.

6.2 Predicting complexes based on largest k-connected sub-graphs

It was shown in our experiments that although the clique criterion was able to predict many real complexes, many complexes are also missed by CMC. When the missed complexes were visualised, they are indeed not cliques. In order to capture these missed complexes, we suggest the prediction of complexes from PPI network based on the largest k-connected sub-graphs. A k-connected sub-graph is a connected sub-graph with size greater than k and will remain connected after deleting k nodes from it.

6.3 Improving the selection of abstracts for augmentation

As mentioned in section 5.3, the selection of abstracts for augmentation could be more precise. Instead of using the names of the proteins of interest to retrieve abstracts, future work could consider detecting if an abstract discusses PPIs first by using Bayesian inference before looking for the abstracts that contain the protein names.

7. Conclusion

This project explores the usage of text mining to supplement PPI data in PPI network for protein complex prediction. We make the following specific contributions:

- Firstly, three rule-based methods of extracting PPIs from text were explored. They are co-occurrences of two proteins in the same sentence (Co), co-occurrences of two proteins in the same sentence together with a dictionary verb from a dictionary of four interaction word (interact, bind, complex, associate) (Dict) and using a trained Bayesian Network (BN) by Chowdhary et al. The results showed that the combined network of real PPIs and Dict fared better in predicting real complexes.
- Secondly, noisy edges were pruned away by removing non-hub proteins from the network iteratively. The iterative pruning led to the prediction of a greater number of complexes that were likely to be real.
- Lastly, the combined network of real PPIs and Dict was selectively augmented with new edges. The augmenting abstracts contain proteins which are found in the reference complexes but not in the predicted complexes. PPIs were extracted and augmented to the PPI network for protein complex prediction. The results showed that even with the small set of augmenting abstracts, there was an improvement in the prediction of some complexes.

8. References

Bader GD, Hogue CW. **An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks.** *Bioinformatics*, 4:2, January 2003.

Chowdhary, R., Zhang, J., Liu JS. **Bayesian Inference of Protein-protein Interactions from Biological Literature.** *Bioinformatics*, 25(12):1536--1542, June 2009.

Guimei Liu, Limsoon Wong, Hon Nian Chua. **Complex Discovery from Weighted PPI Networks**. *Bioinformatics*, 25(15):1891--1897, August 2009.

Kim S, Shin SY, Lee IH, Kim SJ, Sriram R, Zhang BT. PIE: an Online Prediction System for Protein-Protein Interactions from Text. *Nucleic Acids Research*, 36(Web server issue): W411 -- 415, July 2008

Krycer JR, Pang CN, Wilkins MR. (2008). **High Throughput Protein-Protein Interaction Data: Clues for the Architecture of Protein Complexes.** *Proteome Science*, 6:32, November 2008.

LI Zhihui, HYPERLINK "psZ/li-zhihui-hyp08.pdf""Pubmed Abstract Processing for Protein Function Prediction", Honours Year Project Report, Faculty of Science, National University of Singapore, April 2008.

LIM Junliang Kevin, HYPERLINK "psZ/kevin-lim-fyp09.pdf""Inferring Protein Function Module from Protein Interaction Information", Honours Year Project Report, School of Computing, National University of Singapore, April 2009.

Marcotte EM, Xenarios L and Eisenberg D. (2000) **Mining Literature for Protein-Protein Interactions.** *Bioinformatics*, 17(4):359—363, April 2001.

See-Kiong Ng, Soon-Heng Tan, “Discovering protein-protein interactions”, *JBCB*, 1(4):711-741, 2004.

Shoemaker BA, Pancheko AR. **Deciphering Protein-Protein Interactions. Part I. Experimental Techniques and Databases.** *PLoS Comput Biol*, 3(3): e42, March 2007.

van Haagen HHHBM, 't Hoen PAC, Botelho Bovo A, de Morrée A, van Mulligen EM, et al. **Novel Protein – Protein Interactions Inferred from Literature Context.** *PLoS ONE*, 4(11): e7894, November 2009.

Vancouver Prostrate Centre. (2005).
http://www.microarray.prostatecentre.com/Services_Proteins.htm

Wu, M., Li, X., Kwoh, C. (2004). Algorithms for Detecting Protein Complexes in PPI Networks: an evaluation study. Nanyang Technological University, School of Computer Engineering, Singapore Institute for Infocomm Research. Retrieved from http://www1.i2r.a-star.edu.sg/~xlli/publication/PRIB08.pdf

Zhou GD, Su J. **Exploring Deep Knowledge Resources in Biomedical Name Recognition.** *Proceedings of the Joint Workshop on Natural Language Processing of Biomedicine and its Applications (JNLPBA-2004)* 2004, 96-99.

Zhou D., Y. He, C.K. Kwoh. *Validating Text Mining Results on Protein-Protein Interactions using Gene Expression Profiles*. in *The International Conference on Biomedical and Pharmaceutical Engineering 2006 (ICBPE2006)* 2006. Singapore.