Functional Modules Detection from Protein Interaction Network using Graph-based Clustering

Afnizanfaizal Abdullah
Department of Software Engineering
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
+6017-3413411
afnizanfaizal@utm.my

Safaai Deris
Department of Software Engineering
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
+607-5533033
safaai@utm.my

ABSTRACT
The focus of understanding biological functions and processes had moved from study on individual genes and proteins to the study on the biological interactions that may form whole cellular systems. To meet this challenge, development of computational methods for predicting biological functions from genome wide cellular systems is necessary. Thus, the aim of this research is to construct computational model based on machine learning approach to select clusters in biological network of *Saccharomyces cerevisiae* that may correspond to biological functions. In this paper, we proposed a novel graph-based clustering approach for partitioning the protein interaction network into several functional modules. The significances of detected functional modules were evaluated by using protein complexes from MIPS database. Moreover, our algorithm had very low discard rate of detecting overlapping modules.

General Terms

Keywords
Functional modules, graph-based clustering algorithm, protein interaction network, overlapping modules

1. INTRODUCTION
Molecular biology is becoming a highly modular science in which functional modules are considered to be a critical level of biological organization [1]. These modules are entities that perform certain biological functions which relatively independent from each other [2]. Therefore, identifying these modules in biological network is crucial in understanding the organization and dynamics of cell functions [2,3,4]. However, it is a challenging task for identifying and selecting suitable functional modules from genome wide biological network such as protein interaction network.

These interaction networks are usually represented as graph theory where nodes represent the gene or proteins and edge represent the interaction between them. This approach had been applied to describe the network topological properties such as degree distribution, clustering coefficient and network diameter [5]. Hence, graph clustering algorithm is widely used for identifying functional modules from the biological network. By clustering the interaction network, the genes or proteins are divided into groups of common properties such as biological functions and processes based on the highly connected regions detected in the whole network.

The aim of this research is to construct computational model based on machine learning to extract clusters in *Saccharomyces cerevisiae* protein interaction network that may correspond to functional modules that will be used to predict functions. For achieving this aim, three objectives had been set which are (1) to investigate and identify features in protein interaction network features that correspond to biological functions; (2) to develop technique based on machine learning approach for selecting suitable features in protein interaction network that will be use for predicting biological functions and (3) to test and evaluate the performance of the proposed algorithm using precision and recall measure and validate the function prediction accuracy of the selected features.

This study will focus on predicting protein function from genome-scale protein interaction network of yeast *Saccharomyces cerevisiae*. Several studies addressed the confidence issue in the protein-protein interaction dataset of the model yeast *Saccharomyces cerevisiae* obtained from high-throughput experiments. In addition, there are various and diverse databases are available which offered researchers the full access in downloading the recent dataset. However, the dataset are also suffered with noise and partially available which may decrease the performance and accuracy of the computational methods. Therefore, the dataset used in this research is consisted of the unification of core and high quality protein-protein interaction data obtained from [15] via http://www.bioinfo.org.cn/MFGO/. The information of protein function is obtained from Gene Ontology (GO) to extract the biological processes and functions of the proteins listed in the dataset.

According to [19], in a protein interaction network, a set of genes which related to a phenotype could be characterized by projecting them onto the network and testing their network properties. Some general statistics such as average minimal path length between nodes, clustering coefficient and degree distributions are measured to study the inherent of these properties [20]. However, these global metrics provide limited information about networks as a whole and also unable to tell about individual genes and proteins and their relationships [40]. Thus, according to [21],
identifying the modular structure of the biological networks is important in order to understand the organization and interaction of the cellular processes.

Clustering the large-scale protein interaction network is used to extract modules with highly connected proteins which majorly shared common functions. These modules, which called as functional modules, will be used to predict functions to uncharacterized proteins. Functional modules are defined as sets of interacted proteins which shared common functions in a single process [20]. They are discrete entities which functions are separable from other modules [27]. According to [22], the prevalence of the modularity paradigm in molecular cell biology has led to the extensive use of modules in prediction of molecular function.

Therefore, functional modules identification is majorly applied to partition proteins accordingly to their functions since the modules are contained a subset of proteins which closely interacted to each other. The coherent group of proteins is firstly identified then functions of the proteins in the particular group are assigned. However, a given component in a module may belong to different other modules at different times and the module’s function can be affected by signals from other modules [40]. Furthermore, it is possible that a module which had many edges that connected to other modules may be merged into a single cluster. The approaches differ mainly in their module detection techniques and once a module is obtained, usually, simple methods are used to predict function within the module [22].

Furthermore, the researchers found that in this study, clustering protein interaction network for protein function prediction may lead to integration of protein-protein interactions data with gene expression data [23][24] which had high capability in identifying diseases and pathogens [19] and classifying cancer metastasis [25]. Thus, the researchers believed that this work is important and highly significant for further study in the respective area of interest.

2. LITERATURE REVIEW

Advances in molecular biology and sequencing technologies had contributed to the increasing number of accumulated data from high-throughput experiments and this had lead to the completion of genome sequencing in several model organisms. Therefore, the focus of understanding biological functions and processes had moved from study on individual genes and proteins to the study on the biological interactions that may form whole cellular systems.

2.1 Overview of Molecular Biology

Molecular biology is the study of life at a molecular level of resolution [26]. This field of study concerns with understanding the interactions between various systems of a cell, including the interaction of deoxyribose nucleic acid (DNA), ribosome nucleic acid (RNA) and protein and the regulation of the interactions. The central dogma of molecular biology outlines the flow of information that is stored in a gene, transcribed into RNA and finally translated into protein [27]. The DNA works as templates for the RNA transcription. Then, RNA transfers the genetic information to the ribosome where it will be translated into protein. Each step of the dogma is accompanied by recent technological innovation that allow genome-wide analysis and hypotheses can be formulated and revised for the discovering the detail relationships between genotype and phenotype purposes, which therefore uncovering the molecular biology of certain organisms [27].

2.1.1 The DNA

The DNA molecules contain the genetic instructions which specifying the development of all life forms. In prokaryotes cells such as bacteria, DNA is not separated from the cytoplasm. However, in eukaryotic cells such as animals and plants, it located in the cell nucleus. The DNA consists of genes which responsible for expressing and passing on single characteristic to the new generation. During cell division, the DNA is replicated and can be transmitted to offspring in the reproduction process. The entire DNA of an organism is called genome. Figure 1 shows the structure of DNA.

![Figure 1. The DNA structure (image adapted from National Human Genome Research Institute, retrieved from http://images2.clinicaltools.com/images/gene/dna.jpg).](http://images2.clinicaltools.com/images/gene/dna.jpg)

DNA is the fundamental of the molecular biology. The information stored in DNA is used to create messenger RNA (mRNA) by transcription process. In this process, a protein called RNA polymerase is used to copy the DNA into mRNA. Nucleotide sequence in mRNA is converted to amino acid sequence through translation process which in the end constructs the proteins. A gene that transcribed and used for proteins synthesis is expressed whereas plays the fundamental element in the cellular processes.

2.1.2 The RNA

RNA is a nucleic acid which consists of a long chain of nucleotide units. It has very similar structure compare to DNA but differs in few important details. The RNA has the base uracil rather than thymine which is present in the DNA. In chemistry, the RNA contains ribose while the DNA contains deoxyribose. Furthermore, the RNA is usually found in a single stranded. The functional form of RNA molecules requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements or hydrogen bonds. Figure 2 shows the differences between RNA and DNA.
The RNA plays many roles in biological process. For instance, messenger RNA (mRNA) works as information carrier from DNA, transfer RNA (tRNA) transport a specific amino acid to a growing polypeptide chain during translation and ribosomal RNA (rRNA) is the catalytic component of the protein ribosomes. In molecular biology, RNA is transcribed from DNA by RNA polymerases enzymes and further processed by other enzymes. Then, mRNA carries information from DNA to ribosomes. The ribosomes work as a molecular machine that read mRNA and translate the information carried to the proteins.

2.1.3 The Proteins
Proteins are polymer chains composed of 20 different kinds of amino acids. The amino acids are linked together by peptide bonds. The amino acids are small molecules that share a common motif, presented as three substitute chemical groups which arranged around a central carbon atom. One of the substitute groups is always an amino group and another is always carboxyl group. Proteins can fold into very specific three-dimensional shapes that are dependent on their amino acid sequences. Thus, the amino acid sequence determines the shape of the protein and the shape determines the function.

Proteins have different tasks in different components or compartments of the cell [26]. The proteins can be classified according to its function. The enzymes are largest class which catalyzing chemical reaction. Some enzymes act on other proteins to add and remove chemical group. On the other hand, many proteins are involved in the process of cell signaling and signal transduction. Insulin, for example, is extracellular proteins that transmit a signal from the cell which they synthesized to other cells in distance tissues. Antibodies, in addition, are protein component in adaptive immune system. Figure 3 shows an example of protein.
The biological networks models usually represented as a graph theory which nodes are correspond to the genes or proteins and the edges correspond to the physical or functional interactions between the nodes [22][32]. The edges may be represented as directed or undirected. The directed edges are suitable for representing the flow of material from a substrate to a product in a reaction or the flow of information from transcription factor to the gene whose transcription it regulates such as gene regulatory networks. Meanwhile, undirected edges are usually used to represent mutual or physical interactions such as protein-protein interactions [32]. Moreover, these graphs can be incremented by assigning various attributes to the nodes and edges [32] for instance different functions of nodes such as protein interaction network, signed edges for gene regulatory network and weights confidence levels and strength like in gene networks and metabolic networks.

Topological analysis of biological networks is a great topic interest in bioinformatics and systems biology currently and this analysis provides quantitative penetration that characterize various biological systems [29][30]. In biological networks, there are four basic global network measures that allow the comparison of different complex biological networks: clustering coefficient, average degree, mean path length and network centrality. Clustering coefficient measures the local network coherence around a node in the networks or sub-networks. However, this coefficient may be use for global measure by averaging the coefficient values of all nodes in the certain network [1][33]. While, average degree measures the average connectivity of each node. The mean path length counts the average number of links that pass through to travel every two nodes. This measure offers the overall navigability of the network [29]. On the other hand, the network centrality is a quantitative measure of the position of a node which relative to other nodes which essential for identifying the degree of dependency of each node in the certain network.

2.3 Modularity in Biological Network

Theoretically, the decomposition of a large network into modules is an effective way to understand the architecture and behavior of the complex network [30]. The biological networks had been considered to be modular which composed of separate sub-networks that functionally correspond to the specific biological functions [1][20][29][30][31][32]. The modularity of the network may originated from cellular localization or specific protein interactions which involved groups of components that work in a cooperative manner to achieve some defined functions [1].

In a general network representation, modules consist of highly interconnected groups of nodes [29] which every node may share common biological features together. This is due to the fact that, in most cellular metabolisms, a fully interconnected biochemical network consists of hundreds metabolic substrates are densely integrated through biochemical reactions [34]. On the other hand, a subunit of a complex network that appeared to be significantly repeated called network motif that may be a subunit of a module is used [30]. Therefore, clarifying the role of the motifs and identifying the modules in biological network are in major focus in understanding the biological network especially for predicting molecular functions, biological processes and signal pathways.

2.4 Biological Network Modules Detection

The study of biological functions from the whole genome data had become impracticable as the number of distinct sub-networks increase exponentially with the number of nodes that are available in the genomic network. Therefore, identifying groups of biology components such as genes, proteins and metabolites with high interconnectivity that contributed to common functions which called functional modules directly from whole genome had become an alternative approach to overcome the problem. The presence of these modules can be inferred when the identified modules have a high clustering coefficient [20] and focuses predominantly on protein interaction network [31].

Thus, clustering analysis had been suggested to be an obvious option of methodology for extracting the functional modules from biological networks [8]. Clustering works as an unsupervised algorithm that groups or partitions objects or components which share common discrete and measurable features [8][35]. Moreover, a variety of clustering algorithms had been developed and successfully used in diverse fields [8]. In functional genomics, these algorithms had been used in dealing with large scale dataset in biology including gene expression analysis [36], protein families detection [37] and protein sequence data analysis [38].

Many clustering algorithms cannot use the information inherent in weightings on edges in the graphs (Hallinan, 2008). Protein-protein interactions data are usually represented as unweighted graph which consisted of nodes as proteins and edges as interactions between the proteins. Hence, numerous graph clustering algorithms had been applied to the graph representing the binary interaction [22]. Most of the findings in the clustering protein interaction network into functional modules are designed based on partitioning graph theory and identifying clusters with highly connected proteins. In turn, this also will lead researchers to find effective ways to identify functionally relevant modules [39].

In the last few years, several methods had been applied for identifying functional modules by detecting densely connected regions from protein interaction networks [6][7][8]. Consequently, [9] had discovered 48 functional modules using a spectral analysis method and [6] found approximately 50 modules by applying a combination of three methods including superparamagnetic clustering and Monte Carlo simulation. However, most methods are involved in partitioning the interaction network into modules that consist of genes or proteins that belongs to only one specific module [10]. As a result, overlapping modules can not be found and detected directly by those algorithms.

Therefore, [11] proposed an edge betweenness method for detecting overlapping modules. Compared to existing method that used strongly connected communities core, this method used information about edge centrality to detect community peripheries. However, algorithm used in this method is time ineffective especially when dealing with larger network. Thus, [12] introduced a network clustering method based on clique percolation for revealing overlapping module from larger and more complex networks. Nevertheless, the proposed method may be restricted in finding clique as its priority while detecting the modules. To overcome this problem, [10] improved the clique
percolation method with implementation of line graph transformation. This proposed method is deterministic which the resulting modules will be determined by simple processing criteria. Moreover, [13] designed a simple and effective local optimization searching algorithm which treats each group of highly overlapping maximal cliques as a clustering core. On the other hand, [14] introduced graph split and reduction technique before applying proposed graph-based clustering algorithm.

3. RESEARCH METHODOLOGY
This research is basically an applied scientific research which based on problem domain. In this research, a research operational framework as illustrated in Figure 5 is used. Since unsupervised learning algorithm has involved in this research, two types of datasets have been used which are protein-protein interaction dataset as testing data and protein complexes information as validation data. For evaluation purpose, we used precision, recall and p-value to measure the performance.

![Research Operational Framework](image)

Figure 5. Research operational framework.

### 3.1 Protein-Protein Interaction Dataset
Several studies addressed the confidence issue in the protein-protein interaction dataset of the model yeasts *Saccharomyces cerevisiae* obtained from high-throughput experiments. Furthermore, there are various databases that supplied this dataset. Nevertheless, the dataset are suffered with noise and partially available which had decreased the performance and accuracy of the computational methods. Thus, in this work, we used three benchmark dataset from [41], [42] and [43]. Table 1 shows the information of the datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Proteins</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>[41]</td>
<td>779</td>
<td>787</td>
</tr>
<tr>
<td>[42]</td>
<td>926</td>
<td>865</td>
</tr>
<tr>
<td>[43]</td>
<td>1,430</td>
<td>6,531</td>
</tr>
</tbody>
</table>

### 3.2 Protein Complexes Information
As validation, the information of *Saccharomyces cerevisiae* protein complexes from MIPS is downloaded from [http://mips.gsf.de/genre/proj/yeast/Search/Catalogs/catalog.jsp](http://mips.gsf.de/genre/proj/yeast/Search/Catalogs/catalog.jsp) in order to verify the reliability of modules generated from the proposed algorithm. There are 68 protein complexes included in the database which include 2733 proteins assigned to the complexes while 3273 proteins are listed as unsigned.

### 3.3 Performance Measure
The research measure is evaluated based on the discussion from [22]. The researchers had discussed that the evaluation can be made by measuring the precision and recall of an annotation which computed in a leave-one-out setting and also considering the multiple annotations per protein. In this research, the precision and recall is used for estimating the performance of algorithm for detecting functional modules. Precision is a measure that calculates the fraction of true positive predictions out of all the positive prediction and recall calculates the fraction of true positive predictions out of all true predictions (Li et al., 2008). These measures are defined as below:

\[
\text{Precision} = \frac{\sum k_i}{\sum m_i} \quad \text{Recall} = \frac{\sum k_i}{\sum n_i}
\]

where let \(n_i\) be the number of known functions for protein \(i\) and let \(m_i\) be the number of predicted functions for the protein when hiding its true annotations and let \(k_i\) the overlap between the two sets. In order to evaluate the performance of the proposed method, the p-value measure is used. This measure indicates the probability of identifying number of proteins with an annotation by a GO term among total number of proteins in the network. The hypergeometric distribution of the measure is shown as following (Hayashida et al., 2006):

\[
P = 1 - \sum_{i=3}^{k} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} = \sum_{i=k}^{n} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}
\]

where \(N\) is the total of proteins, \(M\) is the total number of proteins annotated by the specific GO term, \(n\) is the number of proteins in cluster and \(k\) is the number of proteins annotated by specific GO term in the cluster. In this research, an online GO term tool called GO-TermFinder ([http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl](http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl)) is used.
4. DESIGN OF PROPOSED METHOD

In this work, a novel graph-based clustering algorithm called Local Neighbor and Overlap Modules Clustering (LNOMC) algorithm is introduced for partitioning Saccharomyces cerevisiae protein interaction network into functional modules. This proposed algorithm has successfully generated modules that consist of highly interacted proteins. Protein complexes from Munich Information of Protein Sequence (MIPS) database is used for verifying the result. Also, the algorithm had successfully detected overlap structures of among these modules. Figure 6 illustrates the framework of the proposed method.

4.1 Graph Modeling

In the framework, we modeled the protein interaction network with undirected graph, G(V,E), where V represents a set of proteins and E represents a set of interactions between proteins. Figure 7 illustrates an example of the process. The dataset formatted by a number of rows and two columns which represent the interaction and interacting proteins respectively. Firstly, the protein-protein interaction dataset is read and each protein in the list is transformed into node and every two nodes will be connected by an edge by the corresponding row in the dataset.

4.2 LNOMC Algorithm

The LNOMC is applied for detecting modules by splitting the whole graph into several sub-networks. This algorithm composed of three major steps: weighting edges, search nearest neighbor proteins and graph splitting. Each step is described in detailed by following subsections.

4.2.1 Weighting Edges

Every node in the dataset will be weighted based on the degree estimation. The degree estimation will be assign by the number of proteins interacted with it. Then, this value will be transformed into the edge weight which reflect the number of interactions occur that involved by the edge. Figure 8 shows the algorithm of the informative interactions selection.

4.2.2 Search Nearest Neighbor Proteins

The algorithm will scanned through the interaction dataset in order to find proteins that interacted with each selected interaction. Then, the algorithm will also find the interactions among searched proteins through the interaction set, E, iteratively. This process may optimize the dense interactions of each generated modules while considering overlap structures among modules. Figure 8 shows the algorithm for searching nearest neighbor proteins.

Figure 6. Proposed method framework.

Figure 7. Example of graph modeling from protein interaction network dataset.

Figure 8. Algorithm for weighting edges.

Figure 9. Algorithm for searching nearest neighbor proteins.
4.2.3 Graph Splitting
After nearest neighbor proteins are found, the algorithm will split the newly found sub-network from the whole network. The algorithm will repeat the previous step (4.2.2) and split the graph until all informative edges have been assigned. Figure 10 shows an example of two sub-networks which have been split by the algorithm.

![Figure 10. An example of two sub-networks which have been split by the algorithm.](image)

4.3 Functional Modules Detection
In the end, the method will produce a set of detected functional modules which consist of partitioned sub-networks from the whole protein interaction network. As for validation, proteins that contained in these modules will be mapped to set of proteins in protein complexes information obtained from MIPS. In addition, GO terms will be used to verify the biological significance of the detected modules.

5. EXPERIMENTAL RESULT
We have conducted an experiment to measure the performance of the proposed method using protein-protein interaction datasets presented in 3.1 as our testing data while protein complexes information presented in 3.2 as our validation data. Parameter cut-off has been used in order to validate the scalability of the method. We also presented the comparison of proposed method with previous methods, Markov clustering (MCL) and Restricted Neighborhood Searching Clustering (RNSC). We measured the detected modules by their network topology using global network measures and computed the biological significance using p-value measure.

5.1 Scalability Evaluation
We experimented the proposed method by changing different parameter cut-offs in order to survey the scalability of the method. In this method, parameter cut-off is a value that we used to trim out the significance number of modules detected to the actual number of protein degree value. This parameter is important to scale the number of detected modules. In this experiment, we used proteins with degree with a range of 1 to 10 for each dataset. Figure 11 shows the effect of parameter cut-off to the average clustering coefficient of each dataset. The increasing average clustering coefficient score through number of cut-off value shows that our method has high scalability through increasing number of proteins in dataset.

![Figure 11. Effect of parameter cut-off of degree, \(k\), to average clustering coefficient, CC.](image)

5.2 Comparison to Existing Methods
We observed the performance of proposed method with two leading existing methods: MCL and RNSC. In the experiment, we test the same datasets to the existing methods and found that our method has produced the smallest number of modules by using suitable parameter cut-off according to the number of proteins in each dataset. This reflected that our method produced better cluster compared to MCL and RNSC whereas the existing methods identify small number of proteins within each cluster (average of 5 proteins per cluster). Moreover, MCL and RNSC failed to detect overlap modules. Table 2 shows the comparison of these methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MCL</th>
<th>RNSC</th>
<th>Our Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[41]</td>
<td>221</td>
<td>423</td>
<td>10*</td>
</tr>
<tr>
<td>[42]</td>
<td>288</td>
<td>491</td>
<td>10*</td>
</tr>
<tr>
<td>[43]</td>
<td>189</td>
<td>474</td>
<td>75**</td>
</tr>
</tbody>
</table>

* Parameter cut-off, \(k \geq 10\)
** Parameter cut-off, \(k \geq 30\)

5.3 Network Topology Measure
We examined the network topology of detected modules based on the degree distribution and average clustering coefficient distribution using parameter cut-off, \(k \geq 30\) for datasets [43]. This is due to the largest number of proteins in the dataset. Figure 12 illustrates these measures which reflected that the detected modules consist of proteins that topologically significance.
function. The p-value measure is used to validate 10 randomly chosen modules from 75 modules detected by our proposed method as shown in Table 3. In this result, the lowest p-value shows the most significant biological function to the corresponding module.

We also observed the biological significance of detected modules. In this experiment, we had set different value of degree, \( k \), as cut-off parameter to measure the scalability of the method. The result shows that this method has high potential in dealing with larger dataset. This has been proven by the used of three different datasets which have different number of interactions. Dataset from [43], which has the largest number of interaction proteins showed increasing number of average clustering coefficient through increasing number of cut-off parameter. This reflected that higher cut-off value may detect more significant modules.

Therefore, we used the result of functional modules detection by our method to compare with two existing methods: MCL and RNNSC. Similar to our method, both existing method also used local search to detect functional modules. However, these methods had obviously neglected functional interaction between different detected modules. This may lead to irrelevant to biology nature where each module have to interact each other in order to perform different biological functions. In this observation, we identified that the existing methods had score large number of modules detected but consist of small number of protein members in each modules. This had reflected computationally ineffective due to fact that each detected modules may consist a significantly large number of elements in order to show that the clustering method had successfully grouped the elements into relevant clusters. On the other hand, our method had showed smallest number of modules detected and highest average number of protein members.

In order to evaluate the network topology structure of detected modules, we used global network measures, degree distribution and average clustering coefficient distribution. The result showed that the detected modules followed power-law distribution, \( P(k) \sim k^{-\gamma} \), which reflect that the detected modules presented in hierarchical manner and built in scale-free structure where only few hub nodes are available. This has been proven by the high coefficient of determination, \( r^2 \), of protein members within the modules while low score for members of different modules. In this evaluation, we used dataset from [43] since this dataset is the largest dataset compared to datasets from [41] and [42].

6. DISCUSSION

In this work, we proposed new graph clustering method using novel LNOMC algorithm to detect functional modules from genome-wide \textit{Saccharomyces cerevisiae} protein interaction network. In general, the method is composed of three major steps which are graph modeling, LNOMC algorithm and modules detection. Firstly, the method transformed the protein-protein interaction data into graph modeling where nodes represent proteins and edges represent interaction between proteins. Then, LNOMC algorithm is used to weight the edges according to the degree of each protein node, search nearest neighbor proteins and split the whole network into several sub-networks which called modules. The major contribution of the proposed method is we used the degree of each protein and converted them to edge weight. This method is not only focused on detecting functional modules but also consider interactions or overlaps of different modules, which distinct to other existing methods.

We had conducted experiment to evaluate the performance of the proposed method using protein-protein interaction datasets obtained from [41], [42] and [43] which had been widely used in similar researches. In this experiment, we had set different value of degree, \( k \), as cut-off parameter to measure the scalability of the method. The result shows that this method has high potential in dealing with larger dataset. This has been proven by the used of three different datasets which have different number of interactions. Dataset from [43], which has the largest number of interaction proteins showed increasing number of average clustering coefficient through increasing number of cut-off parameter. This reflected that higher cut-off value may detect more significant modules.

Table 3. Biological function of detected modules from [43] dataset.

\begin{tabular}{|c|c|c|}
\hline
Module & \textit{GO}-term & P-value \\
\hline
8 & RNA metabolic process & 1.0544\times10^{-30} \\
9 & RNA metabolic process & 7.9579\times10^{-30} \\
22 & RNA metabolic process & 9.6041\times10^{-24} \\
315 & nucleolus & 1.5733\times10^{-25} \\
490 & ribosome & 5.3081\times10^{-23} \\
515 & ribosome & 1.6958\times10^{-29} \\
771 & Transcriptional regulatory activity & 1.9450\times10^{-23} \\
843 & protein catabolic process & 5.0511\times10^{-31} \\
1093 & ribosome biogenesis & 6.4430\times10^{-31} \\
1163 & ribosome biogenesis & 9.7081\times10^{-24} \\
\hline
\end{tabular}
Lastly, we surveyed the detected modules to biological function significance using GO terms. In our evaluation, we calculated p-value of 10 randomly selected modules out of 75 modules detected in [43] dataset with cut-off parameter, $k \geq 30$. The p-value is measured by Cytoscape tool which used hypergeometric test with Benjamin and Hochberg False Discovery Rate correction. The significance level of 0.05 is chosen and we had test cluster versus whole annotation from GO Slim yeast. As a result, we found that three modules shared RNA metabolic process, two modules annotated as ribosome and two modules as ribosome biogenesis. Meanwhile, three other modules are inferred as nucleolus, transcriptional regulatory and protein catabolic process respectively. More than one modules shared same annotation showed that these modules may overlap each other.

However, in our observation, we had examined that there are exist several limitation in applying this proposed method to the genome-wide protein interaction network. One of the limitations is we used benchmark dataset instead of raw dataset with high false positive rate. Therefore, there is no guarantee of producing good result when applying this method to such datasets. In additional, a module may suffer subset interactions to other modules. In other words, there may be exist one module which completely available to other modules. So, this method not only captures interactions among modules but also identify redundant modules. Moreover, the cut-off parameter used in this method had been manipulated manually without any automated mechanisms that can set the value to optimal state.

7. CONCLUSION
In this work, we proposed a method using novel graph-clustering algorithm called LNOMC algorithm for partitioning the protein interaction network into functional modules which also consider overlap structures that appeared among modules. However, the proposed method suffered several limitations. First, the method had been applied with small and high quality dataset and not been test with bigger and raw dataset yet. Even though the used of high quality dataset showed remarkable result, there is no guarantee of producing good result when dealing with raw dataset. Moreover, the several modules yield subset interaction with other modules. In other words, a module may completely available inside other modules. In addition, cut-off parameter can not be optimized automatically. Therefore, further study has to be conducted to overcome these problems.

8. ACKNOWLEDGMENTS
Our thanks to all members of Artificial Intelligence and Bioinformatics Laboratory, academic staffs of Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, especially Dr. Siti Zaiton and Associate Professor Dr. Ito Wasito and also to all friends for giving their supports.

9. REFERENCES


