Clustering of protein-protein interactions (PPI) and gene ontology molecular function using Markov clustering and fuzzy K partite algorithm

To cite this article: W A Kusuma et al 2019 IOP Conf. Ser.: Earth Environ. Sci. 299 012034

View the article online for updates and enhancements.
Clustering of protein-protein interactions (PPI) and gene ontology molecular function using Markov clustering and fuzzy K partite algorithm

Wisnu Ananta Kusuma1,2, Hafidlotul Fatimah Ahmad1 and Mochamad Suryono1

1Department of Computer Science, IPB University, Bogor 16680, Indonesia
2Trophical Biopharmaca Research Centre, IPB University, Bogor, 16128, Indonesia

*Email: ananta@apps.ipb.ac.id

Abstract. Functional disorders of proteins in the human body can cause a certain disease. The function and role of the protein are represented by Gene Ontology (GO). In this study, the GO molecular function was used to enrich the analysis of protein-protein interaction (PPI). The relationship between PPI and GO molecular function was represented in a bipartite graph. In the pre-processing step, the PPI network was reduced using the Markov clustering algorithm to obtain the group of proteins with the highest modularity score. The fuzzy k partite algorithm was used to cluster the PPI network and GO molecular functions into several groups. The result of Markov clustering showed the accuracy of 84.6% compared to that of the same algorithm using the GIANT package on Cytoscape applications, one of the popular software for network analysis. Proteins obtained from Markov clustering results were used as inputs to obtain their related GO molecular function. Their relationship was represented as a bipartite graph which is used as an input for the fuzzy k partite algorithm. With the dataset of Diabetes Mellitus type II, the results of Markov clustering showed that there were 117 proteins and 328 related GO molecular function. With fuzzy k partite algorithm, the minimum cost for the bipartite graph is 594.175 at the 20 clusters of proteins and 29 clusters of GO molecular functions.

1. Introduction
Functional disorders of proteins in the human body can cause certain diseases. Diabetes Mellitus type II (DM) is one of the diseases caused by functional disorders. According to WHO [1], type II DM is a metabolic disorder characterized by elevated levels glucose in the blood (hyperglycemia), which caused chronic complications triggered by abnormal insulin secretion, insulin action or both. Moreover, Bordonaro [2] stated that type II DM is characterized by several characteristics namely, the insulin hormone no longer works properly, the cells fail to respond normally to the hormone insulin, and elevated blood sugar levels compared to normal conditions (hyperglycemia).

In carrying out its functions, proteins interact with each other. This interaction between proteins can be represented in the form of networks (graph). The network of interactions between proteins serves as a powerful tool for revealing the function of proteins, disease genes and the relationship between diseases [3]. Computation approach for identifying significant proteins associated with a disease is proposed as solutions against experimental constraint. One of them is by constructing and analyzing network topology of protein-protein interaction [4][5]. This can be done because of the availability of protein-protein interaction data in large numbers and the technology advances of high throughput [4].
The other alternative is the clustering technique that can be used to reduce the size of PPI and to find the significant proteins so that the drug formulation process becomes easier. Singh and Singh [6] conducted a protein cluster analysis based on a bipartite graph that includes a protein sequence and Derived Sequence Features (SDF) for each sequence of proteins used to predict protein class in humans. However, this study has not involved the ontology (GO) gene information. GO is a framework for biological models with controlled vocabulary hierarchies. GO defines the concepts used to describe gene functions. GO defined into 3 aspects, namely molecular functions, cellular components, and processes biological. GO is used because it was a major bioinformatics initiative to unify the representation of genes. One protein consists of several of GO, it is possible for more than one protein has the same GO.

In this paper, we implement a fuzzy k partite algorithm to obtain significant protein. Significant protein obtained from protein-protein interaction (PPI). Protein data that has a direct association with type II DM is used as input. Fuzzy k partite is employed to clustering between the proteins and Gene Ontology (GO) molecular function. This method is used because it can acquire a separate cluster between PPI and GO molecular function, it also allows the detection of overlapping clusters [7].

2. Materials
This research used both public domain databases: STRING and UniProt. STRING is a database of known and predicted protein-protein interactions. The STRING database currently covers 9,643,763 proteins from 2,031 organisms. We select humans PPI that had a direct association with type II DM by input from the research of Usman [8]. The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. We collect GO molecular function data from UniProt based on proteins that had been reduced in the preprocessing stage.

3. Methods
This research is divided into three main steps including collecting PPI data, preprocessing PPI data, and clustering using fuzzy k partite technique. Data used in this research was protein data that had a direct association with type II DM, and molecular function data from proteins. The whole stages of the research can be seen in Figure 1.

![Figure 1. Research stages.](image)

3.1. Collecting PPI Data
Data used in this research was (1) proteins data that had direct association with type II DM obtained from Usman [8], (2) Interaction proteins data from STRING database using the protein data from Usman [8] as input, with type query “protein by name” and “homo sapiens” organism, (3) GO molecular function from UniProt.
3.2. Preprocessing Data

We reduce PPI data obtained from STRING database into several clusters and select the cluster with the highest modularity score. The preprocessing algorithm used to reduce the PPI data was Markov Clustering (MCL). The process in the Markov Clustering is defined on the space of stochastic matrices. Given a graph G, the algorithm employs the process by applying it to the matrix of random walks on G [9].

Markov Clustering is an algorithm that can divide graph based on node connectivity. The first step of this algorithm is to calculate adjacency matrices from a graph. Then, each column in the matrices is turned to stochastic matrices that represent the probability of a random walk or Markov chain that defined in the graph. Matrices column M had a column as much as vertices of the graph G. M(i,j) element represent the probability of transition from vertex vi to vertex vj or the stochastic flow from vertex vi to vertex vj [9].

The main operation of MCL is expansion operation (M) and inflation operation (M,r). Expansion operation is processed by matrices multiplication (M*M). Expansion operation is aimed to increase flow between vertices and open a new potential flow. Inflation operation (M, r) every element in M matrices is powered by inflation parameter r. Inflation operation is aimed to strengthen the strength flow and debilitate the weak flow. Another operation in MCL is normalizing (M) and prune (M). In the normalize (M) the matrices are transformed into stochastic matrices. Normalize (M) operation is processed after inflation (M, r) operation. The prune (M) operation is aimed to accelerate the convergence of M matrices. The Markov Clustering algorithm can be seen in Figure 2.

**Figure 2.** Markov Clustering algorithm.

Next, we calculate modularity value from every cluster from the result of the MCL algorithm. Modularity is one measure of connectedness in graph clustering. The modularity value is defined by:

$$C_{mod} = \frac{E_{in}}{E_{out}}$$  \hspace{1cm} (1)

**Description:**

- $C_{mod}$: modularity value
- $E_{in}$: total edge inside the cluster
- $E_{out}$: total edge outside the cluster
3.3. Fuzzy K Partite Clustering

Fuzzy K Partite is an algorithm used to clustering the k-partite graph G by making the backbone k-partite graph H from k-partite graph G with the smaller size of the graph. Graph G is transformed into adjacency matrices Aij. Adjacency matrices Aij contained a value between the element of i-th partition and j-th partition. The fuzzy k partite algorithm is used twice iteration, the first iteration is to obtain B_{ij} matrices which contain adjacency between fuzzy cluster node from each partition. The fuzzy cluster node is saved in the C_{i} matrices from the second iteration. C_{i} matrices contain a degree of membership value from each node in each partition. The iteration is executed repeatedly until obtaining the convergence of cost function value. The illustration of fuzzy k partite clustering is on Figure 3.

![Figure 3. Illustration of fuzzy k partite clustering [7].](image)

Before using the fuzzy k partite clustering first, we must choose the number of a maximum cluster for each partition. From [2] the formula for choosing a number of a maximum cluster for each partition is defined below:

\[
M_{\text{molecular function}} = \frac{V_{\text{molecular function}}}{10} \quad (2)
\]

\[
M_{\text{gene}} = M_{\text{molecular function}} \left( \frac{V_{\text{gene}}}{V_{\text{molecular function}}} \right)^{\frac{1}{2}} \quad (3)
\]

Description:
- \(M_{\text{molecular function}}\): number of maximum clusters GO molecular function
- \(V_{\text{molecular function}}\): number of nodes in GO molecular function data
- \(M_{\text{gene}}\): number of maximum cluster protein
- \(V_{\text{gene}}\): number of node protein data

4. Results and Discussion

4.1. Data Collection

There were 2375 PPI data obtained from STRING database by 84 proteins data from Usman research that had a direct association with DM type 2. From the 84 proteins, there were 64 proteins that had interaction data and 20 proteins which have no interaction data. Proteins with no interaction data are not used in the next stage.

4.2. Preprocessing Data

This 2375 PPI data then used as input for Markov Clustering. Markov Clustering was run with granularity parameter = 2, expansion value = 2 and number of iterations = 15. There were 40 clusters as a result of MCL. Cluster with highest modularity value was formed by 117 proteins which contain 17 proteins that had a direct association with type II DM (ACE, AKT2, GCG, GCK, IAPP, IL6, INS,
INSR, IRS1, IRS2, NEUROD1, PDX1, PPARG, RETN, SLC2A2, UCP3, and VEGFA). The graph representing this cluster was shown in Figure 4 and 5.

![Figure 4. Result of clustering using the MCL algorithm.](image)

Figure 4. Result of clustering using the MCL algorithm.

![Figure 5. Proteins interaction in cluster 1 with highest modularity score.](image)

Figure 5. Proteins interaction in cluster 1 with highest modularity score.

Clustering visualization in this research is conducted using network package from Python programming language. At first, the accuracy of MCL implemented with Python was very low compared to that of using Cytoscape. It probably was caused by assigning an improper weight of the edge. After improving the code, the accuracy was increased up to 84.6%.

From 117 proteins data obtained by MCL, 328 GO molecular functions were gained using the UniProt database. Then, both this protein and GO molecular function data were transformed into a bipartite graph for Fuzzy K Partite Clustering input.

### 4.3. Fuzzy K Partite Clustering

Using the previous formula, the number of a maximum cluster for GO molecular function is 33 and the number of a maximum cluster for proteins is 20. From the Fuzzy K Partite algorithm, we obtained the lowest average cost function 594.175 which is in 29 number of cluster.

Based on [7] the threshold value for the degree of membership of each cluster is 0.2. The protein cluster with the least number of members is cluster 12, 14, and 20 which only had 1 protein. Otherwise, The protein cluster with the most number of members is cluster 5 which had 23 proteins. The number of members for each protein cluster showed in Table 1.

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| Number of proteins | 2 | 9 | 14 | 10 | 23 | 10 | 7 | 2 | 3 | 5 | 12 | 1 | 6 | 1 | 4 | 5 | 5 | 7 | 12 | 1 |
In addition, the numbers of clusters obtained for GO molecular function are 29 clusters. With the same threshold we applied to the protein cluster, we obtain 6 clusters that have no member. Moreover, the GO molecular function cluster with the least number of members is cluster 15 with 2 GO molecular functions. On the other hand, the cluster with the highest number of degrees of membership is cluster 17 consisting of 45 GO molecular functions. The number of each GO molecular function cluster showed in Table 2.

Table 2. Number of members each GO molecular function.

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| Number of GO molecular function | 12 | 23 | 13 | 20 | 17 | 30 | 13 | 0 | 38 | 16 | 10 | 6 | 0 | 0 | 2 |

| Cluster | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|---------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Number of GO molecular function | 31 | 45 | 13 | 12 | 0 | 33 | 0 | 10 | 31 | 14 | 9 | 0 | 8 | 28 |

5. Conclusion
Markov clustering could be implemented in the data preprocessing step to reduce the size of the PPI network from 2375 PPI data into 117 proteins in which 17 protein identified proteins that had a direct association with type II diabetes. The fuzzy k-partite clustering algorithm can be applied to PPI data and GO molecular function. Clusters of protein yielded in this study have a relationship with type 2 DM based on members of the GO cluster of molecular functions that are connected to these protein clusters. The clustering accuracy of this study is comparable to that of using Cytoscape, one of the popular software for network analysis.

References
[1] World Health Organization. 1999. Dengue and dengue hemorrhagic fever [internet]. [downloaded at 2016, September 10th]; 73(2):18. Link: http://www.who.int/mediacentre/factsheets/fs117/en/index.html.
[2] Bordonaro M. 2009. Role of wnt signaling in development of type 2 diabetes. Cell & Bioscience [internet]. [downloaded at 2017 February 1st]; 80:563-581. Link: https://ncbi.nlm.nih.gov.
[3] Liu W, Wu A, Pellegrini M, Wang X. 2015. Integrative Analysis of Human Protein, Function, and Disease Networks. Scientific Report. 5(14344): 1.
[4] M Li, H Zhang, J Wang, Y Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst. Biol. 2012; 6(1): 15.
[5] J Ran, H Li, J Fu, L Liu, X Ying, X Li, H Shen, Y Chen, X Jiang, Y Li, H Li. 2013. Construction and analysis of the protein-protein interaction network related to essential hypertension. BMC Syst. Biol.; 7: 32.
[6] Singh M, Singh G. 2011. Cluster analysis technique based on the bipartite graph for human protein class prediction. International Journal Computer Applications. 20(33): 22-27.
[7] Hartsperger ML, Florian F, Stumpflen V, Theis FJ. 2010. Structuring heterogeneous biological information using fuzzy clustering of k-partite graphs. BMC Bioinformatics. 11(522): 2-13.
[8] Usman MS, Kusuma WA, Afendi FM, Heryanto R. 2015. Identification of Significant Proteins Associated with Diabetes Mellitus using Network Analysis of Protein-protein Interactions. Master thesis, Bogor Agriculture University.
[9] Dongen, Stijn Van. 2008. Graph Clustering via a Discrete Uncoupling Process. Society for Industrial and Applied Mathematics. 30(1): 121-141