CLUSTERING ON GENE ANNOTATIONS

Aishwarya AV
School of C&IT
REVA UNIVERSITY
Bangalore,India
aishuav01@gmail.com

Akshita Srikanth
School of C&IT
REVA UNIVERSITY
Bangalore,India
akshita291997@gmail.com

Aishwarya Gajanana Naik
School of C&IT
REVA UNIVERSITY
Bangalore,India
gnaikaishwarya97@gmail.com

Abstract*—This paper proposes the formation of clusters of identical or similar protein complexes . Using the gene annotation dataset as training dataset, which includes gene names and their functionalities or behaviour , we applied and analyzed various similarity techniques to identify the best technique among them to find similar genes (protein complexes) . Based on the results of similar genes , we form the PPI network . To form clusters of these networks , we applied and analyzed various centrality measures to identify the hub node or the most influential node . The clusters so formed help us in easy identification of the category of protein complex they belong to.

Keywords : Gene Annotations , Protein-Protein Interaction (PPI) Network , Semantic Similarity , Clustering

INTRODUCTION

Sections of DeoxyRiboNucleic acid (*DNA*) containing information to code specific proteins are called genes. Gene controls the functions performed by protein. The process of identifying the coding regions in genes and their functionalities are called gene annotation. It helps in finding and attaching structural elements and its related function to each gene location. It has all the biological information to build any given living organism. With the help of gene annotation, it is possible to identify and predict the functions of protein complexes and thus perform further comparative analysis [1].

Proteins play a major role in cellular functions. For understanding cellular processes, identification of essential proteins in protein-protein interactions (*PPI*) networks has great significance [2]. Identifying protein complexes not only help in understanding the building blocks of an organism but also the characteristics of proteins help in prediction of related disease or of target cells that might be associated with it [3].

Essential proteins play an important role in maintaining cellular life. The common methods for essential protein detection include mutagenesis, and gene knockout [4]. The native methods of identification of essential proteins were time consuming and tedious. Thus there has been a significant effort to identify protein complexes in protein-protein interaction networks.

Gene ontology (*GO*) provides the genes and the function associated with it is most widely used as a resource for gene annotations. *GO* is a bioinformatics resource. It can be called as a biological vocabulary that displays the functionalities of the main categories including cellular component, molecular function and biological process. Semantic Similarity between two sentences is defined based on the structure and syntax of the sentences and if there exists a similarity between them , it is considered that both the sentences convey a similar meaning .To give a numeric value (measuring value) to the *GO* term, we perform semantic similarity of GO terms.

Using the ontological data resource (nominal data), similarity techniques can be applied to find the similar genes based on their functionalities [5]. The common

text similarity measures include cosine similarity, pairwise similarity, Jaccard's similarity , Levenshtein, based on distance and ratio measures,Latent Semantic Indexing (LSI), Word Mover's Distance,Latent Dirichlet Allocation . Most of these measures find the common terms among the two texts and also find the frequency of those terms to quantify their significance in the sentence , in some cases this is also called weighting (adding weights to terms) .The result provided by these measures are usually a numeric value between 0 to 1 . These measures help us in comparing two texts and understanding the extent to which the sentences are similar.

The protein-protein interaction network is a graphical representation of proteins connected to each other through edges , it can be an undirected or directed graph. The edges may contain weights representing the closeness between the protein nodes . Mapping similar genes/proteins to each other in a graph , the PPI networks can be formed . These networks usually depict a biological function .

Centrality measures are used to find the most influential node in a network . Centrality measures include Degree centrality centrality ,Closeness centrality , Betweenness centrality,Eigenvector centrality, Harmonic centrality , Second Order centrality , Group centrality and many more . These measures provide a numerical value for every node in a network which represents the contribution of that node in the network. The node with highest centrality value is identified as the central node meaning that it is connected to most of the nodes in the network . The PPI network with the hub node is considered as a cluster and specifying the biological function of hub node.

Clustering helps in analysing the data by grouping them into a specific group/category the data belongs to . There are various clustering algorithms like centroid - based clustering , density - based clustering , distribution based clustering and hierarchical clustering .The degree of extent to which the nodes in a network cluster together can be found by calculating clustering coefficient . Average coefficient clustering provides the average clustering over the entire network and how complete the cluster is . Global clustering coefficient , square clustering coefficient, triangle clustering coefficient are other clustering coefficients .

LITERATURE REVIEW

The idea to form a dense network of molecular interaction in a cell requires proteins, nucleic acids and small molecules. Interaction between the molecules are nodes and edges. Network architecture of molecular structure, tells about the protein function and organisation. Presence of highly connected nodes form clusters of proteins with interactions of protein networks.[3].Later in 2007, [4] proposed a novel method where an algorithm is implemented for determining the semantic similarity of GO(gene ontology) terms. Applying this algorithm, to measure the similarity of genes functionally. Results of clustering the genes based on the similarity values are gained by using similarity of genes online based tools. Further we reviewed that in [5] , they evaluated the impact with the progress of technologies and researches in genomics, they disclosed the dynamic structure and function of genes. The process of recognizing the genomic elements and their function known genome annotation. This can be stored in the form of text format. . Hence we are able to explore the query or visualize the genomics data.

Using a method to compare several commonly used gene clustering algorithms using a figure of merit based on the mutual information between cluster and known gene attributes was designed as stated in [7]. By using various datasets, the k-means algorithm was implemented for clustering of genes.
Clustering of genes on similarity of their expression patterns in a set. genes having similar expression patterns are likely to have similar biological function. Clustering's main objective is to get genes of similar function together.In 2003 [6] , a method to investigate the semantic similarity measures are correlated with sequence similarity. A metric for semantic similarity is based upon ontological annotation of data resources. Many bioinformatics resources have data and annotations. Annotations are used by the bioinformatician to increase the knowledge about the data in the sequence database.

The base paper [1] provides an approach to protein complex identification. There is an algorithm present for inferring the protein complexes and more effort has been put into the making, as the structure is complex. They design the complex subgraph by a probabilistic Bayesian Network(BN). There are training sets of known complexes which are used to learn the parameters of the BN model. It extracts the features which are used to distinguish complex versus the non-complexes. This experiment shows that EGCPI using the evolutionary graph

clustering can provide better results for identifying the protein complexes.

In 2011 , a paper [8] was published which proposed that the node centrality measures which plays an important role in a large number of graph applications and biological network analysis. Centrality has been proposed, ranging from very simple (e.g., node degree) to more elaborate to scalable methods. Centrality is widely-used for measuring the relative importance of nodes within a graph of small nodes. The types of centrality measures are close centrality, centrality degree, betweenness centrality, and eigenvector centrality. Also in 2018 , paper [2] for predicting the essential proteins by combining network topology for functioning of cells. Based on PPI(protein protein interactions), network levels are important. For identifying essential proteins on PPI networks, centrality methods and GO(gene ontology) similarity measures are utilised.

Similarity Analysis

Aggregation of similar genes based on the semantic similarity analysis of GO terms .To understand the best similarity technique to be applied on the GO data resource , few known similarity techniques were applied on the dataset .

General Steps followed across all similarity measures:

1. The similarity measures were applied on the same data resource.
2. If the similarity value >= .60 (60% threshold) then
   similarity = 1
   else
   similarity = 0
3. The similarity values(1's and 0's) between the genes were mapped into two - dimensional matrices .

A. *Cosine similarity*

Cosine similarity measures the cosine of the angle between two non-zero vectors and it can be used for measuring similar documents regardless of their size and is one of the common approaches used to find the similarity between the two documents. Similarity between the two documents is identified based on counting number of maximum words present.

$$Cosine\ Similarity\ (A, B) \quad =$$

$$A.\ B \ = \ ||A||\ ||B||\ cos\ \boldsymbol{\theta}$$

In python we start with finding word count in the sentence using CountVectorizer or TfidfVectorizer (also provides frequency count of each word) provided by scikit learn and then applying cosine similarity function on it . This function can take inputs as a sparse matrix or as a pandas dataframe . The final output is a matrix with resultant similarity values.

B. *Pairwise document similarity method*

Pairwise document similarity method is a text document similarity technique based on the weights of terms in each document and the common terms (information)shared by two documents [11]. The weight of a term is assigned by a weighting scheme and indicates the significance of the term in that document. Weighting schemes like Term frequency (tf) , inverse document frequency (idf) , or multiplication of tf and idf (tf-idf) can be used .We have used the tf-idf form to measure pairwise similarity.Using the TfidfVectorizer module provided by scikit learn .

```
data = ['apples and oranges', 'oranges and bananas']
vector = TfidfVectorizer()
tfidf = vector.fit_transform(data)
Pairwise_similarity = tfidf * tfidf.T
```

C. *Jaccard similarity*

Jaccard similarity or union over intersection is measured as the size of intersection divided by the size of union of two words or sentences. In this,we can find the similarity between two documents.Two values used are "0" and '1' which represent a degree similarity. Value '0' means the documents dissimilar and '1' means the documents are identical.

Jaccard similarity =

$$J(A, B) \ = \ |A \cap B| \div |A \cup B|$$
$$= \ |A \cap B| \ \div (\ |A| \ + \ |B| \ - \ |A \cap B|\ )$$

Jaccard similarity for each document holds only a unique set of words. First by performing lemmatization to reduce words to the same root word, in order to calculate similarity using Jaccard similarity.

D. Levenshtein ratio and distance measure

Levenshtein distance with an unequal length, performs on strings.This is also a string metric for measuring the difference between two sequences. Levenshtein distance between two strings function can be '0' and 1'.

Levenshtein distance uses dynamic programming approach such as : i) String matching
ii) Spelling checking

E. Sequence matcher using difflib

Sequence matchers can be used for comparing pairs of input sequences or strings. It focuses on comparing the longest contiguous matching subsequence between the two input sequences and finding the longest matching subsequence that contains no junk values. This is a class available in the python module named 'difflib'.

ALGORITHM sequence_matcher : **diff**(*a*, *b*)

Parameters:

a : *sequence of comparable*
Initial sequence
b : *sequence of comparable*
Changed sequence

Returns: An *iterable* of operations.

Conclusion :

Based on the output of similarity values (1's and 0's) , the technique with maximum number of 1's was taken to be the best method for similarity analysis because more the number of 1's , there would be more similar genes .Here , with respect to the data resource used , Levenshtein Distance measure produced the most related values .

	BUB1B	CENPE	INCENP	CENPA	CCNA2	MAD2L1
BUB1B	1	0	0	1	0	0
CENPE	0	1	0	0	0	1
INCENP	0	0	1	0	0	0
CENPA	1	0	0	1	0	0
CCNA2	0	0	0	0	1	0
MAD2L1	0	1	0	0	0	1
NEK2	0	0	0	0	0	0

*fig:*1.0 Levenshtein Similarity Result

The figure fig 1.0 represents part of result obtained , where the similar genes (row and column headers) have a value 1 and dissimilar have a value 0 at their intersection.

PROTEIN-PROTEIN INTERACTION NETWORK

The similar genes can be linked to form a network of nodes , where the nodes represent the gene and the edges are formed between the nodes only if the similarity weight is 1 .

Based on the results obtained from Levenshtein distance measure , the similar genes were grouped together and formed a network . Used a popular software package in python called Networkx. We formed an undirected graph of genes which is called **Protein-Protein Interaction Network .**

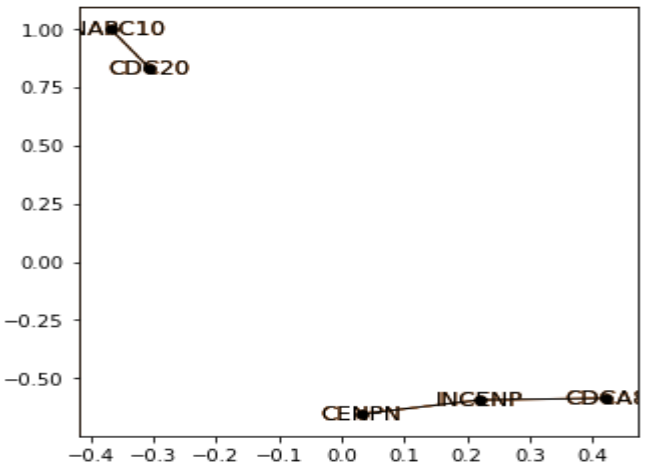

fig:2.0 The partial PPI network .

CENTRALITY ANALYSIS

Centrality identifies how important a node or edge is for the connection or the flow of the network within a graph. With respect to the PPI network formed , we applied centrality measures to identify the hub or influential node .

Identification of hub nodes is important as this node will be connected to most number of nodes in the network and is having a strong influence over all the other nodes . The functionality of the gene , identified as hub node , can be considered as the function of that whole network .

Four approaches were used for analysing the network of gene interaction.

A. Closeness Centrality

Closeness centrality of a node is a measure of centrality in a network, which is calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph.This centrality indicates how close a node is to all other nodes in the network.

Closeness :

$$C(x) = \frac{1}{\Sigma y\, d(y,x)}$$

where d(y,x) = distance between vertices ‘x’ and ‘y’.

Results: [‘Gene/Protein’ : Centrality value]
‘INCENP’: 0.5, ‘CDCA8’: 0.25, ‘CENPN’:0.25,
‘ANAPC10’: 0.25

This is a partial output of closeness centrality measure with respect to data resource used.

B. Degree Centrality

One of the simplest centrality measures to compute is Degree centrality. The degree of vertex of a graph is the number of edges incident to the vertex, counted twice with loops.

Degree Centrality :
The degree centrality of a vertex ‘υ’, for a graph

G: =(V, E)

with |V| vertices and |E| edges, is defined as

$$C_D(\upsilon) = deg\,(\upsilon)$$

Results:[‘Gene/Protein’ : Centrality value]
‘INCENP’: 0.5, ‘
CDCA8’:0.3333333333333333,
‘CENPN’:0.3333333333333333,
‘CDC20’: 0.25,
‘ANAPC10’: 0.25

INCENP is recognised as the hub node with highest value.

This is a partial output of close centrality measure with respect to data resources used.

C. *Betweenness Centrality*

Betweenness centrality is a measure of centrality in a graph based on the shortest paths. Every pair of vertices with at least one shortest path the vertices , the number of edges the path passes through or the sum of the weights edges.

Betweenness:

$$g(\upsilon) = \Sigma\, s \neq \upsilon \neq t \; \frac{\sigma st\,(\upsilon)}{\sigma st}$$

where $\sigma st$is the total number of shortest paths from node s to node t and is the number of paths that pass through υ..

Results:[‘Gene/Protein’ : Centrality value]
'INCENP': 0.16666666666666666,
'CDCA8': 0.0,
'CENPN': 0.0,
'CDC20': 0.0,
'ANAPC10': 0.0

INCENP is recognised as the hub node with highest value.

This is a partial output of betweenness centrality measure with respect to data resources used.

D. *Eigenvector centrality*

Eigenvector centrality or eigen centrality is a measure of the influence of a node on a network. If a node is pointed

by many nodes, then that node will have high eigen-vector centrality.

Eigenvector :

$$x_\upsilon = \frac{1}{\lambda} \sum_{t \epsilon M(\upsilon)} x \quad = \sum_{t \varepsilon G} a_\upsilon, \quad x$$

where M(υ)is a set of neighbours of υand λis a constant.

Results: [‘Gene/Protein’ : Centrality value]
‘INCENP’:0.7071067811066628,
’CDCA8’:0.49999999994351296,
’CENPN’:0.49999999994351296,
‘CDC20’:1.0628924235733579e-05,
’ANAPC10’: 1.0628924235733579e-05

INCENP is recognised as the hub node with highest value.

This is a partial output of eigenvector centrality measure with respect to data resources used.

Conclusion :

We used four different centrality measures to verify if the hub node so formed was the same in all the measures . We noticed that the hub node (here INCENP for partial result) was returned as the hub node in all the methods.

CLUSTERING

The PPI network for the data resource used , had several smaller networks formed . Each network was identified with a hub node . This network is now a cluster and the hub node represents the central node of the cluster . The cluster as a whole represents all the nodes(genes/proteins) with the similar biological function.

The network with three nodes or more than that was considered as the cluster . Below fig 3.0 , depicts a partial output of the data resource , with two networks but with one cluster and its hub node .The hub node is represented with a unique color .

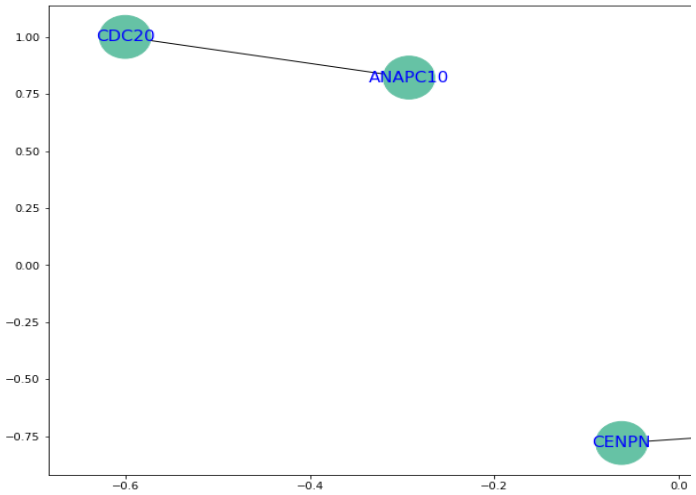

fig 3.0 Partial network formed with 2 networks and in which one is cluster with hub node

CLUSTERING COEFFICIENTS

Clustering coefficient is a measure of the degree where nodes in a graph tend to form a cluster together. The clustering coefficient of a graph closely depends on the transitivity of a graph.

Clustering coefficient metric differs from measures of centrality. When clustering coefficient is high,these nodes in network connections are dense.

Clustering coefficient is used for both directed graphs and mostly undirected graphs. The graph obtained here was undirected and using the Networkx package in python , we performed the following clustering coefficients calculation.

A. *Average clustering coefficient*

The level of clustering in a network is measured by the average of the local clustering coefficients of all the vertices 'n'.

Average coeff. :

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} Ci$$

B. *Squaring clustering coefficient*

Clustering coefficients of squaring are the nodes.For every node that returns the fraction of possible squares exist at the node.

APPLICATIONS

Gene annotation represents the functional information about the gene. The correlation of semantic similarity to sequence similarity helps to predict protein function. Genes with similar expression patterns can be made into clusters and further analysis can be done on this. Gene prediction is a key function of annotation, and it helps to investigate the protein binding sites within a genome. The main limitation of any ontology approach is that incomplete GO annotation cannot be used to cover any statistical information.

RESULTS

The dataset used is from Gene Ontology database (www.geneontology.org) [12]. It consists of Electronic Annotations that are Swiss –Prot reviewed and manually curated Annotation or un reviewed Annotations. The above mentioned analysis were performed on the dataset and the below fig 4.0 was the final clusters formed .

Part of dataset used is given below:

GENE	FUNCTION
B5KUL2	oxidation-reduction process, catalytic activity,FMN hydroxy acid dehydrogenase domain-containing protein
BUB1B	Mitotic checkpoint serine/threonine-protein kinase BUB1 beta
CENPE	Centromere-associated protein E; Microtubule plus-end-directed kinetochore motor which plays an important role in chromosome congression
INCENP	Inner centromere protein; Component of the chromosomal passenger complex, a complex that acts as a key regulator of mitosis.

Table 4..0 Part of Dataset

The final clusters obtained are shown in fig 4.0

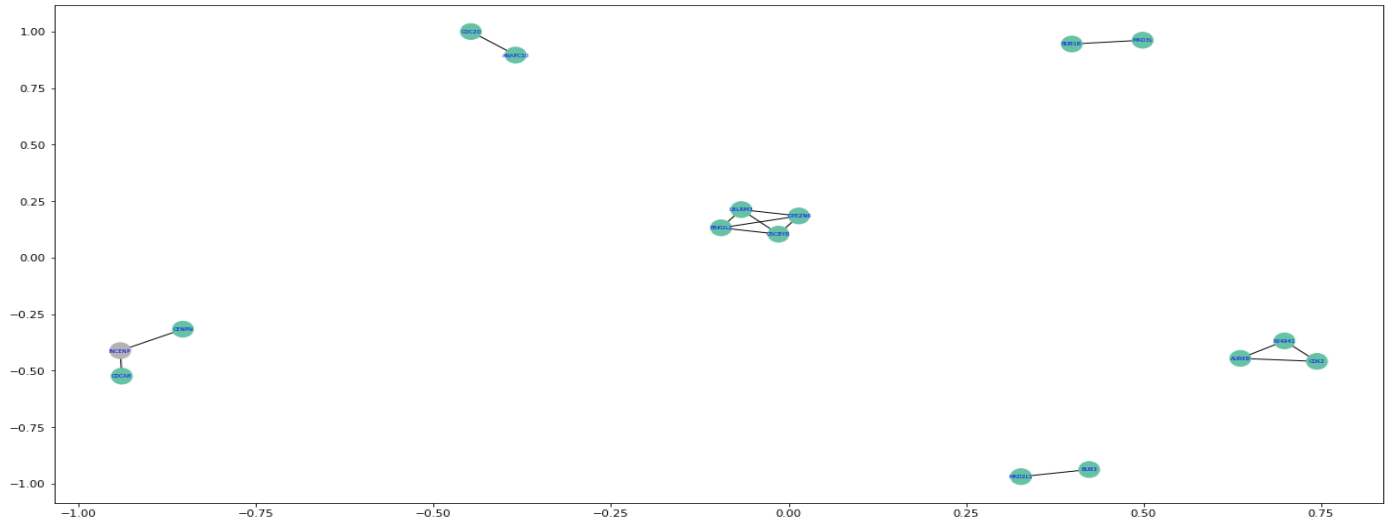

fig 4.0 Clusters formed (network with 3 and more nodes)

The Table 4.1 represents the enrichment analysis for the clusters . It represents the annotation cluster , representative annotation term that is the most common function among the cluster nodes and the enrichment score stating the number of nodes in the cluster. Any new gene/protein provided by Gene Ontology can be applied with the following analysis and then add it to the respective cluster it belongs to . This analysis helps us to identify the other common functions the gene/protein might have and also the other features it holds based on the cluster properties .

Table 4.1: Enrichment Analysis for clusters.

Annotation cluster	Representative Annotation Term	Enrichment Score
1 [P24941, CDK2,AURKB]	Cyclin-dependent kinase,Serine/threonine-protein kinase involved in the control of the cell cycle	3
2 [CENPN,INCENP,CDCA8]	Component of the chromosomal passenger complex, a complex that acts as a key regulator of mitosis	3
3 [D7EZN6,U5LRM3,G5CBY8,B5KUL2]	Catalytic activity,FMN hydroxy acid dehydrogenase domain-containing protein	4

CONCLUSION

For the data resource used ,we have applied and analysed semantic text similarity methods to find the best optimum similarity methods which helps to determine functionally similar genes. Furthermore formed PPI networks of similar

genes and identified the hub nodes of the protein complexes by using various centrality measures. Therefore, this analysis can be used for bioinformatics dataset to categorise and group the genes/proteins into their respective groups . Based on characteristics of the identified cluster group , further behaviour of the gene can be predicted.

Reference

[1] Tiantian He, and Keith C.C. Chan ,"Evolutionary Graph Clustering for Protein Complex Identification", IEEE Trans. on Comp. Bio. and Bioinfo., vol.15 , no.3, pp.892-904,May-Jun. 2018.

[2] Wei Zhang , Jia Xu, Yuanyuan Li, and Xiufen Zou, "Detecting Essential Proteins Based on Network Topology, Gene Expression Data, and Gene Ontology Information", IEEE Trans. On Comp. Bio. and Bioinfo., vol. 15, no.1, pp.109-116, Jan/Feb. 2018.

[3] Victor Spirin and Leonid A. Mirny, "Protein Complexes and Functional Modules in Molecular Networks", Science, vol.100, no. (21), pp. 12123-12128, Oct. 2003.

[4] James Z. Wang, Zhidian Du,Rapeeporn Payattakool, Philip S. Yu and Chin-Fu Chen, A new method to "Measure the Semantic Similarity of GO terms",Science ,Bioinformatics, vol.23 , no.10 , pp.1274-801 , Jun.2007.

[5] Ayllón-Benítez A, Mougin F, Allali J, Thiébaut R, Thébault P,"A New Method for Evaluating the impacts of Semantic Similarity Measures on the Annotation of Gene Sets", Science, PLOS ONE, vol. 13, no.11, Nov.2018.

[6] P.W. Lord , R.D. Stevens, A.Brass and C.A. Goble,"Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship Between Sequence and Annotation", Science, Bioinformatics, vol. 19, no. 10, pp. 1275–1283, 2003 .

[7] Francis D. Gibbons and Frederick P. Roth,"Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation", Science, Genome Research, vol.12, no.10, pp. 1574–1581, Oct. 2002.

[8] Kang, Spiros Papadimitriou, Jimeng Sun and Hanghang Tong,"Centralities in Large Networks: (Algorithms and Observations)", Computer Science , Data mining , pp. 119-130 , Dec. 2011.

[9] Leila Ranandeh Kalankesh, Robert Stevens and Andy Brass,"The Language of Gene Ontology:a Zipf's Law", Science, BMC Bioinformatics , vol.13,no.127 , 2012.

[10] Imad Abugessaisa, Takeya Kasukawa, and Hideya Kawaji, "Genome Annotation",Science+Business , Springer, vol. 1525, 2017.

[11] Oghbaie, M., Mohammadi Zanjireh, M., "Pairwise Document Similarity Measure based on present term set. *J Big Data* 5, 52 (2018).

[12] Harris, M. A., Clark, J., Ireland, A., et al. Gene Ontology Consortium 2004. *The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res*, *32*, D258-D261.DOI: 10.1093/nar/gkh036