Original Research Sciences **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

GENE ONTOLOGY BASED FUNCTIONAL ANALYSIS AND GRAPH THEORY FOR PARTITIONING GENE INTERACTION NETWORKS

***SREEJA ASHOK AND DR U.KRISHNAKUMAR**

**Assistant Professor ,Department of Computer Science & I.T., School of Arts & Sciences*
Director, School of Arts & Sciences, Amrita University, Kochi

ABSTRACT

Exploring gene-disease associations improves the understanding of the underlying cause of the disease, which leads to further improvements in the diagnosis and treatment. Since genes that belong to the same topological, functional or disease module has an increased tendency of being involved in the same disease or phenotype, cluster analysis is the efficient approach to identify functionally similar genes. The aim of this work is to identify biologically relevant gene clusters using graph theory, which is an essential and influential scientific tool for modeling and exploring interconnected groups. The current work exhibits a computationally efficient algorithm for improving the performance of community detection in graphs using edge pruning techniques. The algorithm does not demand a priori judgment on the size of communities; it helps in automatic detection of gene communities with better performance. The optimized and streamlined approach is applied on cancer dataset and is compared and validated with standard clustering solutions using different validation measures.

KEYWORDS : **Community detection, Gene Ontology, Similarity Matrix, Node Connection Score, Silhouette Index, Modularity**

SREEJA ASHOK

Assistant Professor, Department of Computer Science & I.T., School of Arts & Sciences.

Received on: 16-01-2017
Revised and Accepted on 15-02-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.2.b183-192

INTRODUCTION

With the development of high-throughput devices and highly efficient technologies, the volume of biological data is increasing with a great pace. Traditional methods for gene mapping such as linkage analysis and association studies are used for identifying disease-associated genes[1].But these methods only associate a disease with chromosomal regions that usually contain tens or even hundreds of genes[2]. Hence it is essential to develop efficient computational methods to identify the groups of strongly associated genes from a long list of candidate genes. Several methods have been proposed to address this problem by identifying, ranking and grouping candidate set of genes. Guilt-by-direct-association[3] principle is used to facilitate the prioritization of candidate genes in many problems where disease susceptibility genes are ranked based on their relevance to the genes that are already known to be associated with the disease genes. Even though different clustering solutions are widely used for pattern extraction and for logically grouping homogeneous genes, clustering techniques with the ability to visualize is more commonly used by experts. This increases the understanding of the connections between datasets within groups and makes it simpler to comprehend the results of the clusters. Graph-based models represent complex interactions where data objects/observations are represented as nodes and relationships between observations as edges. The most significant and expressive feature of graphs in demonstrating real-time system is the community structure. The goal of the current approach is to discover communities or clusters by partitioning the vertices in a large semantic similarity gene network into different clusters based on neighborhood similarity. The algorithm uses gene ontology information to identify the functionally accurate clusters. The biological significance of these newly identified clusters is evaluated using gene-gene interactions.

Related Work

Different clustering solutions were developed so far to address the problem of grouping functionally similar genes[43-45]. Graph based clustering is the most realistic approach to demonstrate the biological reality. Numerous studies focus on topological characteristics of gene networks for extracting homogeneous groups. Jonsson *et al.*[4,5] performed cluster analysis using PPI networks to conclude that disease genes belong to communities with a higher degree of connectivity and act as a key central node. Even though Cai[6] infer similar findings in their studies, they could observe an inverse relationship between disease genes and clustering coefficient. Rahmani[7] used network properties to predict disease genes in PPI networks. Two parameters were considered for analysis in this approach; the functional context of the proteins and the rank of target proteins with respect to the selected proteins set using ANOVA analysis. Zhang *et al.*[8]studied online community detection for large, complex networks by proposing a node split based on the degree of the nodes for creating new clusters. The method has been applied to a set of real world network datasets and was compared with Agglomerative[9] and Eigen vector approach[10] for community detection. Fast greedy algorithm [11] uses greedy method while walktrap algorithm[12] uses a random walk process to find the distance between two nodes. Both approaches used hierarchical agglomerative clustering for deriving meaningful communities. To maximize the density of the clusters obtained, Monte Carlo algorithm was proposed by Spirin *et al.*[13]where the algorithm demands the cluster size as input parameter for detecting the protein complexes. Highly connected sub graph algorithm is used in several studies for deriving meaningful sub graphs from a connected network by Hartuvetal.[14]. A sub graph with n nodes are said to be a highly connected sub graph if more than n/2 edges must be removed in order to disconnect it. Przulj etal.[15] used this approach in determining the protein complexes in PPI network. The algorithm partitions the graph by finding a minimum cut in the graph. A cut in a graph is a partition of nodes into two non-overlapping sets. Recursive process is initiated until highly connected components are obtained. Krogan *et al.*[16]and Enright *et al.*[17]used Markov clustering (MCL) algorithm for complex detection in PPI network. An adjacency matrix is computed for a graph G= (V, E), A|V|×|V|= $\{a_{ij}\}$ where a_{ij}=1 if and only if v_i and v_i are neighbors is a matrix. High and low flow regions are identified and the process converges towards high flow regions of sub graphs known as protein complexes separated by no flow regions. Clique percolation (Adamcsek *et al)*[18], spectral methods (Bu *et al.)*[19], graph flow simulation (Pereira-Leal *et al.)*[20], edge-betweenness clustering (Dunn *et al.)*[21]etc are the additional graph clustering techniques applied to explore sub graphs from a network. Most of the research effort on building knowledge based gene clustering algorithm is focused on integrating gene ontology's for extracting the semantic similarity between genes. Graph based gene ontology structures were used by Cheng *et al.*[22] for inferring the degree of relationship between genes. The similarity metric is then applied to a community detection algorithm for finding the group of genes that are related by their biological significance and then integrated with an expression-based metric to perform co-cluster analysis for performance improvement. Dotan-Cohen *et al.*[23] integrated semantic similarity measures for deriving clusters using hierarchical clustering approach. The knowledge of bimolecular functions engraved in the GO was used for cutting the dendogram into meaningful clusters. Graph clustering algorithms make extensive requests on computational assets and demands cluster size as input parameter which increases the computational complexity and performance of the system. The present research proposes an Edge Reduction and Community Detection Algorithm (ERCD) that improves the performance of the graph clustering by finding out the highly connected edges in the graph. The objective is to partition the nodes of a graph based on a threshold value and to create the optimum clusters from the dataset automatically into k independent groups that maximizes the intragroup similarities and minimizes the intergroup similarities. The method uses beta-PERT distribution for modeling the distributions of variables and finding a threshold value which represents the upper bound to segregate the strong and weak connections between the nodes. The similarity threshold

value for each node analytically guarantees the connectivity of each node with adjacent nodes and creates more accurate groupings/clusters than the existing clustering methods.

MATERIALS AND METHODS

Gene Ontology (GO) is a major initiative in bioinformatics that includes complete and ample knowledge about genes and gene-products in a structured and controlled format. They are symbolized as directed acyclic graphs (DAGs) in which the nodes represents the terms and the edges represents the semantic relations. GO is subdivided into three non-overlapping ontology's, Molecular Function (MF), Biological Process (BP) and Cellular Component (CC).Genes and gene products are connected to GO terms using annotations which are linked to a source file which can be a computational, database evidence or a literature reference (Gene Ontology Consortium. 2004[24]; Smith *et al.)*[25]. GO is extensively used in data analysis and broadly incorporated into various data analysis platforms and applications. GO annotations help in explaining the similarity of a group of genes using different measures. Recent studies on gene ontology using gene annotations enlighten the functioning of complex biological systems. Functional similarity between proteins based on GO helps in extrapolation of disease genes[26].Vafaee *et al.*[27] introduced semantic similarity measures for predicting gene functional associations. Arnaud *et al.*[28] conducted study on ontology to look up genes with similar functionality or location within the cell. Rhee *et al.*[29] and Lovering *et al.*[30] did a study to infer the location or function of genes that are over- or under-expressed using GO annotations. De Bodt *et al.*[31]evaluated the protein–protein interactions (PPI) utilizing the cellular component ontology and inferred that proteins are likely to interact if they are in the same location. Since Gene Ontology represents highly enlightening and comprehensive annotations of genes products and due to its ability to quickly compute pairwise similarity between gene annotations, we used GO annotations as a measure to compute the similarity between genes in the current work.

Edge reduction and community detection (ERCD)-Process Flow

Here we propose a graph based clustering approach that uses the weight of the edges to explore sub graphs of functionally similar genes from Semantic Similarity Network. The process of building a gene semantic similarity network and extracting communities are detailed in Figure 1. ERCD algorithm tries to partition a weighted undirected graph G into homogeneous sub groups or communities such that nodes in the same group are thickly associated, having numerous edge "within" the community than edges linking nodes "between" different communities by

- Cutting or removing edges of each node which are weakly connected based on statistical approach.
- Detecting dense sub graphs by grouping nodes with maximum neighborhood connectivity

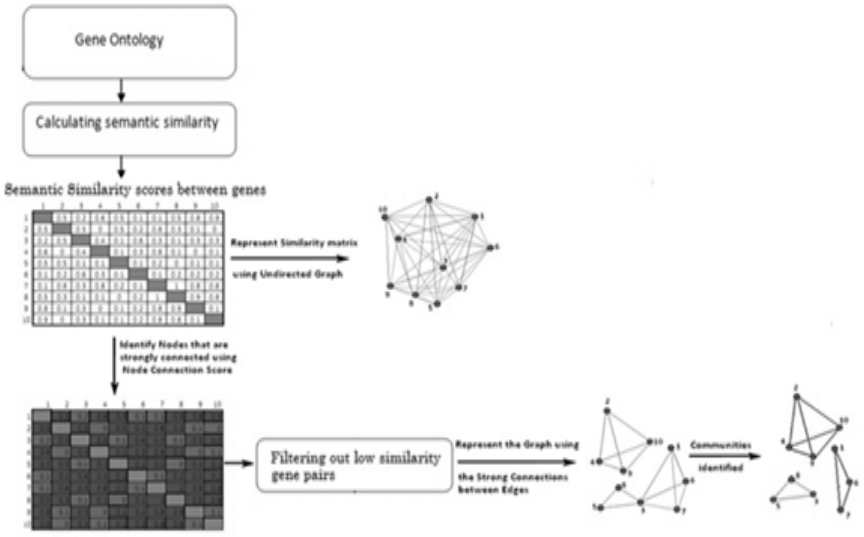

Figure 1
ERCD Process Flow

Components of the ERCD process includes a semantic similarity matrix constructed based on the knowledge source used, the matrix that is transformed into a graph with genes as nodes and edge weights as the association between nodes, optimum semantic graph constructed based on strong associations only and the community extraction process from the optimum graph. The process is detailed in the following steps

Step 1 - Select a Measure of Similarity

There are various measures to express similarity between pairs of objects. If the attribute values are discrete, the similarities between two nodes are determined by examining each of the attributes and counting the number of attribute values they have in common. For continuous attributes, different well known measurements are utilized to compute the separation between two data objects like Euclidean, Manhattan, Murkowski, Canberra, correlation etc. (Pandit*et al.*)[32].

Semantic Similarity between GO terms

Usually the analysis happens at the gene level and each gene is associated with multiple GO terms. Different approaches are used to calculate the similarity based on GO terms associated with it. Information Content (IC) based approach is one of the typical methods used to measure semantic similarity of genes. IC-based similarity measures mainly depend on the factors like frequency of the GO terms and their closest common ancestor term. Equation (1) defines the method for calculating the frequency of a term t;

$$p(t) = \frac{n_{t'}}{N} \mid t' \in \{t, children\ of\ t\} \qquad (1)$$

where *n* is the number of term *t* and *N* is the total number of terms in GO corpus. IC of a GO term is calculated by negative logarithmic probability of the term, defined in equation (2):

$$IC(t) = -\log(p(t)) \qquad (2)$$

Since multiple parents for each concept are allowed in GO allows multiple parents, it is possible to have two terms with common parents with multiple paths. IC based methods finds the similarity of GO terms based on their common ancestor term's IC. Such a common ancestor term is also known as Most Informative Information Ancestor (MICA). There are four IC based similarity measures that are commonly used, proposed by Resnik[33], Jiang et al.[34], Lin and D.[35] and Schlicker et al.[36].

Resnik Method

$$sim_{Resnic}(t_1, t_2) = IC(MICA) \qquad (3)$$

Lin Method

$$sim_{Lin}(t_1, t_2) = \frac{2IC(MICA)}{IC(t_1)+IC(t_2)} \qquad (4)$$

Rel Method

This is a combination of Resnik's method and Lin's method. It uses the following equation to find out the similarity.

$$sim_{Rel}(t_1, t_2) = \frac{2IC(MICA)(1-p(MICA))}{IC(t_1)+IC(t_2)} \qquad (5)$$

Jiang Method

$$sim_{Jiang}(t_1, t_2) = 1 - min(1, IC(t_1) + IC(t_2) - 2IC(MICA)) \qquad (6)$$

At the end of this process, an N×N similarity matrix is formed where entries in the matrix quantify the similarity of node pairs.

Step2 – Computation of the Threshold value, Node Connection Score

Any Similarity matrix can be visualized as a network or as a graph. A graph is said to have community structure if it comprises of subsets of genes, with high cohesion within the subset having many edges connecting nodes of the same subset but few edges lying between subsets[37]. Finding communities within a graph is an efficient way to identify groups of related nodes. It is expensive if we construct the graph for all pairs of similarity index/distances, since for that we have (N* (N-1))/2 edges. The performance of the process can be improved by extracting the optimum connections from the graph. For that we followed PERT modeling, which is based on beta distribution and is useful in real-world analysis because it explains the degree of randomness of the dataset. Typically real dataset follows skewed distribution and beta distribution can handle a variety of skewness in data, both positive and negative efficiently[38]. An optimum threshold value, NCS is computed for identifying the optimum edge weights/node connections. The method uses three estimation parameters, minimum distance, maximum distance and average path length of each node. For each node 'i', a_i: denotes the shortest path of all connected nodes of node i; bi: denotes the largest path of all connected nodes of node i; mi: denotes average path of all connected nodes of node i, NCS is computed using the weighted average mean of all the three parameters for each node shown in equation (7). NCS for each observation reflects the degree of closeness of the nodes in terms of their attribute values

$$NCSi = \frac{ai+4mi+bi}{6} \qquad (7)$$

Step 3 – Partitioning the Semantic Similarity Graph into Communities

To produce reliable clusters, this approach prunes the similarity matrix, removing edges weights whose similarity indices are less than NCS. Edge weights with higher values represent a high degree of similarity between two nodes. The reduced Similarity matrix determines the intra cluster edges that represent the thickly associated components in the dataset. From the reduced similarity matrix, vertices are ordered based on the length of total connections on the nodes with each of the adjacent nodes. Pick the nodes from the sorted list and find the intersection of node connection with each of the node. Based on the confidence level set, remove or keep the nodes in the community list. Sub groups or communities are generated from the community list. Each community is distinguished by setting different colors for clusters from1 to k.

Implementation Of The Proposed Algorithm

The algorithm is implemented in R (http://www.r-project.org), a widely used open source platform for data analysis. Table 1 shows the notations and definitions used in ERCD for obtaining natural communities/clusters. Table 2 represents the pseudo code of the proposed algorithm and Table 3 represents the procedure for building the optimum graph from semantic similarity network using GO terms.

Table 1
Notations used in the Algorithm

D	Gene Dataset
N	Number of genes in the gene set
S	N x N similarity matrix, each cell quantifies the similarity between pair of genes
G	(V, E) be a weighted matrix .The set of nodes/genes and the set of edges of a graph G are denoted by V (G) and E (G), respectively
V(G)	{v1. . . vN} is the node set
E(G)	$\{e_{ij}\}$ is the edge set with an edge e_{ij} connects vertices v_i and v_j if they are adjacent or neighbors
W_{ij}	The positive weight of an edge (i, j) representing the amount of similarity between genes, $gene_i$ and $gene_j$.

Table 2
ERCD Algorithm

Input: D=Geneset, T1 = Confidence level
Output: k sub graphs or communities $G_1 \ldots G_K$ such that $G_i = (V_i, E_i)$, where $V_i \cap V_j = \emptyset \forall i \neq j$
Step 1: Graph G = *Generate_OptimumGraph (D)*
Step 2: For each node in Graph G, *NodeConnected_List[i] = getconnectednodes(G_i)*, i = 1 to N
Step 3: Sort the nodes of the Graph by descending order of their degree
Step 4: Select the *NodeConnected_List* of the first node (V_1) from the sorted list to form communities
Step 5: Compute confidence level of each node, $V_{k = 1 \text{ to } N}$ with the sorted *NodeConnected_List* using ($V_1 \cap V_k$)
Step 6: If confidence level > T_1; Put all the nodes into a community and remove node from the initial list of nodes
Else create a new community and add the nodes to the new list.
Step 7: Repeat Step 4 to 6 until the list is empty

Table 3
Generate Optimum Graph from Semantic Similarity Network

Generate_OptimumGraph(D)
Identify strongly connected nodes using Node Connection Score
Input: D= Geneset
Output**:** Optimum Weighted Graph, G
Begin

1. Compute the semantic similarity matrix S based on the biological knowledge source
2. For each row in S, i = 1 to N compute

   i. Edge with minimum value , Z_{iMin} = Minimum (S_{ij}) $\forall i \neq j$, j = 1 to N

   ii. Edge with Maximum Value, Z_{iMax}= Maximum (S_{ij})$\forall i \neq j$, j = 1 to N

   iii. Average Path Length, $\text{ZiAverage} = \frac{\sum_{i \neq j} d(i,j)}{\frac{n(n-1)}{2}}$

   iv. Similarity threshold value , $\text{ZiNCS} = \frac{ZiMin + ZiMax + 4 * ZiAverage}{6}$

3. # Remove edges which has reduced weight or similarity from Similarity matrix #
   For each row in S , i = 1 to N

   If ($\mathbf{S_{ij} < Z_{iNCS}}$) $\forall i \neq j$, j = 1 to N Delete $\mathbf{S_{ij}}$

4. Create Graph, G (V, E) using the reduced Similarity Matrix S, which contains only the connections with strong association between genes.

End

Experiment and Result Analysis

The proposed approach is evaluated using a set of 93 cancer genes from Illumina Inc., a research center focusing on the analysis of genetic variation and biological function. Information Content based Resnik approach is employed to find out the semantic similarity between the genes from the cancer gene list. Figure 2 represents the community structure formed using the proposed methodology. The performance of clustering results is validated using three main components based on the study conducted by Jiang *et al.*[39]; the precise judgment of the cluster quality, performance evaluation based on ground truth and evaluation of the reliability of the communities formed. Cluster quality can be measured in terms of homogeneity of data objects within the cluster and the separation of data objects between different clusters. Assessment based on ground truth involves comparison with respect to domain knowledge such as protein interactions or known gene functions relating to the clusters. Checking the PPI interactions of genes in each cluster reveals the network similarity. The final aspect of validation gives emphasis on the reliability of the clusters.

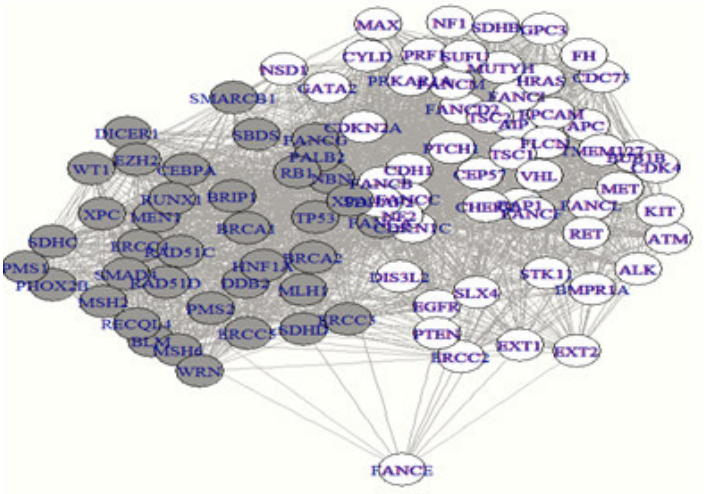

Figure 2
Automatic detection of two communities using ERCD Algorithm

Validation using Modularity measure

Modularity is one of the commonly used metric to evaluate the quality of network's partition into sub graphs called communities (Newman et al[40]; Newman and M. E.[41]). Graphs with a high value of modularity indicate a good community structure where nodes within groups are strongly connected whereas nodes in different group are sparsely connected. Modularity is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left[Aij - \frac{ki*kj}{2m}\right]\delta c_i\, c_j \qquad (8)$$

where m is the total number of edges in the network. A_{ij} represents the element of the adjacency matrix with row i and column j, k is the degree of node, c is the label of the community to which node is assigned and delta(x, y) is 1 if x=y and 0 otherwise. The performance of the proposed approach is compared with three different benchmark community detection algorithms like fast greedy, walktrap and leading eigenvector. Table 4 shows the modularity value and the size of clusters formed when different community detection algorithms are applied on the same cancer dataset. Modularity of new approach, ERCD is higher when compared to other algorithms and the result (Figure 2) clearly distinguishes the communities formed where as no clear distinction can be inferred from the outputs plotted using other algorithms (Figure 3). Identifying the significant edges, 1844 from a total edge count of 3796 significantly improves the performance of the proposed algorithm and provides an interactive exploration of the clustering results.

Table 4
Comparison of Modularity measure using different community detection algorithms

Algorithm	Modularity	Community Sizes	Edge count
ERCD Algorithm	0.1799317	55 , 38	1844
Fast greedy Algorithm	0.05338913	31 , 7, 3, 34 ,18	3796
Walktrap Algorithm	0.04921651	28, 54, 11	3796
Leading eigenvector Algorithm	0.04985941	48, 38, 7	3796

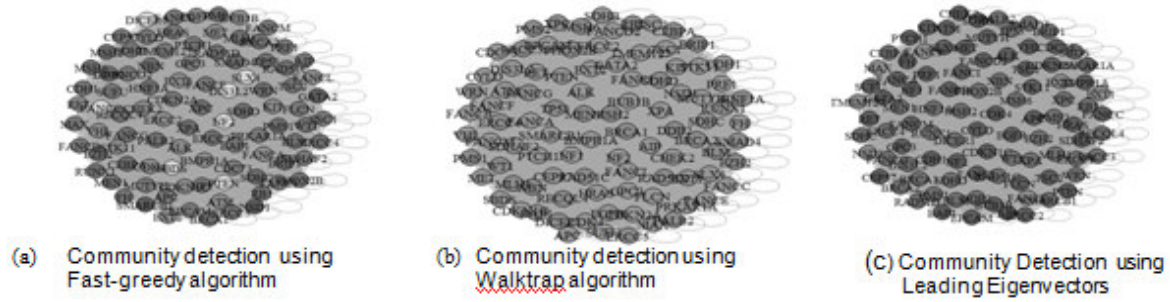

Figure 3
Communities formed when running the three algorithms Fast-greedy (a) Walktrap (b) and Leading Eigenvectors (c) on the cancer dataset

Validation of Clustered genes based on conformity with gene-gene interactions

Several measures have been identified to find the degree of similarity between clusters formed and the ground truth using indices like Rand, Jaccard coefficient, Minkowski etc. If N denotes the number of data objects and C denotes a set of k clusters {C_1, C2… C_k}, an N*N binary matrix B can be constructed where B_{ij} = 1 if object pairs D_i and D_j belong to the same cluster and B_{ij} = 0 otherwise. Similarly, a ground truth matrix M = {M_1,

M_2,….. M_s} can be constructed to represent the ground truth. Here we extracted the gene interactions of 93 cancer genes from NCBI's database (http://www.ncbi.nlm.nih.gov/gene) and checked the conformity of the clusters formed with the gene interactions. Total of 314 interactions were identified which has association with the genes in the training list. The underlying hypothesis is that the number of predicted clusters matching the gene interactions should be maximum for better clustering solutions. Accuracy is computed by taking the ratio of sum of the count of object pairs showing value "1" in both matrix B and M for each clusters to the total number of gene interactions. Table 5 and 6 lists the gene interactions in the each cluster. We got an accuracy of 78.34% for the clusters derived using ERCD which shows the conformity of the clustering solutions with the ground truth.

Table 5

Gene Interactions of cluster 1 genes that are associated with the genes in the training set. Total of 88 interactions were detected in cluster 1.

Genes	Count of interactions
AIP	2
EGFR	
RET	
APC	1
BUB1B	
ATM	5
CHEK2	
FANCD2	
FANCI	
STK11	
VHL	
BAP1	1
PTEN	
BUB1B	1
APC	
CDH1	3
EGFR	
MET	
PTEN	
CDK4	3
CDKN1C	
CDKN2A	
PRKAR1A	
CDKN1C	1
CDK4	
CDKN2A	2
CDK4	
VHL	
CHEK2	2
ATM	
VHL	

Genes	Count of interactions
FANCB	3
FANCC	
FANCL	
FANCM	
FANCC	5
FANCD2	
FANCE	
FANCF	
FANCL	
FANCM	
FANCD2	5
ATM	
FANCC	
FANCE	
FANCI	
FANCL	
FANCE	4
FANCC	
FANCD2	
FANCF	
FANCM	
FANCF	4
FANCC	
FANCE	
FANCL	
FANCM	
FANCI	3
ATM	
FANCD2	
FANCL	

Genes	Count of interactions
HRAS	1
NF1	
MET	2
CDH1	
EGFR	
NF1	1
HRAS	
NF2	1
TSC1	
PRKAR1A	2
CDK4	
EGFR	
PTCH1	2
EGFR	
EXT1	
PTEN	3
BAP1	
CDH1	
STK11	
RET	2
AIP	
EGFR	
STK11	2
ATM	
PTEN	
TSC1	2
NF2	
TSC2	
TSC2	1
TSC1	

Genes	Count of interactions
EGFR	7
AIP	
BMPR1A	
CDH1	
MET	
PRKAR1A	
PTCH1	
RET	
EXT1	2
EXT2	
PTCH1	
EXT2	1
EXT1	
FANCL	6
FANCB	
FANCC	
FANCD2	
FANCF	
FANCI	
FANCM	
FANCM	5
FANCB	
FANCC	
FANCE	
FANCF	
FANCL	
VHL	3
ATM	
CDKN2A	
CHEK2	
Grand Total	88

Table 6

Gene Interactions of cluster 2 genes that are associated with the genes in the training set. Total of 158 interactions were detected in cluster 2.

Genes	Count of interactions
BLM	11
BRCA1	
BRIP1	
FANCA	
FANCG	
MLH1	
MSH2	
MSH6	
NBN	
RAD51D	
TP53	
WRN	
BRCA1	15
BLM	
BRCA2	
BRIP1	
EZH2	
FANCA	
MLH1	
MSH2	
MSH6	
NBN	
PALB2	
PMS2	
RB1	
SMAD4	
SMARCB1	
TP53	
BRCA2	6
BRCA1	
FANCG	
PALB2	
PMS1	
PMS2	
TP53	
FANCA	5
BLM	
BRCA1	
ERCC4	
FANCG	
MSH2	

Genes	Count of interactions
BRIP1	5
BLM	
BRCA1	
MLH1	
PMS1	
PMS2	
CEBPA	4
HNF1A	
RB1	
RUNX1	
SMAD4	
DDB2	3
EZH2	
XPA	
XPC	
ERCC3	3
ERCC5	
TP53	
XPC	
ERCC4	4
FANCA	
FANCG	
MSH2	
XPA	
ERCC5	1
ERCC3	
EZH2	3
BRCA1	
DDB2	
WT1	
MSH2	11
BLM	
BRCA1	
ERCC4	
FANCA	
MLH1	
MSH6	
NBN	
PMS1	
PMS2	
TP53	
XPA	

Genes	Count of interactions
NBN	6
BLM	
BRCA1	
MLH1	
MSH2	
MSH6	
WRN	
PALB2	2
BRCA1	
BRCA2	
PMS1	5
BRCA2	
BRIP1	
MLH1	
MSH2	
MSH6	
PMS2	6
BRCA1	
BRCA2	
BRIP1	
MLH1	
MSH2	
MSH6	
RAD51C	1
RAD51D	
RAD51D	2
BLM	
RAD51C	
RB1	4
BRCA1	
CEBPA	
SMARCB1	
TP53	
RECQL4	1
TP53	
RUNX1	2
CEBPA	
SMARCB1	
SMAD4	2
BRCA1	
CEBPA	

Genes	Count of interactions
MLH1	8
BLM	
BRCA1	
BRIP1	
MSH2	
MSH6	
NBN	
PMS1	
PMS2	
XPA	4
DDB2	
ERCC4	
MSH2	
XPC	
WRN	3
BLM	
NBN	
TP53	
WT1	3
EZH2	
MEN1	
TP53	
XPC	5
DDB2	
ERCC3	
SMARCB1	
TP53	
XPA	
TP53	13
BLM	
BRCA1	
BRCA2	
ERCC3	
MEN1	
MSH2	
MSH6	
RB1	
RECQL4	
SMARCB1	
WRN	
WT1	
XPC	

Genes	Count of interactions
MSH6	8
BLM	
BRCA1	
MLH1	
MSH2	
NBN	
PMS1	
PMS2	
TP53	
FANCG	4
BLM	
BRCA2	
ERCC4	
FANCA	
HNF1A	1
CEBPA	
MEN1	2
TP53	
WT1	
SMARCB1	5
BRCA1	
RB1	
RUNX1	
TP53	
XPC	
Grand Total	158

Validation based on Silhouette Index

Quality of the clustering is also measured using the internal validation index, Silhouette Index[42].For a given community or cluster, G_i (i = 1...k), silhouette width is a quality measure that indicates the membership of a sample in a cluster. It's defined as

$$s = \frac{b-a}{\max(a,b)} \qquad (9)$$

where a is the average distance between a sample and all other data points in the same cluster and b is the average distance between a sample and all other data points in the neighboring clusters. A higher Silhouette score relates to a model with better defined clusters. K-means clustering is applied on the same dataset by setting k = 2 to 10 to extract the optimum Silhouette index. Silhouette index is maximized in the second cluster solution with an average silhouette width of 0.64 shown in Figure 4. The automatic detection of two clusters using our approach shows a better Silhouette index of 0.754 which shows the performance of the new clustering solution.

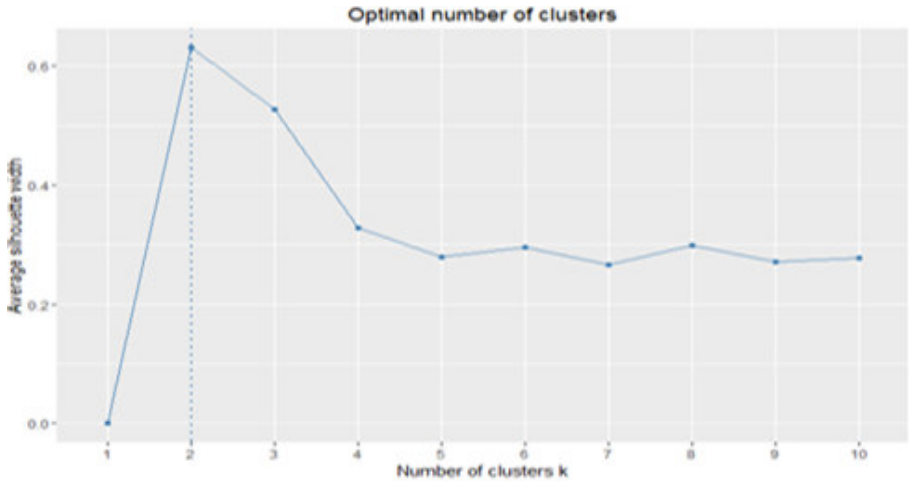

Figure 4

Average Silhouette Width of cancer dataset for different cluster sizes

CONCLUSION AND FUTURE WORK

Many networks in biological system, including gene networks are found to divide naturally into communities or clusters. Discovering and characterizing this community structure is one of the major challenges in the study of graph/networked systems. The reliance of many clustering solutions depends on the prior knowledge of cluster or community size, which degrades the performances of the system. Edge Reduction and Community Detection Algorithm overcomes this by exploring the communities automatically using the edge reduction approach which gives a true set of distinguished clusters. Results from ERCD is compared with existing community detection algorithms like Fast greedy, Walk trap, Leading eigenvector and K-means clustering to demonstrate the data clustering capability. Different validation measures are used to check the performance of the system. The proposed algorithm exhibits excellent performance, with an added advantage of low computational complexity that enables one to analyze large systems. As an enhancement we propose to use integrated knowledge to extract functionally similar genes that contribute to a single phenotype or disease type from the set of promising candidate genes.

Abbreviations

GO Gene Ontology
PIP Protein-Protein Interactions
ERCD Edge Reduction and Community Detection
MF Molecular Function
BP Biological Process
CC Cellular Component
DAG Directed Acyclic Graphs
NCS Node Connection Score

FUNDING/ACKNOWLEDGEMENT

This work is supported by the Cognitive Science Research Initiative (CSRI) of the Department of Science and Technology (DST), Government of India, as part of the funded Project, SR/CSI/81/2011 at Department of Computer Science, School of Arts and Sciences, Amrita University, Kochi.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics. 2003 Mar 1;33:228-37.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53.
3. Aravind L. Guilt by association: contextual information in genome analysis. Genome Research. 2000 Aug 1;10(8):1074-7.
4. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006 Sep 15;22(18):2291-7.
5. Jonsson PF, Cavanna T, Zicha D, Bates PA. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC bioinformatics. 2006 Jan 6;7(1):2.
6. Cai JJ, Borenstein E, Petrov DA. Broker genes in human disease. Genome biology and evolution. 2010 Jan 1;2:815-25.
7. Rahmani H, Blockeel H, Bender A. Predicting genes involved in human cancer using network contextual information. Journal of Integrative Bioinformatics (JIB). 2012 Mar 1;9(1):44-71.
8. Pan G, Zhang W, Wu Z, Li S. Online community detection for large complex networks. PloS one. 2014 Jul 25;9(7):e102799.
9. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. Physical review E. 2004 Dec 6;70(6):066111.
10. Newman ME. Modularity and community structure in networks. Proceedings of the national academy of sciences. 2006 Jun 6;103(23):8577-82.
11. Newman ME. Fast algorithm for detecting community structure in networks. Physical review E. 2004 Jun 18;69(6):066133.
12. Pons P, Latapy M. Computing communities in large networks using random walks. InInternational Symposium on Computer and Information Sciences 2005 Oct 26 (pp. 284-293). Springer Berlin Heidelberg.
13. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences. 2003 Oct 14;100(21):12123-8.
14. Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. Information processing letters. 2000 Dec 31;76(4-6):175-81.
15. Pržulj N, Wigle DA, Jurisica I. Functional topology in a network of protein interactions. Bioinformatics. 2004 Feb 12;20(3):340-8.
16. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006 Mar 30;440(7084):637-43.
17. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic acids research. 2002 Apr 1;30(7):1575-84.
18. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics. 2006 Apr 15;22(8):1021-3.
19. Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, Li G. Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic acids research. 2003 May 1;31(9):2443-50.
20. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. PROTEINS: Structure, Function, and Bioinformatics. 2004 Jan 1;54(1):49-57.
21. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC bioinformatics. 2005 Mar 1;6(1):39.
22. Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, Siani-Rose MA. A knowledge-based clustering algorithm driven by gene ontology. Journal of biopharmaceutical statistics. 2004 Dec 29;14(3):687-700.
23. Dotan-Cohen D, Kasif S, Melkman AA. Seeing the forest for the trees: using the gene ontology to restructure hierarchical clustering. Bioinformatics. 2009 Jul 15;25(14):1789-95.
24. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004 Jan 1;32(suppl 1):D258-61.
25. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007 Nov 1;25(11):1251-5.
26. Jiang R, Gan M, He P. Constructing a gene semantic similarity network for the inference of disease genes. BMC systems biology. 2011 Dec 14;5(2):S2.
27. Vafaee F, Rosu D, Broackes-Carter F, Jurisica I. Novel semantic similarity measure improves an integrative approach to predicting gene functional associations. BMC systems biology. 2013 Mar 14;7(1):22.
28. Arnaud MB, Costanzo MC, Shah P, Skrzypek MS, Sherlock G. Gene Ontology and the annotation of pathogen genomes: the case of Candida albicans. Trends in microbiology. 2009 Jul 31;17(7):295-303.
29. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nature Reviews Genetics. 2008 Jul 1;9(7):509-15.
30. Lovering RC, Dimmer EC, Talmud PJ. Improvements to cardiovascular gene ontology. Atherosclerosis. 2009 Jul 31;205(1):9-14.
31. De Bodt S, Proost S, Vandepoele K, Rouzé P, Van de Peer Y. Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-

expression. BMC genomics. 2009 Jun 29;10(1):288.
32. Pandit S, Gupta S. A comparative study on distance measuring approaches for clustering. International Journal of Research in Computer Science. 2011 Jan 1;2(1):29.
33. Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR). 1999 Jul 11;11:95-130.
34. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008. 1997 Sep 20.
35. Lin D. An information-theoretic definition of similarity. InICML 1998 Jul 24 (Vol. 98, No. 1998, pp. 296-304).
36. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics. 2006 Jun 15;7(1):302.
37. Girvan M, Newman ME. Community structure in social and biological networks. Proceedings of the national academy of sciences. 2002 Jun 11;99(12):7821-6.
38. Moitra SD. Skewness and the beta distribution. Journal of the Operational Research Society. 1990 Oct 1;41(10):953-61.
39. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. IEEE Transactions on knowledge and data engineering. 2004 Nov;16(11):1370-86.
40. Newman ME. J, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review E.;69(2):026113.
41. Newman ME. Modularity and community structure in networks. Proceedings of the national academy of sciences. 2006 Jun 6;103(23):8577-82.
42. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987 Nov 1;20:53-65.
43. Ashok S, Judy MV. A novel iterative partitioning approach for building prime clusters. International Journal of Advanced Intelligence Paradigms. 2015;7(3-4):313-25.
44. MALLICK PK, MISHRA D, PATANAIK S, SHAW K. A novel supervised gene clustering approachby mining interdependent gene patterns. International Journal of Pharma and Bio Sciences.;7.
45. Sonumol NS, Uma VR, Ashok S, Judy MV. Community detection in multidimensional genomic dataset. International Journal of Artificial Intelligence™. 2015 Sep 22;13(2):109-17.