CHAPTER 6

FUNCTIONAL CATEGORIZATION OF GENES BASED ON SPECTRAL GRAPH THEORY AND INFORMATION FUSION

6.1. Introduction

Edge pruning and community detection technique presented earlier demonstrated better clustering efficiency when compared to conventional approach of selecting entire gene association for analysis. It was also observed that the complex dependency relationships of various genomic features are better modeled using graph properties and is considered as the efficient technique for exploring biologically relevant and accurate gene group. Graph clustering based on spectral methods is proved to be a more reliable and effective approach for finding meaningful communities and it is appropriate for wide range of data types and similarity functions (White and Smyth 2005). Spectral clustering utilizes information derived from the eigenvalues and eigenvectors of their associated graph matrix for forming meaningful subgraphs. One remarkable advantage of spectral clustering is its simplicity and the ability to group data objects based on their ‘relatedness’ rather than using their geometric closeness. Flexibility of extracting clusters with arbitrary shapes is another major advantage of spectral clustering, which is a significant requirement of many realistic problems. It can be applied for large datasets and proved to be more efficient than traditional clustering approaches (Luxburg 2007). However there are few open issues observed related to spectral clustering while deciding the set of parameters for implementation like

1. Defining a suitable similarity measure.
2. Automatic determination of the number of clusters.
3. Significance of eigenvector subset selection and determining the key eigenvectors.
In an attempt to overcome these issues and for partitioning the gene network into biologically relevant group of genes, an enhanced clustering process using spectral graph theory is proposed in this chapter by incorporating multiple preprocessing and optimization steps.

Integrative approaches that consolidate several types of information are being perceived as a way to gain comprehensive view of the disorders. Grouping information from a wide range of biological sources broadens the scope of applications and has turned out to be tremendously valuable at determining a variety of insights such as shared functions or regulations. Following two main challenges has to be addressed when integrating multiple knowledge sources. Firstly, construction of a similarity matrix depends on the definition of suitable similarity function for each knowledge source. For that knowledge based semantic similarity measures have been utilized for each knowledge source for better accuracy. Secondly, linear aggregation of pair wise similarity of genes for all knowledge sources can be challenging due to the individual significance of each information source in defining the functional associations, for which feature scoring is implemented to allocate weights for each information source. KEGG pathways are also incorporated into this analysis in addition to the three knowledge sources, Gene Ontology, MeSH and Disease annotations for improved performance.

The efficiency of spectral clustering is primarily based on the assumption that it doesn't make any presumption on the form or shape of the clusters. This property originates from the mapping of the initial data space to an eigenspace. Performance differs fundamentally in the quantity of eigenvectors used for clustering. When large datasets are analyzed, similarity graph construction and eigenvector decomposition is computationally complex. Hence it is important to design strategies to keep only significant associations or relationships in the similarity graph. Here a sparse matrix is constructed from the similarity matrix such that each gene is connected to a small number of its significant neighbors and extracts a small subspace of highly significant eigenvectors for post processing. In the previous chapter, the optimum threshold for
edge pruning was defined based on a weighted average that uses both the average and extreme edge weights. One drawback of this process is that an evaluation of the best or worst case scenario may skew the value so much that the evaluation may not be precise. To efficiently deal with this challenge, the significance of each edge is measured based on its statistical significance rather than using a weight threshold and the insignificant edges are automatically pruned to form an optimized gene semantic similarity network.

Gene clustering based on spectral graph theory defined in this chapter has three main objectives. The first aim is to build an aggregated association score for representing the pair wise similarity of genes. The next aim is to introduce an edge pruning method for removing insignificant associations between genes and the third objective is eigenvector subset selection using entropy measures for improving cluster efficiency. Once the appropriate gene clusters are derived, it’s possible to explore distinct patterns within each subgroup and can further investigate information about functional similarities. Different gene similarity measures are analyzed to compute the similarity scores between genes and the clusters are systematically evaluated for strong associations. The biological significance of these newly identified clusters is also evaluated by using internal and external validation measures.

**Major highlights of this work**

- Graph-based clustering approach for exploring functionally couples genes.
- Fused knowledge framework based on MeSH terms, gene to disease associations, gene ontology, KEGG pathways helps in extracting biologically enriched gene clusters.
- An edge pruning strategy is proposed to significantly reduce the computational complexity.
- Automatic clustering using eigenvector subset selection based on entropy measures enhances the performance of the system.
- Ground truth validation to assess the biological significance of clusters formed.
6.2. Spectral Clustering Algorithm

The general approach to spectral clustering is based on the principle of forming a gene network from the biological information where each node represents a gene and edge weight represents the gene interactions in terms of its similarity measure. The eigenvectors and eigenvalues of the matrix derived from such a network are closely related to the connectivity of the graph, which helps in partitioning the graphs. Spectral Clustering algorithm introduced by Ng et al. 2001 consists of three main stages. They are

- **Preprocessing:** The algorithm builds a similarity graph using the data objects as nodes and edge weights as pair wise similarity between nodes.

- **Spectral Mapping:** Laplacian Matrix associated to the similarity graph is constructed first. Then eigenvector decomposition is done to study the graph spectrum formed. Different definitions of the Laplacian Matrix includes
  - Unnormalized Spectral Clustering: Laplacian matrix is defined as
    \[
    L = D - W
    \]  
    (6.1)
  - Normalized Spectral Clustering: Laplacian matrix is defined as
    \[
    L_{symmetric} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}
    \]  
    (6.2)

    where I is the identity matrix ; D represents degree matrix which is an N*N diagonal matrix whose (i,i) element represents the total weight of edges incident to node \( x_i \), \( D(i, i) = \sum_j w_{ij} \) and W represents the similarity graph.

- **Post Processing:** Building clusters from the eigenvectors extracted using standard algorithms.
Properties of Normalized Laplacian matrix, $L$ with eigenvalues $\lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1}$ are (Auffarth 2007)

- $L$ is always positive-semi definite ($\forall i, \lambda_i \geq 0$).
- The multiplicity, $k$ of the eigenvalue 0 represents the number of connected components of the graph.
- $\lambda_1$ is termed the algebraic connectivity.
- The smallest non-trivial eigenvalue of $L$ is called the spectral gap.

Properties of Unnormalized Laplacian symmetric matrix, $L_{\text{symmetric}}$ includes

- $L_{\text{symmetric}}$ is always positive semi-definite and have $n$ non-negative real-valued eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$.
- The multiplicity, $k$ of the eigenvalue 0 represents the number of connected components in the graph.

The algorithm proposed in this chapter uses eigenvector decomposition of unnormalized similarity matrix and uses its properties to partition the genes.

6.3. Gene clustering based on spectral graph theory and information fusion

The current algorithm extends spectral clustering algorithm to provide fused learning of multiple knowledge sources. Given a set of genes and a measure of similarity between two genes, the objective is to identify functionally related genes where genes in the same cluster are closely related to each other and perform similar function where as genes in separate clusters are maximally far apart from each other. A complex disorder is imagined as a set of gene clusters $C_1, C_2 \ldots C_n$ where each gene cluster $C_k$ is defined as a set of genes that are grouped by their functional similarity where sequence of at least one of them when critically modified can contribute to the severity of the disease. Weighted graphs are used to model the gene interactions where $G = (V, E, W)$ where $V$ is the set of nodes or vertices that represent the genes, $E$ is the set of edges
corresponding to the functional relationship between pairs of genes. The degree / strength of the relationship between two vertices is denoted as the edge weights, $W$. The proposed clustering solution includes the following steps:

**Step 1: Information fusion**

Using all the six knowledge sources, MeSH, Molecular Function Ontology, Biological Process Ontology, Cellular Component Ontology, DisGeNET and KEGG Pathways, the first step is to construct a fused similarity network of gene-gene interactions whose vertices include all the genes and edges represents the aggregated weight of all gene pairs. The aggregated scoring is done based on weighted sum of similarity scores. Weights are assigned by evaluating the proportion of variance of each knowledge source. The process flow is detailed in Figure 6.1 and step wise execution is mentioned below.

![Fig 6.1: Fused semantic similarity network of gene-gene interactions](image-url)
**Step 1.1: Semantic Similarity Computation for different data sources**

Given a set of genes \{g_1, g_2, g_3 \ldots g_N\}, use the semantic similarity function to compute similarity of all genes and form N*N similarity matrix where each cell \((i, j)\) represents the similarity of gene pairs \(g_i\) and \(g_j\). The entries in the matrix quantify the strength of the similarity between genes.

**Measuring the similarity between MeSH terms**

Genes can be rendered as a collection of MeSH terms. Gene2MeSH (Ade et al. 2007) utilizes a measurable way to reliably deal with concepts defined in MeSH, the National Library of Medicine's controlled vocabulary for biology and medicine. Two genes \(\text{Gene}_1\) and \(\text{Gene}_2\) can be represented as a set of terms \(\text{Gene}_1 = \{M_{11}, M_{12} \ldots , M_{1i}, \ldots , M_{1n}\}\) and \(\text{Gene}_2 = \{M_{21}, M_{22} \ldots , M_{2j}, \ldots , M_{2m}\}\). Weighted Jaccard similarity measures (Hamers et al. 1989; Manning et al. 2008) can be computed as the sum of the score of each MeSHID in the intersection of MeSH terms of \(\text{Gene}_1\) and \(\text{Gene}_2\) divided by sum of the score of each term in the union of MeSH terms of \(\text{Gene}_1\) and \(\text{Gene}_2\). Similarity between two genes \(\text{Gene}_1\) and \(\text{Gene}_2\) based on MeSH terms is represented as \(\text{MeSHSim}\) and is defined using equation (6.3)

\[
\text{MeSHSim}(\text{Gene}_1, \text{Gene}_2) = \frac{\sum_{k \in \text{MeSHIds}(\text{Gene}_1) \cap \text{MeSHIds}(\text{Gene}_2)} \text{Score}(k)}{\sum_{k \in \text{MeSHIds}(\text{Gene}_1) \cup \text{MeSHIds}(\text{Gene}_2)} \text{Score}(k)} \quad (6.3)
\]

\(0 \leq \text{MeSHSim}(\text{Gene}_1, \text{Gene}_2) \leq 1\)

**Measuring the similarity between disease terms**

Similar to MeSH terms, \(\text{Gene}_1\) and \(\text{Gene}_2\) can be represented as a set of disease IDs such as \(\text{Gene}_1 = \{D_{11}, D_{12}, \ldots , D_{1i}, \ldots , D_{1n}\}\) and \(\text{Gene}_2 = \{D_{21}, D_{22}, \ldots , D_{2j}, \ldots , D_{2m}\}\). Weighted Jaccard similarity measures can be used to compute the semantic similarity of two genes by taking the sum of the score of each disease IDs in the intersection of Disease terms of \(\text{Gene}_1\) and \(\text{Gene}_2\) divided by sum of the score of each disease IDs in the union of Disease terms of \(\text{Gene}_1\) and \(\text{Gene}_2\) as shown in equation (6.4)
DiseaseSim(Gene1, Gene2) = \frac{\sum_{k \in DiseaseIds(Gene1) \cap DiseaseIds(Gene2)} \text{Score}(k)}{\sum_{k \in DiseaseIds(Gene1) \cup DiseaseIds(Gene2)} \text{Score}(k)} \tag{6.4}

0 \leq \text{WeightedDiseaseSim} (\text{Gene 1, Gene 2}) \leq 1

**Measuring the similarity between GO terms**

For computing the similarity of genes based on GO terms, graph based distance measure proposed by *Wang et al. (2007)* is used where the semantic closeness is processed by considering the topology of GO graph taking into account all parent terms. Wang’s measure yields more precise outcomes than Resnik's measure in clustering gene pairs. Hence GOSemSim, an R package with parameter as Wang’s measure is used to compute pair wise similarity between gene with respect to all the three ontologies, molecular function, cellular component and biological process. Semantic similarity between two terms g1 and g2 in Wang measure is defined using the equation (6.5)

\[
\text{Sim}_\text{Wang} (g_1, g_2) = \frac{\sum_{p \in (P_{g1} \cap P_{g2})} (S_{g1,p} + S_{g2,p})}{\sum_{g1 \in P_{g1}} S_{g1,p} + \sum_{g2 \in P_{g2}} S_{g2,p}} \tag{6.5}
\]

Where g1 and g2 denotes gene ontology terms, Pg1 and Pg2 are the sets of all parents of g1 and g2 GO terms. Sg,p is the maximal semantic contribution of the paths from g to p.

**Measuring the similarity between Pathway terms**

Jaccard similarity measure is used to compute the similarity of two genes by taking the ratio of the intersection of pathways associated with both genes to the union of pathways associated with both genes, defined using equation (6.6)

\[
\text{PathwaySim} = \frac{\text{PathwayIds}(\text{Gene1}) \cap \text{PathwayIds}(\text{Gene2})}{\text{PathwayIds}(\text{Gene1}) \cup \text{PathwayIds}(\text{Gene2})} \tag{6.6}
\]

**Step 1.2: Feature scoring based on analysis of variance**

Distinctive knowledge sources have diverse level of relationship with functional similarity. Therefore to achieve efficient integration, it is important to allocate weights for each information source in light of some basic measuring stick. Different
unsupervised feature ranking methods are widely used in applications, including saliency, smoothness, dependability, entropy etc (Guyon and Elisseeff 2003). A similar method is adopted in this approach where each information source will carry a weight value based on the significance of the attribute with respect to a defined criterion. Edges are considered as data objects and similarity measure of edges based on each knowledge source is considered as features. A variable is relevant or significant if it has a high variance or a vast range, contrasted with others (Ding, C. H. 2003). Proportion of variance is used as a measure for ranking and scoring the knowledge sources using equations (6.7) - (6.10)

\[
\text{Variance} = \frac{\text{SSTR}(\text{Treatment Sum of squares})}{\text{SST}(\text{Total Sum of Squares})}
\]

(6.7)

\[
\text{SST} = \text{SSE}(\text{Error sum of squares}) + \text{SSTR} 
\]

(6.8)

\[
\text{SSE} = \sum_{i=1}^{n} (X_i - \bar{X})^2 
\]

(6.9)

\[
\text{SSTR} = \sum_{i=1}^{k} n_i (\bar{X}_i - \mu)^2 
\]

(6.10)

Where \( n \) is the number of observations, \( k \) is the number of features, \( \mu \) is the overall mean of the dataset. The normalized variance scores are assigned as weights of each knowledge source for further analysis.

**Step 1.3: Integrated semantic similarity network using data fusion**

Pair wise similarity of genes with respect to each knowledge source is represented as an N*N similarity matrix, N is the number of genes. Multiple similarity matrices generated from heterogeneous data sets are integrated into a solitary one by taking a linear weighted sum of the semantic similarity matrices using equation (6.11)

\[
S_{\text{Integrated}} = m_1S_1 + m_2S_2 + \cdots + m_nS_n 
\]

(6.11)

where \( S_i \) is the similarity measure associated with the \( i^{th} \) data source, \( n \) is the number of knowledge sources used, \( m_i \) is the weights assigned to each knowledge source. The simplicity of this approach makes it exceptionally adaptable to wide amount of information. Every information source can be progressively re-weighted as more
information is distinctly accessible. At the end of this process, a fused semantic similarity matrix, $S_{N \times N}$ is formed. This matrix is converted into a semantic similarity graph using R (igraph) package.

**Step 2: Simplifying semantic similarity network using Edge Pruning**

The parameters that directly impact the computational complexity in a graph are the number of nodes “$n$” and the number of edges “$m$”. If the number of edges $m \gg$ the number of nodes $n$, the dispersion of edges among the nodes is exorbitantly homogeneous for the clusters to make meaningful inferences. To efficiently deal with the performance of a graph is to reduce the size of the graph while retaining the significant information (Guzzi et al. 2013). An optimized edge pruning technique is proposed for building compact gene network using statistically significant factor by identifying the critical edges based on a threshold value.

A threshold value, $\tau$ is derived for each node to ensure the degree of relationship with its connected nodes using the equation (6.12)

$$\tau = \mu + \alpha \times \sigma \quad (6.12)$$

where $\mu$ and $\sigma$ represents the mean and the standard deviation of the weights of the adjacent nodes.

Low standard deviation specifies that the data points will be closer to the mean value where as a high standard deviation indicates that the data points are spread out over a wider range of values. “$\alpha$” represents the sigma performance levels. It can take any numeric values from 1 to 6. Different threshold values are computed for each gene to measure its statistical significance and for reducing the noise and potential bias towards numerous associations with weak connections. This ensures that the edges identified are statistically significant for producing functionally related communities. Thus the pruning process generates a new semantic similarity network from the existing network by removing edges whose semantic similarity score is below the threshold value $\tau$. 
Step 3: Post processing the eigenvectors to find gene clusters

The main objective of any clustering solution is to partition the dataset into homogenous groups of data objects with maximum intracluster similarity and minimum intercluster similarity. Gene communities are derived using a k-way partitioning by finding eigenvector decomposition of Laplacian matrix \( L = N^*N \) such that \( L = D - A \) (Degree – Similarity) and using k-means algorithm on selected eigenvectors to form clusters. Degree matrix, \( D = N^*N \) holds the information about the degree of each node where

\[
D (i, j) = 0 \text{ if } i \neq j \\
D (i, j) = \text{degree} (v_i) \text{ if } i = j
\]

Every eigenvector incorporate information about all clusters, however only some of them are truly enlightening and prompt to improve the efficacy of the clustering (Varshavsky et al. 2006). Hence to get an optimized subset of eigenvector combination \( V^i \), \( i = 1….m \) that upgrades the clustering quality instead of considering all eigenvectors, a feature selection method called Entropy is utilized, which is a statistical measure of randomness or disorder in a system for positioning and weighing the attributes in a dataset. Shannon entropy is the widely used and highly recognized measure of uncertainty or irregularity in an information source (Alvim, Andrés and Palamidessi, 2010; Lima et al. 2012). To assess the significance and get the priority list of eigenvectors, each eigenvector is excluded from the list and the entropy of remaining set of eigenvectors is processed using equation (6.13).

\[
E = - \sum_{i=1}^{n} P_i \log_2 P_i
\]

(6.13)

where \( P_i \) is the probability of occurrence of a data i.

If the ejection of an eigenvector creates more issues in the framework than another, it indicates more significance and higher rank of this eigenvector (Dash et al. 2000). Since the principal components of the dataset are the eigenvectors with the largest
eigenvalues, for feature detection only the top positioning eigenvectors are chosen for computing entropy. Eigenvectors with higher entropy scores are selected for post processing. The selected eigenvectors encode knowledge about the groups, which can be utilized by k-means algorithm for partitioning the genes. Zero\textsuperscript{th} eigenvalue explains whether the graph is connected or not and if a graph has k connected components, then eigenvalue 0 has got a multiplicity k (Marsden, A. 2013). The same approach is used for determining the cluster size, k. The process is detailed in Figure 6.2.

Fig 6.2: Matrix decomposition and eigenvector selection

6.4. Pseudo code of the proposed algorithm

Table 6.1: Psuedo code of Spectral graph clustering

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Create a pair wise similarity matrix N\times N for each information source</td>
</tr>
<tr>
<td>2</td>
<td>Form the fused weighted semantic similarity matrix, W based on significance of knowledge source</td>
</tr>
<tr>
<td>3</td>
<td>Form optimized matrix Ws using Edge pruning technique</td>
</tr>
<tr>
<td>4</td>
<td>Define diagonal matrix, D</td>
</tr>
<tr>
<td>5</td>
<td>Form the matrix, L = D - Ws</td>
</tr>
<tr>
<td>6</td>
<td>Find X1, X2,...,Xp, the p largest eigenvectors of L</td>
</tr>
<tr>
<td>7</td>
<td>Set the cluster size k as the number of ‘0’ eigenvalues</td>
</tr>
<tr>
<td>8</td>
<td>Identify the optimum eigenvectors k using Entropy measure</td>
</tr>
<tr>
<td>9</td>
<td>Form the columns of the new matrix X where the dimension has reduced from NxN to Nxk</td>
</tr>
<tr>
<td>10</td>
<td>Cluster into k clusters via K-means</td>
</tr>
</tbody>
</table>
6.5. Experiment and Result Analysis

6.5.1. Clustering of autism genes using proposed approach

ASD top priority 654 genes (Appendix A.1) from the prioritized set of autism genes are considered for analysis.

6.5.1.1. Similarity computation based on all knowledge sources

Total of 30331 MeSH terms from gene2MeSH and 62544 diseaseIDs from DisGeNET that are associated with the disease genes are extracted. Similarity 3205 pathways associated with the genes are extracted from KEGG database. Pair wise similarities of genes are computed using weighted Jaccard’s index for both MeSH and gene-disease annotations and simple Jaccard’s index for pathways identifiers. For computing the similarity of genes based on GO terms, Wang’s graph based distance measure is used. The similarity score were stored in an N*N similarity matrix for each knowledge source.

6.5.1.2. Information fusion

The similarity score of all edges with respect to each knowledge source is given as input for computing the proportion of variance. Figure 6.3 explains the variance measured for each knowledge source. The last three knowledge sources are enough to explain 80% of the variance in the original dataset.
The variance score of each feature is taken as the weight factor. A linear weighted sum of the semantic similarity matrices is computed which gives an integrated semantic similarity matrix that represents the fused knowledge of all information sources.

6.5.1.3. Optimizing the semantic similarity network

The outcomes are tested utilizing diverse estimations of threshold value for edge pruning; with $\alpha = 1$, 2 and 3. The edge weights that are less than the threshold value, $\tau$ is removed for constructing the optimized semantic similarity matrix. Table 6.2 lists the number of edges obtained for each threshold setting.

Table 6.2: Edge count based on threshold setting

<table>
<thead>
<tr>
<th>Optimizing parameter</th>
<th>$\alpha = 1$</th>
<th>$\alpha = 2$</th>
<th>$\alpha = 3$</th>
<th>Without Edge pruning</th>
</tr>
</thead>
<tbody>
<tr>
<td>No: of Edges</td>
<td>40219</td>
<td>11423</td>
<td>3360</td>
<td>213334</td>
</tr>
</tbody>
</table>

Fig 6.3: Proportion of variance explained by each knowledge source
6.5.1.4. Eigenvector selection based on Entropy measure

After the eigenvector decomposition, it is vital to identify the optimal eigenvectors that clearly represent the characteristics of data. For each $\alpha$, different eigenvectors are generated and the optimum set is selected using Entropy measure. Figure 6.4 represents the ranking of eigenvectors based on different threshold settings for edge pruning and without edge pruning. It is preferred to utilize the top ranked eigenvectors for clustering rather than random selection for better performance.

Fig 6.4 (a)

Fig 6.4 (b)
6.5.2. Cluster validation based on ground truth derived from gene interactions

After obtaining the distinct non-trivial eigenvectors, k-means clustering algorithm was applied on the eigenspace by setting k equal to the number of zero eigenvalues. k-value varies based on the testing scenarios used. The performance of the clustering solutions are examined using eigenvector selection for four different cases; 3 test cases with edge pruning for threshold settings $\alpha = 1, 2$ and $3$ and one case without edge pruning. True
positives are computed by taking the number of genes that are correctly detected in each cluster. When a gene is not identified in the cluster expected, that will be a false negative observation. When a gene is detected where it shouldn’t be, it will be false positive and when a gene is not detected where it shouldn’t be are true negative observations. To decide if a gene is correctly marked in a cluster / community, the standard measure, Rand Index is used which is the proportion of sets of genes accurately classified out of all possible combinations and it measures the probability of a gene being effectively characterized based on the ground truth. The cluster efficiency is also assessed using Silhouette Index, $88.82\%$ which is better for threshold $\alpha = 2$. Validation using biological significance has also demonstrated that, the proposed approach with $\alpha = 2$ resulted in forming significant groups with maximum efficiency of $86.59\%$. The significant improvement in Rand Index clearly exhibits the efficiency of the clustering process in deriving distinctive partitions of biologically significant groups. Table 6.3 shows the performance index for different settings and the cluster membership vector derived for $\alpha = 2$ is represented in Appendix A.3.

Table 6.3: Performance comparison of proposed approach at different test scenarios using biological significance and cluster efficiency

<table>
<thead>
<tr>
<th>Optimizing parameter</th>
<th>Proposed approach</th>
<th>Without Edge Pruning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 1$</td>
<td>$\alpha = 2$</td>
<td>$\alpha = 3$</td>
</tr>
<tr>
<td>No of connected components</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Rand Index</td>
<td>0.7600377</td>
<td><strong>0.8659352</strong></td>
</tr>
<tr>
<td>Silhouette Index</td>
<td>0.7997412</td>
<td><strong>0.8881561</strong></td>
</tr>
</tbody>
</table>
The clusters derived from the gene semantic similarity network using the proposed approach with optimum threshold value $\alpha = 2$ is represented in Figure 6.5(a) - (i).

Fig 6.5 (a) – Cluster1 (104)                      Fig 6.5 (b) Cluster2 (129)

Fig 6.5 (c) – Cluster 3 (105)                Fig 6.5 (d) Cluster 4 (39)
Fig 6.5 (e) – Cluster 5(38)  

Fig 6.5 (f) – Cluster 6(61)  

Fig 6.5 (g) – Cluster 7(65)  

Fig 6.5 (h) – Cluster 8(64)
Fig 6.5 (i) Cluster 9(49)

Fig 6.5: Clusters derived using spectral graph theory in ASD genes. Fig (a) – (i) represent clusters 1 to 9 with sizes 104, 129, 105, 39, 38, 61, 65, 64 and 49 respectively.

### 6.5.3. Performance comparison of the proposed solution with k-means algorithm

To compare the performance of proposed approach with standard k-means algorithm, Rand Index for ground truth checking and Silhouette index for cluster efficiency is computed for different parameters. Starting with \( k = 2 \), and increases by 1 on each iteration, clusters are derived using k-means algorithm from the aggregate similarity matrix. Table 6.4 presents the average Silhouette index for cluster numbers \( k=2 \ldots 20 \) for k-means clustering on the fused gene semantic similarity matrix. Edge pruning was not applied here. The results show that the Silhouette index is maximum of cluster size \( = 2 \) with index value 0.7997 and Rand index is highest for cluster size \( = 13 \) with value 0.8132. The proposed approach with \( \alpha = 2 \) shows better results than the k-means with a Silhouette index of 0.8882 and Rand Index of 0.8659.
Table 6.4: Performance comparison using standard k-means on fused semantic similarity matrix

<table>
<thead>
<tr>
<th>cluster size (k)</th>
<th>Silhouette Index</th>
<th>Rand Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.799741361</td>
<td>0.344502672</td>
</tr>
<tr>
<td>3</td>
<td>0.713010342</td>
<td>0.419051566</td>
</tr>
<tr>
<td>4</td>
<td>0.666799832</td>
<td>0.555076453</td>
</tr>
<tr>
<td>5</td>
<td>0.710978105</td>
<td>0.570689314</td>
</tr>
<tr>
<td>6</td>
<td>0.625416009</td>
<td>0.672759459</td>
</tr>
<tr>
<td>7</td>
<td>0.699844212</td>
<td>0.625511052</td>
</tr>
<tr>
<td>8</td>
<td>0.680418082</td>
<td>0.717539842</td>
</tr>
<tr>
<td>9</td>
<td>0.733345068</td>
<td>0.777275899</td>
</tr>
<tr>
<td>10</td>
<td>0.749194384</td>
<td>0.729039343</td>
</tr>
<tr>
<td>11</td>
<td>0.658856295</td>
<td>0.79114508</td>
</tr>
<tr>
<td>12</td>
<td>0.596910447</td>
<td>0.787564803</td>
</tr>
<tr>
<td>13</td>
<td>0.442975212</td>
<td><strong>0.813167971</strong></td>
</tr>
<tr>
<td>14</td>
<td>0.324698317</td>
<td>0.780085796</td>
</tr>
<tr>
<td>15</td>
<td>0.284735903</td>
<td>0.809654149</td>
</tr>
<tr>
<td>16</td>
<td>0.254100564</td>
<td>0.797075507</td>
</tr>
<tr>
<td>17</td>
<td>0.241280136</td>
<td>0.784543368</td>
</tr>
<tr>
<td>18</td>
<td>0.220323266</td>
<td>0.695965036</td>
</tr>
<tr>
<td>19</td>
<td>0.219164735</td>
<td>0.585838591</td>
</tr>
<tr>
<td>20</td>
<td>0.205144493</td>
<td>0.609530562</td>
</tr>
</tbody>
</table>

6.5.4. Performance comparison of the proposed solution with community detection algorithms

Topological cluster quality measure, modularity and the ground truth significance is also compared for the clusters derived using current approach with the benchmark community detection algorithms like Fast greedy algorithm (Newman, M. 2004) and Walktrap algorithm (Pons et al. 2005) and Leading eigenvector approach (M.E.J. Newman 2006). A good cluster should have larger number of internal edges and a smaller number of intercluster edges. A modularity index represents values between 0 and 1, where 1 signifies the strength of communities. The proposed approach outperforms the benchmark community detection algorithms in terms of modularity and biological validity measure. A high value of modularity, **0.5472** is observed using
proposed approach with \((\alpha = 2)\). Table 6.5 represents the performance comparison of all the algorithms.

Table 6.5: Performance comparison of ASD genes using standard community detection algorithms and spectral clustering set using Modularity and Rand Index

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modularity</th>
<th>Rand Index</th>
<th>Community Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Algorithm ((\alpha = 2))</td>
<td>0.5471566</td>
<td>0.8659352</td>
<td>104, 129, 105, 39, 38, 61, 65, 64, 49</td>
</tr>
<tr>
<td>Fast greedy Algorithm</td>
<td>0.06764758</td>
<td>0.5005596</td>
<td>282,366,6</td>
</tr>
<tr>
<td>Walktrap Algorithm</td>
<td>0.06756441</td>
<td>0.6444966</td>
<td>282,153,219</td>
</tr>
<tr>
<td>Leading eigenvector Algorithm</td>
<td>0.05948131</td>
<td>0.4884818</td>
<td>376,278</td>
</tr>
</tbody>
</table>

6.5.5. Performance comparison of Individual Vs Aggregated knowledge source when \(\alpha = 2\)

To check whether the integrative approaches that combine several types of data are being perceived as an approach to attain novel insights into gene relationships and forming functionally similar genes in each cluster, the ground truth check is performed for individual and aggregated gene similarity network using proposed approach with the optimum threshold setting, \(\alpha = 2\) for edge pruning. The compliance check for biological relevance using Rand Index in Figure 6.6 shows that the clusters formed using aggregated knowledge is performing well than individual knowledge sources.
6.5.6. Functional evaluation of each cluster using Homogeneity Measure

To check the functional relevance of each cluster, the genes connected within the clusters are tested to determine whether they take part in the same molecular function, biological process and cellular component by computing functional homogeneity index. The optimum clusters with cluster size = 9 with threshold setting $\alpha = 2$ is taken for analysis. Figure 6.7 shows the homogeneity measures obtained for each cluster.
Most of the clusters show an average biological process of 0.80, molecular function of 0.65 and cellular component of 0.40 suggesting that 75-85% of the genes belonging to each cluster participates in the same biological process, 60-70% of the genes belong to each cluster participates in the same molecular function and 30-45% of the genes belong to each cluster participates in same cellular function. For complex disorders, it’s a not a single cellular/biological process but rather a set of processes are relevant [Fernández et al. 2009; Berger et al. 2009]. Hence disease treatments and drug improvements considering the diversity of cellular / biological processes could be of more advantage for complex disorders.

6.5.7. Prevalence of co-occurring conditions in each cluster

Since comorbid conditions have a significant impact on the functioning of people with ASD (Kohane et al. 2012), the literature reviews are explored in terms of comorbid conditions with ASD and evaluated the percentage of co-occurring conditions associated with the candidate genes in each clusters. Table 6.6 details the list of comorbid conditions that are associated with ASD based on literature reviews. The number of comorbidities in ASD is also said to be related to the disease severity; hence
the comorbidity percentage in each cluster is evaluated to classify the clusters based on disease severity. Figure 6.8 represents the percentage of diseases associated with the genes in each clusters. Clusters, C4 and C1 reported highest occurrence of co-occurring conditions leading to increased disease severity, suggesting that subjects with severe autistic cases can have strong associations of genes grouped in these clusters.

Table 6.6: Conditions comorbid to autism spectrum disorders (ASD)

<table>
<thead>
<tr>
<th>Comorbid conditions</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Retardation/Mental Disorders</td>
<td>(Kar et al. 1997)</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>(Tuchman et al. 2002 ; Tuchman et al. 2011)</td>
</tr>
<tr>
<td>Depressive Disorder</td>
<td>(Magnuson et al. 2011 ; Matson et al. 2014)</td>
</tr>
<tr>
<td>Developmental Disabilities</td>
<td>(Levy et al. 2010 ; Matson et al. 2014)</td>
</tr>
<tr>
<td>Intellectual Disability(ID)</td>
<td>(Tuchman et al. 2011)</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>(Hommer et al. 2015)</td>
</tr>
<tr>
<td>Psychology</td>
<td>(Levy et al. 2010; Close et al. 2012; Höglund et al. 2013)</td>
</tr>
<tr>
<td>Seizures</td>
<td>(Volkmar et al. 1990)</td>
</tr>
<tr>
<td>ADHD</td>
<td>(Leitner and Y. 2007)</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>(Joshi et al. 2013)</td>
</tr>
<tr>
<td>Nervous System Diseases</td>
<td>(Goines et al. 2010)</td>
</tr>
<tr>
<td>Autoimmune Diseases</td>
<td>(Enstrom et al. 2009)</td>
</tr>
<tr>
<td>Anxiety Disorders</td>
<td>(van Steensel et al. 2013)</td>
</tr>
<tr>
<td>Memory Disorders</td>
<td>(Williams et al. 2006)</td>
</tr>
<tr>
<td>Mood Disorders</td>
<td>(Matson et al. 2014)</td>
</tr>
<tr>
<td>Behavior Disorders</td>
<td>(Höglund et al. 2013)</td>
</tr>
<tr>
<td>Personality Disorders</td>
<td>(Lugnegård et al. 2012)</td>
</tr>
<tr>
<td>Obsessive Compulsive Disorder</td>
<td>(Jacob et al. 2009 ; Ruzzano et al. 2015)</td>
</tr>
<tr>
<td>Down Syndrome</td>
<td>(DiGuiseppi et al. 2010)</td>
</tr>
<tr>
<td>Fragile X Syndrome</td>
<td>Farzin et al. 2014</td>
</tr>
</tbody>
</table>
6.6. Conclusion

In this chapter, a clustering approach based on spectral graph theory is proposed by fusing multiple information sources, selecting significant edges and identifying optimum eigenvectors with a specific goal to distinguish consistent gene subsets. The preprocessing steps are modified and parameters are tuned to achieve the optimum performance. The integrative approaches that combine several types of data is proved to be an efficient approach to attain novel insights into gene relationships and forming biologically significant clusters and deriving the symptom severity of genes in each
cluster. The outcomes show that diverse knowledge sources contain specific yet complementary data, and efficient data aggregation procedures can be retained to utilize such complementary data in order to extract groups of highly associated gene sets.

The experimental result shows that the clusters derived using the proposed strategy presents high cohesion within the group than those created by standard k-means algorithm. Moreover the pair wise gene associations help in uncovering global features through eigenvector decomposition. By incorporating the pathway information in the algorithm, the spectral clustering process is additionally guided towards more biologically meaningful partitions. Edge pruning approach based on statistical significance thresholding also shows a significant influence on the formation of sensible and biologically interpretable groups. It is also demonstrated that the clusters generated are biologically relevant in their enrichment in GO terms and cluster size shows a more characteristic appropriation than that of standard clustering solutions.

This novel approach gives more optimized and accurate results when compared with benchmark community detection algorithms like Fast greedy, Walktrap and Leading eigenvector algorithm. Moreover this approach reveals potential benefits in biomedical translational research by understanding and demonstrating disease comorbidities in each cluster that leads to severity of the disease. The approach may also be suitable for categorizing highly associated genes with functionally relevant characteristics in other complex disorders with a strong genetic component.