Chapter 6

Subspace Module Extraction and Analysis-An application to Parkinson's Disease

6.1 Introduction

A living organism's body consists of multiple cells. These cells comprises of number of genes within themselves. However, not all genes are turned on or express themselves at the same time. The duration during which the genes are turned on can be referred to as the expression point of that gene. This systematic on-off mechanism in genes makes the cells behave differently in different organs of the body. Gene regulation is the process which precisely captures the on-off activity of genes within the organism. Recording the expression level of genes take place mostly during the transcription phase, where in information from DNA passes into the mRNA. The advent of microarray technology has enabled parallel investigation of genes depending on their expression values. The expression level of genes can be used in analyzing their roles under different conditions. A group of genes work in coordination to accomplish different functions in the living body. Such set of genes are said to be coexpressed and are represented by coexpression networks. Analyzing such a network unfolds the regulation mystery of genes, prioritization basis for disease genes and can also be associated with the annotation related to genes. However, these networks are effective only while identifying the group of genes active under similar conditions. It does not provide any clue about their behavior at different conditions. A gene might undergo mutation causing other genes to behave abnormally. This dysfunctionality can be referred to as disease condition. Using a coexpression network study, it is very difficult to regulate the action of such genes. In order to analyze the varying nature of genes, differential coexpression analysis has come to rescue. This kind of study can be used in determining the participating genes under different conditions or under different stages. Realizing the dynamicity of genes at different conditions can be extended towards determining their role in the phenotypic changes occurring during the different conditions.

Abnormality in the behavior of genes leads to the formation of malfunctioned proteins. Such proteins then starts behaving peculiarly and disrupts the normal functioning of the body leading to certain diseases. Diseases can be broadly classified into groups depending on the body parts they affect. One such class of disease is the neurodegenerative disease which perturbs the brain and spinal cord in a person. The term neurodegeneration is actually made up of two words-neuro meaning nerve cells and degeneration meaning gradual damage. It therefore considers all disorders affecting the nerve cells under its umbrella. However, research has focused only on few notable diseases such as Parkinsons Disease, Alzheimers Disease and Huntington disease. These three diseases manifest themselves with different phenotypic traits but at the cellular level they all lead to deteriorating cognitive abilities such as loss of memory, inability to make decisions in a human etc. Statistical data reveals the number of people affected by this class of disease. Alzheimers Disease is known to affect nearly 4 million people of age more than 65 years. In the United States, it is known to affect nearly 4-6 people out of every 100,000 people. Statistics revealing the occurrence of Parkinsons Disease varies among individuals. On an average, it is known to affect nearly 7-19 people out of every 100,000. The cause of such diseases is still debatable. Only 5 % cases are known to be associated with genetic mutations, the rest of the cases are still under study. Reserchers have suggested the accumulation of toxic materials as one of the causes leading to the death of neurons, thereby disrupting the normal function in the brain. The prevalence of such diseases is on rise, yet there is no effective treatment available till date. The available medicines can only alleviate the symptoms and help the patients in leading a better lifestyle. For instance, memantine and donepezil administered to an Alzheimers patient can only reduce the rate of progression of dementia, Levodopa when administered to a patient having Parkinsons disease would lead to an increase in the dopamine level in the brain thereby providing temporary relief. [1].

Parkinsons Disease is one of the neurodegenerative disorder affecting a number of day-to-day activities in the person. Apart from memory loss and decisive

[1] https://www.news-medical.net/health/What-is-Neurodegeneration.aspx

inability, he can also develop difficulty in swallowing food, might show sleep related disorders, bladder problems or blood pressure changes. It is known to affect the overall well-being of a person [2]. Therefore, this disorder calls for a thorough understanding of its symptoms and ways to overcome them . In this work, I have made one such effort to discover new biomarkers which may be associated with the disease. These biomarkers have been established from literature sources and therefore can be used by the drug designers to develop drugs targeting them.

6.2 Motivation

In co-expression analysis, it is important to consider the heterogeneity of the samples. Tissue-specific or condition-specific co-expression modules may not be detectable in a co-expression network constructed from multiple tissues or conditions because the correlation signal of the tissue/condition-specific modules is diluted by a lack of correlation in other tissues/conditions. However, limiting co-expression analysis to a specific tissue or condition also reduces sample size, thereby also decreasing the statistical power to detect shared co-expression modules. Therefore, methods that do not distinguish between tissues or conditions should be used for identification of common co-expression modules, while differential co-expression comparing different conditions or tissues will be better for identifying modules unique to a specific condition or tissue.

6.3 Related Works

Gene coexpression analysis is widely gaining popularity in the research community. A number of studies have been proposed in the literature to perform anlaysis of gene coexpression analysis and modular structure of gene networks. For example, Tulika et al. [60] used the concept of border genes to identify interesting genes on Alzheimer's Disease. Their method relied on two parameters to point out the novel genes asociated with the disease. Medina and Pilav [94] used Weighted Gene Co-Expression Network Analysis (WGCNA) to identify modules. WGCNA works by constructing the network by considering a Pearson correlation similarity matrix, which is then transformed into a 0-1 matrix using a threshold. The next step involves calculation of Topological overlap measure, TOM [117], which then pro-

[2] https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055

duces a disimilarity TOM. Hierarchical clustering is performed on this dissimilarity TOM to produce modules. Medina and Pilav effectively used this method on Type 1 Diabetes caused by the destruction of pancreatic cells, which are responsible for producing insulin. Yang et al.[163] use WGCNA to construct network and identify modules. These modules are then used to identify various properties of biomarkers associated with cancer. Most methods rely on the choice of a coexpression similarity measure such as Pearson, Spearman or a similar measure and the appropriate use of a threshold. Kumar et al. [73] show that statistical biases introduced by each of these measures result in different networks for the same data. It has been noted that the Pearson measure often results in low performance when identifying coexpressed genes, especially, in the presence of noise. Moreover, coexpression networks are used to infer only a part of the information, i.e., which genes actively participate during any given time, but do not provide much information about their regulation. An extension of this is the differential coexpression analysis [146]. This approach deals with genes having different coexpression partners under different conditions. For examples, the approach explores how a gene changes its function in a healthy person to give rise to a certain disease or how it is expressed in different tissues. The role of such genes can be elaborately studied using module analysis and module correspondence among the stages of the disease. Individual gene analysis is a time-consuming process and often is inadequate to derive unambiguous conclusions about their behavior. Therefore, module analysis among the stages of the disease can serve both the purposes. Some works have been carried out in literature which are based on this logic. Ray and Maulik[118], Ray et al. [119], Hossain et al. [50] have studied the behavior of genes during the progression of HIV-1 disease from acute to chronic stage. Deshpande et al. [26] analyzed the gene expression network using module analysis in the progression of artherosclerosis. Work based on identifying modules from gene coexpression network and analyzing the pattern of genes is reported by He et al. [45]. They analyzed the progression of chronic hepatitis B and C to hepatocellular carcinoma. Ahmed et al. [1] study a variation of gene coexpression network suggesting that two genes may be connected over multiple sets of samples among the given dataset. This leads to the formation of a multi-edge network, where the strength of each edge is determined by the variations among the sample pairs it is made up of. They proposed a measure called TSOM (Topological Subspace Overlap Matrix, an extension of TOM) over multi-edge networks. They extended their TSOM network to find groups of coexpressed modules and validated them. This provided a framework to analyze the effectiveness of TSOM over some disease datasets and studied their progression stages.

6.4 Proposed Method

In order to analyze the differential behavior of genes across the stages, the approach used in my method is based on the multi-edge concept of network construction. The overall process involves a series of steps which are discussed in detail in the following subsections.

6.4.1 Preprocessing

The raw gene expression data is first normalized using the log2 transformation. In order to determine a subset of active genes, we used the variance parameter as suggested by [133]. The variance highlights the most important genes present in the sample. The dataset comprises a number of samples for both control and diseased stage, represented in the form of a matrix called GM having a number of rows and b number of columns. The rows of the matrix corresponds to genes and the columns correspond to the number of conditions. A discretization technique suggested in [1] is then used on the GM matrix, which results in the transformation of $GM = a \times b$ matrix into $D_m = a \times (b \times (b-1)/2)$ matrix. The calculation of this new matrix, D_m is based on an adaptive discretization strategy that is used to convert the $a \times b$ matrix to a $a \times (b \times (b-1)/2)$ matrix, where a represents the number of genes and b represents the number of conditions. Since, we are interested in finding multiple edges between genes, we have used the same line of work as discussed in [1]. For each pair of conditions (p, q) for gene g_k in both the stages, the *arctan* value is computed for the difference in the expression values obtained from the original gene expression matrix, GM for the two conditions. Mathematically, it is represented as

$$\omega = arctan(GM(g_k, p) - GM(g_k, q)). \tag{6.1}$$

The arctan values for each sample pair is then discretized using a threshold value η, which is calculated based on the standard deviation of the expression matrix. The standard deviation is commonly used to understand the variability in data [3]. The arctan values closer to η are given the same discretized value. In this

[3] http://www.dummies.com/education/math/statistics/why-standard-deviation-is-an-important-statistic/

way, we get a new discretized matrix, D_m of order $a \times (b \times (b-1)/2)$.

6.4.2 Multi-edge network construction

The concept of multi-edge networks has been borrowed from [1], where authors are of the opinion that two genes may be correlated, considering different sample subsets. Using this concept, the entry for each pair of genes in D_m is analyzed to obtain sample subsets. These sample subsets represent those conditions under which the D_m matrix shows the same value for two genes. For example, assume genes g_a and g_b have the value 1 for sample pairs d_{sp1} and d_{sp3}, and a value 2 for samples d_{sp4}, d_{sp6} and d_{sp7}. Then g_a and g_b are connected via two edges: one edge corresponds to sample subset $\{d_{sp1}, d_{sp3}\}$ and other edge to sample subset $\{d_{sp4}, d_{sp6}, d_{sp7}\}$.

Example 1: Suppose we consider two genes g_1 and g_2 with expression values for seven samples as given in Table 6.1. Using Pearson correlation coefficient, we see that correlation value between g_1 and g_2 is 0.422. If we use a correlation threshold value 0.5 to consider the pair (g_1, g_2) as connected, we see that these two genes do not satisfy this requirement, and hence they are not connected (low correlation is shown with dotted line in Figure 6.1).

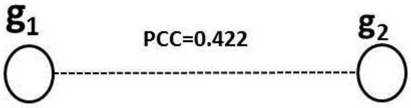


Figure 6.1. Genes g_1 and g_2 are shown connected with dotted line since Pearson correlation value between them is less than 0.5

Table 6.1 Gene expression value for dummy genes g_1 and g_2 for seven samples

Gene name	S1	S2	S3	S4	S5	S6	S7
g_1	0.33	0.29	0.129	0.72	0.11	0.03	0.86
g_2	0.45	0.14	0.34	0.56	0.56	0.5	0.59

To determine whether g_1 and g_2 are connected or not, we find the absolute difference between the expression values of the genes at each condition. If the difference between the values is less than some threshold, say, $\Gamma = 0.9$ (as used in [1]), we can use the same label for the two genes. This is shown in Table 6.2.

Table 6.2 Labels for dummy genes g_1 and g_2 for seven samples using our edge finding approach

Gene name	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
g_1	c1	c1	c2	c1	c3	c3	c2
g_2	c1	c1	c2	c1	c3	c3	c2

In Table 6.2, we see that g_1 and g_2 are connected via three sets of samples, viz., $\{S1, S2, S4\}$, $\{S3, S7\}$ and $\{S5, S6\}$ (Figure 6.2). Using Pearson correlation, we see that the subsets of samples comprising the three edges have a value greater than 0.5, hence all the three edges connect genes g_1 and g_2. The same process is repeated for each pair of genes for all pairs of conditions giving rise to a multi-edge network or graph.

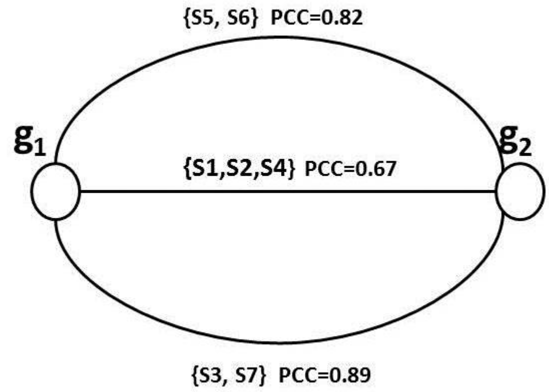


Figure 6.2. Genes g_1 and g_2 are shown to be connected via three subset of edges.

6.4.3 Calculating the topological subspace overlap matrix

It is well established that some information may be lost in a biological network due to the presence of spurious interactions or due to incompleteness of data. To alleviate this issue, Ravasz [117] proposed the concept of a topological overlap matrix incorporating m^{th}- order neighbors (in addition to direct neighbors) between any pair of nodes. Studies have also shown that two proteins with a higher topological overlap are more likely to belong to the same functional group in a PPI network [166]. However, the use of topological overlap matrix (TOM) has been restricted to single edge networks, whether weighted or unweighted. In this work, we analyze the gene network from multi-edge perspective, and hence we use a measure called TSOM as proposed in [1]. We calculate the Topological Subspace Overlap Metric,

T_{SbOM} for a pair of nodes a and b in a network with n nodes as follows.

$$T_{SbOM_{ab}} = 0.5 \times \alpha A + 0.5 \times \beta B. \tag{6.2}$$

where αA and βB represent the similarity between neighbors and connectivity with direct connection respectively and are given by Equations [6.3] and [6.4], respectively.

$$\alpha A = \frac{\sum_{i=1}^{V} |E^p_{g_a g_i} \cap E^q_{g_i g_b}|}{\sum_{i=1}^{V} min(|E^p_{g_a g_i}|, |E^q_{g_i g_b}|)} \tag{6.3}$$

where $E^p_{g_a g_i}$ represents the edge between genes g_a and g_i; p and q represent the sets of samples for which the Pearson correlation coefficients between (g_a, g_i) and (g_i, g_b) are maximum; and V_g is the total number of genes in GM.

$$\beta B = \frac{|E^r_{g_a g_b}|}{sm} \tag{6.4}$$

where, $1 \leq r \leq E_{g_a g_b}$, and $E^r_{g_a g_b}$ is the edge between g_a and g_b with as maximum number of sample subsets and sm is the total number of samples in the original matrix, GM. A factor of 0.5 is assigned to both the terms so that equal emphasis is given to edges directly connecting the two nodes g_a and g_b, and also the common edges between both g_a and g_b. This calculation gives rise to a subspace overlap matrix, T_{SbOM} having the same order as GM, i.e., $V \times E$, but with more interesting values indicating connectivity between gene pairs.

6.4.4 Network module extraction

After the subspace overlap matrix is obtained in the previous step, we convert it into a 0-1 matrix, Adj using a threshold value.

$$Adj_{g_a, g_b} = \begin{cases} 1 & \text{if } T_{SbOM_{g_a, g_b} > n_{th}} \\ 0 & \text{otherwise} \end{cases}$$

The choice of n_{th} is made wisely using the requirement that two nodes which are connected by more than 50% connectivity has to be taken to be present in the network. This is also supported by many other sources [133].

A network module in general is a set of nodes with high topological and neighborhood similarity. To describe the module formation process, the following definitions are required.

Definition 30 (Degree of node). *The degree of a node g_v is the total number of neighbors of g_v in the network. In other words,*

$$d(g_v) = |Adj_(g_v, g_k) = 1| \; \forall g_k \in V \text{ and } k = \{1, 2, ...\}. \tag{6.5}$$

Definition 31 (Nodedistance). *The distance between two node, $Nodedistance(g_v, g_k)$ comprises of shortest path from each node, to every other node, $g_j, \forall g_j = \{V - \{g_v, g_k\}\}$*

$$Nodedistance(g_v, g_k) = \sum_{g_j} |(Adj_(g_v, g_j) == 1)\&(Adj_(g_k, g_j) == 1)| \tag{6.6}$$

Definition 32 (Constrained neighbor). *A node g_k is a constrained neighbor of node g_v only if (i) $Nodedistance(g_v, g_k) >= 1$ and (ii) $degree(g_k) \geq 2$.*

Definition 33 (Constrained node score). *The constrained node score of a node g_v is the ratio of the number of constrained neighbors of g_v in the network to the degree of g_v. Mathematically, it is represented as*

$$CnS(g_v) = \frac{|N_{CN}(g_v)|}{degree(g_v)} \tag{6.7}$$

where $N_{CN}(g_v)$ is the constrained neighbor set of g_v.

Definition 34 (Simpson Index). *The Simpson Index [10] for a pair of nodes g_v and g_k is given by the ratio of the common neigbors of g_v and g_k to the minimum number of neigbors of g_v and g_k. Mathematically, it is given as*

$$SI(g_v, g_k) = \frac{N(g_v) \cap N(g_k)}{min(|N(g_v)|, |N(g_k)|)} \tag{6.8}$$

where $N(g_v)$ and $N(g_k)$ represent the sets of common neighbors of g_v and g_k.

Definition 35 (Seed node). *A node g_v is chosen as a seed node for module extraction iff $CnS(g_v) \geq \gamma$, where γ is a user defined threshold.*

Definition 36 (Network module). *A group of nodes $\{g_v, g_{v_1}, g_{v_2}, ...g_{v_k}\}$ with seed node g_v is defined as a network module, m_i iff*

(i) $\forall g_{v_i} \in m_i$, $T_{SbOM}(g_{v_i}, g_v) \geq T_{SbOM}(g_{v_m}, g_v)$, where $g_{v_m} \notin m_i$ and

(ii) $SI(g_{v_i}, g_v) \geq \delta \; \forall g_{v_i} \in m_i$

Definition 37 (Concerned network module). *A module m_i is said to be a concerned network module iff $\exists g_{v_i} \in m_i$ such that $g_{v_i} \in L$, where $L = \{c_{g_1}, c_{g_2}, ...c_{g_k}\}$ is the set of causal genes associated with the disease as given in a database.*

The module extraction process begins by taking the T_{SbOM} and the *Adj* matrix as input along with two user-defined thresholds, γ and δ. γ is used during the seed selection process. A node g_i with $CnS(g_i) \geq \gamma$ is chosen as the seed node as shown in Line 8 of Algorithm 5. The seed is then expanded to form a cluster (Lines 9-20 of Algorithm 5). During the cluster expansion process shown by Line 10, a node g_j with $T_{SbOM_{g_i,g_j}} \geq T_{SbOM_{g_i,g_k}} \forall g_k$ is chosen as a possible candidate for seed pair expansion. The membership of g_j in the *partialCluster* is further strengthened by the Simpson Index measure. If $SI(g_i, g_j) > \delta$, then g_j can be added into the *partialCluster* and the expansion process continues with the next gene with highest T_{SbOM} value (Lines 12-16 of Algorithm 5). The process stops when there is no node g_l with $SI(partialCluster, g_l) > \delta$. The elements in the *partialCluster* are declared modules if $|partialCluster| \geq 3$ (Line 17). A minimum limit of three elements has been set in order to declare a cluster a module, because lower size clusters cannot be used effectively for p-value analysis and for inferring the behavior of unknown genes. This constraint has been suggested in [101], where it was assumed that a cluster of proteins has to be of minimum size three. The new cluster formation process begins with the node having the next highest *CnS* score. This process also ensures the non-exclusive nature of genes in real life. The psueudocode for module extraction method is given in Algorithm 5.

To establish the effectiveness of our method, we present the following propositions.

Proposition 8. *Two genes $g_v, g_j \in m_i$, the i^{th}module, iff both are strongly connected.*

Explanation: A gene g_j can be a member of a module m_i given by our method only when it is strongly connected with the seed gene of m_i, say, g_v. Strongly connected genes can be determined with the help of both subspace overlapping scores given by T_{SbOM} and the Simpson Index score. A node $g_j \in m_i$ iff $T_{SbOM_{g_v,g_j}} \geq T_{SbOM_{g_v,g_k}} \forall g_k \in V$ and $SI(g_v, g_j) > \delta$. To satisfy both the criteria at the same time, the node g_j needs to have maximum number of common neighbors with g_v and also the two nodes should share maximum correlation in terms of expression values. This correlation is considered while calculating the subspace overlapping matrix, T_{SbOM}. Hence, genes g_v and $g_j \in m_i$ are strongly associated. □

Proposition 9. *A module M_{c_i} from the control stage has high correspondence with a module M_{d_j} from the disease stage iff (i) both have large number of common genes and (ii) both share a maximum number of common pathways.*

Explanation: A gene $g_{v_1} \in m_i$, where $m_i = \{g_v, g_{v_1}, g_{v_2}, ...g_{v_k}\}$ is i^{th} module and

```
Input  : $Adj = \{V, E\}$ (Gene gene network); $T_{SbOM} = \{V, E'\}$ (TSOM
         network); $\gamma$ (Constrained neighbor score); $\delta$ (Simson Index
         threshold);
Output: $Modules = \{C_1, C_2, \cdots, C_N\}$

Initialize clusterExpNode $= V$, NodeList $= V$, Modules $= NULL$, count $=1$;
.........Calculate constrained node score for each gene........
foreach $g_v \in V$ do
    $CnS(g_v) = \frac{|N_{CN}(g_v)|}{degree(g_v)}$
end
..........Module expansion..........
while |NodeList| $> 2$ do
    choose $g_a$ from NodeList such that $CnS(g_a) > CnS(g_k) \forall g_k \in NodeList$
      and $CnS(g_a) > \gamma$
    $partialCluster = partialCluster \bigcup g_a$;
    choose $g_y$ such that
      $T_{SbOM}(partialCluster, g_y) > T_{SbOM}(partialCluster, g_x) \forall g_x \in \{V - g_a\}$
    $SI(g_y, partialCluster) = \frac{N(g_y \cap N(g_k)}{min(|N(g_y)|,|N(partialCluster)|)}$;
    while $SI(g_y, partialCluster) > \delta$ do
        $SI(g_y, partialCluster) = \frac{N(g_y \cap N(g_k)}{min(|N(g_y)|,|N(partialCluster)|)}$
          $clusterExpNode = V - partialCluster$;
        $partialCluster = partialCluster \bigcup g_y$;
        choose another $g_y$ such that $T_{SbOM}(partialCluster, g_y) >$
          $T_{SbOM}(partialCluster, g_x) \forall g_x \in clusterExpNode$;
    end
    Mark $partialCluster$ as $C_{count}$ only when $|partialCluster| \geq 3$;
    $Modules = Modules \bigcup C_{count}$;
    $count++$;
    NodeList $= V - g_a$;
end
Return $Modules$ ;
```

Algorithm 5: Algorithm for module extraction from gene gene network

g_v is the seed node iff (i) $\forall g_{v_i} \in m_i$, $T_{SbOM}(g_{v_i}, g_v) \geq T_{SbOM}(g_{v_m}, g_v)$, where $g_{v_m} \notin m_i$ and (ii) $SI(g_{v_i}, g_v) \geq \delta \; \forall g_{v_i} \in m_i$ (as per Definition 36). A gene has to satisfy these criteria in order to become a member of a module. If two modules, M_{c_i} (from the control stage) and M_{d_j} (from the disease stage) share a maximum number of common genes, it indicates that the network structure of the two stages are similar, which may be attributed to the expression patterns of the member genes, with high correlation among them. A pathway gives a sequence of biochemical reactions with a purpose occuring in the body. Therefore, if two genes share common pathways, it either implies that they correspond to the same gene or they are functionally very similar in nature and hence they validate that the expression patterns of these genes are more coherent in nature. Therefore, modules M_{c_i} and M_{d_j} high correspondence [52]. □

Proposition 10. *A non-causal gene g_k is interesting for a given disease, D_i w.r.t. (i) a set of causal genes for D_i and (ii) sets of corresponding concerned modules from control and disease stages if any one of the following conditions is satisfied.*

(i) g_k has strong topological association with causal genes or

(ii) g_k's pathway is related to that of the causal genes' pathways, or

(iii) g_k has been shown to be associated with D_i in the literature.

Explanation: In order to find new and interesting genes for the disease D_i from the topological perspective, we use the members of the concerned network modules (Definition 37). Suppose we consider a module, $m_i = \{g_i, g_j, g_k, ...g_p\}$, where g_i and g_j are already established to be causal genes w.r.t. D_i. We find the topological partners of g_i and g_j within m_i using the STRING tool [31] and find g_k to be interacting with the causal genes. From pathway point of view, gene g_k's association with the disease, D_i can be analyzed in terms of the number of common pathways it shares with either gene g_i or g_j. A gene, g_k can also be said to be associated with disease D_i, if there are ample evidences for its role in the disease established in literature sources. □

A conceptual framework of the proposed module extraction method is given in Figure 6.3.

6.5 Experimental Results

We implemented our method in MATLAB running on an HP Z 800 workstation with two 2.4 GHz Intel(R) Xeon (R) processors and 12 GB RAM, using the Win-

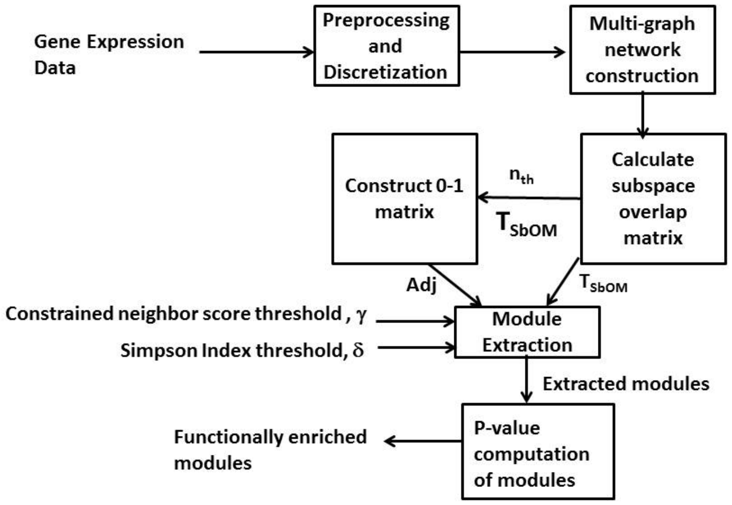

Gene Expression Data
Preprocessing and Discretization
Multi-graph network construction
Calculate subspace overlap matrix
n_{th}
T_{SbOM}
Construct 0-1 matrix
Adj
T_{SbOM}
Constrained neighbor score threshold , γ
Simpson Index threshold, δ
Module Extraction
Extracted modules
P-value computation of modules
Functionally enriched modules

Figure 6.3. Conceptual framework of module extraction method

dows 7 operating system. The main objective of this work is to find functionally enriched modules in both control and diseased stages. The functional enrichment of a gene module is analyzed in terms of its p-value, which gives the probability of random occurrence of functionally enriched genes together. We report the results for gene subset at variance threshold of 0.9. We perform our module extraction technique at different γ and δ values. The obtained cluster (i.e., module) set is then fed into an online tool called Funcassociate [13] which gives the p-value for each of these modules. We report the p-value of these sets in Table 6.3.

6.5.1 Comparison with existing work based on p-value

In order to establish our module extraction method, we have used for comparison our previous work on module extraction [133] on the same dataset. This work uses two parameters, CCT and SST, one being the topological threshold and the other is the functional threshold, respectively. The p-value obtained using this method are given in Table 6.4.

As seen in Table 6.4, p-values of both control and disease stages at CCT=0.5, SST=0.7 are better than at other thresholds. The p-value obtained for the control stage is at par with that of our proposed method, while the p-value in the disease stage is better than the proposed method. One reason for the improvement in p-value in case of the previous work can be attributed to the use of semantic

Table 6.3 p–values of modules obtained using different thresholds in both control and disease stages

ModuleType	p-value			
	$\gamma = 0.03$			
δ	0.2	0.4	0.6	0.8
Control	6.281E-17	5.449E-16	6.251E-16	6.898E-16
Disease	6.983E-8	9.205E-8	1.378E-7	8.262E-8

ModuleType	p-value			
	$\gamma = 0.05$			
δ	0.2	0.4	0.6	0.8
Control	1.825E-17	4.638E-17	4.662E-15	3.154E-13
Disease	4.202E-8	7.397E-8	8.725E-8	4.880E-8

ModuleType	p-value			
	$\gamma = 0.07$			
δ	0.2	0.4	0.6	0.8
Control	4.958E-17	6.648E-17	1.259E-15	6.776E-9
Disease	4.202E-8	4.241E-8	6.988E-8	9.740E-8

Table 6.4 Best p–value reported by modules obtained using different thresholds in both control and disease stage using the method given in [133]

Stages	**CCT=0.5, SST=0.5**	**CCT=0.7, SST=0.5**	**CCT=0.5, SST=0.7**	**CCT=0.7, SST=0.7**
Control	3.17E-13	8.15E-10	4.13E-16	2.16E-14
Disease	4.67E-5	5.53E-8	1.79E-10	5.719E-5

similarity during the module expansion process. We have also used another gene module extraction technique called Module Miner [90] which uses the spanning tree concept to identify modules. Table 6.5 reports the top p-value obtained from our proposed method and two other existing methods both for control stage and disease stage.

Table 6.5 Best p–value reported using my proposed method and other existing works.

Stage	**Proposed work**	**My existing work [133]**	**Module Miner [90]**
Control	4.638E-17	4.13E-16	3.67E-10
Diseased	7.397E-8	1.79E-10	5.45E-6

As can be seen from Table 6.5, the best p-value obtained in both control and disease stage is given by our proposed method. Thus, we can say that our module extraction process proposed here can take us a step further towards the use of multi-edge information when constructing the gene gene network and thereafter in finding functionally similar modules.

6.5.2 Comparison with biclustering techniques

Grouping of genes under subset of samples has been dealt with using biclustering techniques. It is a two-way clustering technique performed considering two dimensions, i.e., genes and conditions simultaneously. Since our proposed method also considers genes under subset of conditions, therefore it has to be compared with biclustering techniques. In this work, we have used two biclustering techniques, namely, Cheng and Church (CC) algorithm and the BiMax algorithm for finding the top three modules based on their p-value. The CC algorithm is based on a greedy approach and uses the mean values of genes at different conditions to find the Mean Squared Residue (MSR), which is then used for evaluating the quality of biclusters The BiMax is based on a divide-and-conquer strategy for declaring maximal bicliques as biclusters. The p-value of top three biclusters obtained from the two algorithms using the BicAT tool [7] at default parameters for both control and disease state are reported in Table 6.6.

As seen from Table 6.6, it is found that biclustering techniques reports functionally coherent modules better than our proposed method both in the control and disease stage. We can fine tune our method to get better quality modules which are at par with those obtained using the biclustering techniques. We now find the analogy between the modules obtained using our method in both the stages so as to extend our analysis.

Table 6.6 Best p–value reported by modules obtained using CC and BiMax biclustering technique in both control and disease stage

Algorithm	Control stage			Disease stage		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**
Cheng and Church (CC)	4.34E-10	1.25E-8	5.23E-7	1.39E-6	6.34E-4	1.95E-4
BiMax	4.14E-23	6.93E-17	5.13 E-11	1.39E-6	6.34E-4	1.95E-4

6.5.3 Module correspondence between the two stages

In order to analyze the behavior of genes responsible for the progression of a disease from a healthy person to a diseased patient, we initially find the module correspondence between the two stages. We chose the modules at $\gamma = 0.05$ and $\delta = 0.2$ due to better p-value w.r.t. other thresholds. We analyze the module correspondence from two aspects: topological and pathways. A visual representation of module correspondence is given in Figure 6.4. We now discuss these aspects in detail.

6.5.3.1 Topological aspect of module correspondence

Topology is concerned with the structure of the gene network. Our idea is somewhat similar to the concept of maximum matching ratio [101]. For every module in the control stage, M_{c_i}, we found its matching ratio with every module in the disease stage, M_{d_j}. This is repeated for all the 22 control modules w.r.t. the 7 diseased modules. The matching ratio between the modules is given by Equation 6.9. The diseased module, d_j with which c_i is found to be maximally matching is considered the corresponding module for c_i.

$$MR(c_i, d_j) = \frac{(|M_{c_i} \cap M_{d_j}|)^2}{|M_{c_i}| \times |M_{d_j}|} \tag{6.9}$$

The control and disease module correspondences at $\gamma = 0.05$ and $\delta = 0.2$ $< C_i, D_j >$ from topological perspective are observed as follows: $< 1, 5 >$, $< 2, 5 >$, $< 3, 7 >$, $< 4, 7 >$, $< 5, 7 >$, $< 6, 7 >$, $< 7, 5 >$,$< 8, 7 >$, $< 9, 7 >$,$< 10, 7 >$,$< 11, 7 >$,$< 12, 7 >$, $< 13, 7 >$, $< 14, 7 >$, $< 15, 7 >$,$< 16, 7 >$, $< 17, 7 >$,$< 18, 7 >$, $< 19, 7 >$, $< 20, 7 >$,$< 21, 7 >$, $< 22, 7 >$. We observe here that all the 22 control modules either correspond to module D_5 or D_7 in the disease stage.

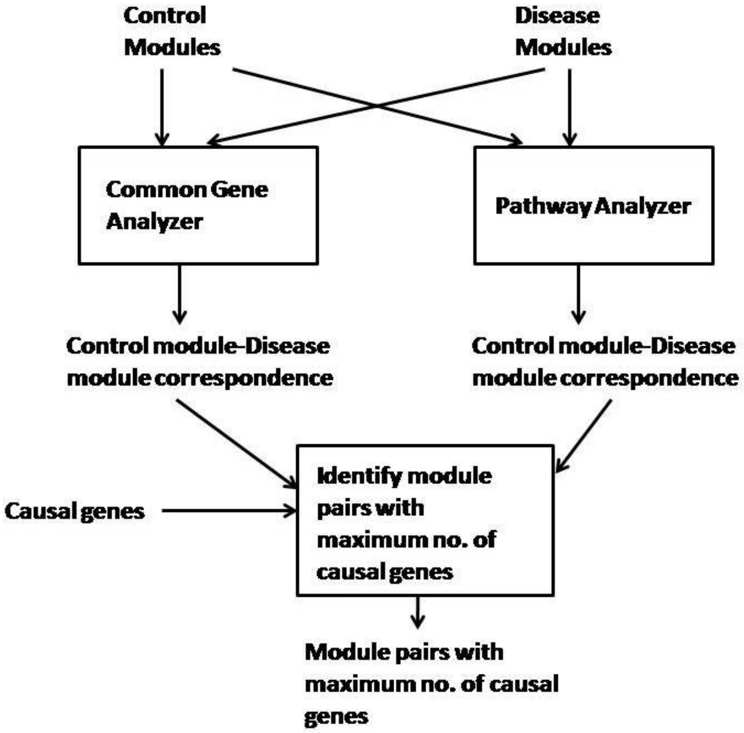

Control Modules
Disease Modules
Common Gene Analyzer
Pathway Analyzer
Control module-Disease module correspondence
Control module-Disease module correspondence
Causal genes
Identify module pairs with maximum no. of causal genes
Module pairs with maximum no. of causal genes

Figure 6.4. Conceptual framework of module correspondence between control and disease stage

6.5.3.2 Module correspondence from pathway point of view

A pathway in a biological domain represents a series of related biochemical reactions occurring in the living body. Molecules involved in promotion or inhibition of any activity in the living body need to be studied carefully. The number of common pathways found among the members of modules in different stages can be an interesting feature. We therefore use the PANTHER tool [97] to find the pathways in each module in the control as well as disease stages. Table 6.7 represents the number of common pathways as given by PANTHER among the modules in both the stages.

In Table 6.7, we observe that disease modules numbered D_7, D_4 and D_5 are the top 3 modules which showed the maximum overlap with all the control modules in terms of pathway. In this case, D_7 emerges as the winner corresponding module for each control module. We further analyze D_7 for better understanding the progression of the disease.

Table 6.7 Number of common pathways for all the 22 control modules w.r.t. the 7 disease modules

Control modules	**Number of common pathways**						
	Diseased modules						
	D1	D2	D3	D4	D5	D6	D7
C1	22	13	8	27	28	24	30
C2	22	13	9	28	29	25	31
C3	22	13	9	28	29	25	31
C4	22	13	7	26	27	24	29
C5	22	13	9	28	29	25	31
C6	22	13	9	28	29	25	31
C7	22	13	9	28	29	25	31
C8	22	13	8	26	27	23	29
C9	20	12	8	26	28	24	30
C10	22	13	9	28	29	25	31
C11	20	12	8	26	28	24	30
C12	20	12	8	26	28	24	30
C13	20	12	8	26	28	24	30
C14	22	13	8	27	28	24	30
C15	20	12	8	26	28	24	30
C16	22	13	8	27	28	24	30
C17	20	12	8	26	28	24b	30
C18	20	12	8	26	28	24	30
C19	22	13	9	28	29	25	31
C20	22	13	8	27	28	24	30
C21	20	12	8	26	28	24	30
C22	22	13	9	28	29	25	31

6.6 Identification of interesting biomarkers

We analyzed initially all the 22 modules for finding interesting features. However, almost in all the modules there were overlaps among the elements in the range (48-52)%. Therefore, in order to get a subset of modules for our analysis, we took the help of GeneCard [120]. It is a repository which stores the list of causal genes for around 5000 diseases. From this repository, we got 54 genes associated with Parkinson's Disease. We used this list of causal genes to identify modules with a high number of disease genes.We observed the inclusion of number of causal genes in each control module, i.e., $< module\ no., no\ of\ causalgenes >$ as follows: $< 1, 5 >$, $< 2, 6 >$,$< 3, 6 >$,$< 4, 2 >$,$< 5, 6 >$, $< 6, 6 >$,$< 7, 6 >$,$< 8, 4 >$, $< 9, 6 >$, $< 10, 6 >$, $< 11, 6 >$, $< 12, 6 >$, $< 13, 6 >$, $< 14, 4 >$, $< 15, 6 >$, $< 16, 3 >$, $< 17, 6 >$, $< 18, 6 >$, $< 19, 5 >$ $< 20, 4 >$, $< 21, 6 >$, $< 22, 5 >$.

We saw that the maximum number of causal genes found among the modules in the control stage is 6 and $\{C_2, C_3, C_5, C_6, C_7, C_9, C_{10}, C_{11}, C_{12}, C_{13}, C_{15}, C_{17}, C_{18}, C_{21}\}$ module set showed the presence of maximum number of causal genes. Hence, further analysis of these modules was carried out. For each of these 14 modules, we performed an extensive study of the causal genes and their interacting partners among the module members. The interacting partners of the causal genes were found using the STRING tool [31], which obtains the partner genes using informations from coexpression data, other experimental data and text mining. In addition to the causal genes listed in GeneCard, we analyzed each interacting partner with the causal gene in terms of pathways. Table 6.8 -6.14 report the causal genes along with their interacting partners and the number of pathways they share in common with the causal gene. Apart from these information, these tables also higlights the genes which can be possible suspected genes for the disease.

We carried out an analysis of the roles of suspected genes which were not yet known to be sharing any pathway with the six causal genes found in the modules from GeneCard. From literature sources, we gathered some information on the association of such genes with the mechanisms involved in the progression of the disease. We now discuss the role of these genes here.

a) ADCY2: Dopamine neurons are rare in the brain and are associated with many day-to-day activities such as movement and learning [38]. The ADCY2 gene, which is an isoform of adenyl cyclase is known to be expressed in the brain. Sources such as [57, 85] have reported mice expressing certain kinds of motor dysfunctionality that affect the striatal dopamine signalling.

b) CNR1: About 40% patients suffering from Parkinson's Disease show a tendency to undergo depression. Experiments have established the role of cannabinoid receptor gene (CNR1) to be associated with depression brought about by the disruption in the monoamine transmission [8].

c) GNB5: Ample evidence is available to describe the role of GNB5 in causing attention deficit hyperactivity disorder, which is one of the symptoms of Parkin-

Table 6.8 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
2	DRD2	ADCY2 (3), CNR1 (1), GNB5 (2), HTR2A (2), ACTL6B (0), SYT1 (0), SLC18A2 (4), SLC6A3 (4), GRIN2A (6)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1*
	SLC6A3	DRD2 (4), TH (5), GRIN1 (3), DDC (4), PTK2B (0), GCH1 (0)	
	TH	HTR2A (0), SLC6A3 (5), GRIN1 (3), DDC (7)	
	DDC	CACNG3 (0), SLC17A7 (0), GCH1 (0), SLC6A3 (4), GRIN2A (4), SLC17A6 (0), MOXD1 (0)	
	SLC18A2	SLC6A3 (5), DRD2 (5), STX1A (2), GCH1 (0), SLC17A7 (1), SLC17A6 (1)	
	GCH1	SYT1 (0), SLC6A3 (0), ADCY2 (0), DDC (1), SLC18A2 (0), NRXN1 (0)	
3	DRD2	ADCY2(3), GNB5 (2), FOS (3), ACTL6B (0), SYT1 (0), SLC18A2 (4), SLC6A3 (4), PTK2B (0), GRIN2A (6)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN*
	SLC6A3	DDC (4), HTR2A (0), DRD2 (4), GRIN2A (3), GRIN1 (1)	
	TH	DDC (7), DDN (0), GRIN1 (3), SLC17A6 (0), HTR2A (0), FOS (2), SLC17A7 (0)	
	DDC	MOXD1 (0), TH (7), SLC17A6 (0), SLC17A7 (0)	
	SLC18A2	SLC6A3 (5), STX1A (2), HTR2A (1), DRD2 (5), SLC17A7 (1), SLC17A6 (1), DDC (5)	
	GCH1	ADCY2 (0), SYT1 (0), SLC6A3 (0), NRXN1 (0)	

Table 6.9 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
5	DRD2	SLC6A3 (4), GRIN2A (6), CNR1 (1), ADCY2 (3), GNB5 (2)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN, DUSP5*
	SLC6A3	CACNG3 (0), STX1A (1), TH (5), SLC18A2 (5), DRD2 (5), HTR2A (0)	
	TH	DDC (7), SLC18A2 (5), SLC6A3 (5), SYT1 (0)	
	DDC	MOXD1 (0), TH (7), GRIN2A (4), SLC18A2 (6), SYT1 (0), GCH1 (0)	
	SLC18A2	SLC17A7 (0), DDC (5), STX1A (2), SYT1 (0), TH (5), NRXN1 (0), SLC6A3 (5)	
	GCH1	SYT1 (0), DUSP5 (0), ADCY2 (0), FOS (0)	
6	DRD2	ADCY2 (3), GNB5, CCKBR (0), FOS (3), SLC18A2 (4), CNR1 (1), CXCR4 (0)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN, CCKBR, CXCR4, KALRN, CACNB2, ACHE, RET, AGTR1, ATP1A3*
	SLC6A3	SLC18A2 (5), TH (5), DDX3Y (0), KALRN (0), HTR2A (0), DRD2 (5), CACNB2 (0)	
	TH	GRIN1 (3), DDC (6), ACHE (0), CACNG3 (0), RET (0), SLC6A3 (5)	
	DDC	MOXD1 (0), SLC17A6 (0), SLC17A7 (0), TH (7), SLC6A3 (4)	
	SLC18A2	PTK2B (0), ACHE (0), FOS (2), SLC17A6 (1), DDC (5), SLC6A3 (5), DRD2 (4), STX1A (2)	
	GCH1	CACNG3 (0), SLC17A7 (0), AGTR1 (0), SYT1 (0), ATP1A3 (0)	

Table 6.10 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
7	DRD2	SLC6A3 (4), SLC18A2 (4), ADCY2 (3), CNR1 (1), FOS (3), CXCL9 (0), DDC (4)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN, CXCL9, DDX3Y, RET, RAB3B, KALRN, ACHE, CACNB2*
	SLC6A3	DDX3Y (0), DRD2 (5), STX1A (1), SLC18A2 (5), TH (5), GRIN2A (4), DDC (4), RET (0), RAB3B (0)	
	TH	SLC6A3 (5), SLC18A2 (5), KALRN (0), STX1A (1), GRIN1 (3), ACHE (0), GCH1 (0), DDC (6), SLC17A6 (0)	
	DDC	CACNB2 (0), GRIN2A (4), GCH1 (0), TH (6), SLC6A3 (4), DRD2 (4), MOXD1 (0), RET (0)	
	SLC18A2	HTR2A (1), SLC17A6 (1), SLC17A7 (1), STX1A (2), DRD2 (4), TH (5), SLC6A3 (5)	
	GCH1	TH (1), ADCY2 (0), DDC (1)	
9	DRD2	PTK2B (0), TH (4), CACNB2 (0), GRIN2A (6), CCK (0), CHRM3 (1), CNR1 (1), CXCR4 (0), CXCL9 (0)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN, CACNB2, CCK, CHRM3, CXCR4, CXCL9, DDX3Y, CHRNB3, ACHE, KALRN, GABRA1*
	SLC6A3	GCH1 (0), SLC17A7 (0), DDX3Y (0), STX1A (1), SLC18A2 (5), PTK2B (0), HTR2A (0), GRIN1 (3)	
	TH	DDN (0), CHRNB3 (0), SLC17A6 (0), ACHE (0), GCH1 (0), SLC18A2 (5), GRIN1 (3), DRD2 (4)	
	DDC	MOXD1 (0), SLC18A2 (5), GCH1 (0), GRN1 (0), GRIN2A (4), CACNB2 (0)	
	SLC18A2	SLC6A3 (5), GCH1 (0), TH (5), DDC (5), CACNB2 (0), PTK2B (0), KALRN (0), SYT1 (0)	
	GCH1	SLC17A6 (0), GABRA1 (0), SLC17A7 (0), SLC6A3 (0), ADCY2 (0), GRIN1 (0), TH (1)	

Table 6.11 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
10	DRD2	PTK2B (0), GRIN2A (7), TH (4), CXCL9 (0), AGTR1 (1), FOS (3), GNB5 (2)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN, CXCL9, AGTR1, DDX3Y, ACHE, RET, CCK, CCKBR, CACNB2, ATP1A3*
	SLC6A3	DDX3Y (0), STX1A (1), SLC18A2 (5), GRIN1 (3), ACHE (0)	
	TH	RET (0), DDC (7), SLC17A6 (0), GRIN2A (4), SLC18A2 (5), DRD2 (4), CCK (0), CCKBR (0), GCH1 (0)	
	DDC	MOXD1 (0), SLC17A6 (0), GRIN1 (3), CACNB2 (0), GRIN2A (4), TH (6), GCH1 (0)	
	SLC18A2	GRIN1 (3), SLC17A7 (1), SLC6A3 (5), STX1A (2), SYT1 (0), PTK2B (0), TH (5), CACNB2 (0)	
	GCH1	DDC (1), TH (1), FOS (0), ATP1A3 (0)	
11	DRD2	PTK2B (0), TH (4), GRIN2A (7), CACNB2 (0), CCK (0), CHRM3 (1), CNR1 (1), CXCR4 (0), CXCL9 (0)	*ADCY2,CNR1,GNB5, HTR2A, ACTL6B, SYT1, GRIN2A, GRIN1, PTK2B, GCH1, CACNG3, SLC17A7, MOXD1, STX1A, NRXN1, FOS, DDN, CACNB2, CCK, CHRM3, CXCR4, CXCL9, ACHE, DDX3Y, DDN, CHRNB3, CACNB2, KALRN, GABRA1*
	SLC6A3	SLC17A7 (0), ACHE (0), GCH1 (0), GRIN1 (3), HTR2A (0), PTK2B (0), SLC18A2 (5), STX1A (0), DDX3Y (0)	
	TH	DDN (0), CHRNB3 (0), SLC17A6 (0), ACHE (0), GCH1 (0), SLC18A2 (0), GRIN1 (3), DRD2 (4), DDC (7)	
	DDC	MOXD1 (0), GRIN2A (4), CACNB2 (0), SLC18A2 (5), GRIN1 (3), GCH1 (0)	
	SLC18A2	SLC6A3 (5), GCH1 (0), TH (5), DDC (5), CACNB2 (0), PTK2B (0), KALRN (0), SYT1 (0)	
	GCH1	SLC17A6 (0), GABRA1 (0), SLC17A7 (0), SLC6A3 (0), SLC18A2 (0), ADCY2 (0), GRIN1 (0), CACNG3 (0), TH (1)	

Table 6.12 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
12	DRD2	SLC6A3 (4), CACNB2 (0), DDC (4), SLC18A2 (4), SLC17A7 (0), ADCY2 (3), STX1A, CXCL9 (0), GRIN1 (6), GRIN2A (7), GABRA2 (1)	*CACNB2, SLC17A7, ADCY2, STX1A, CXCL9, GRIN1, GRIN2A, GABRA2, RET, DDX3Y, SLC17A6,RAB3B, HTR2A, KALRN, DDN, GCH1, ACHE, MOXD1, SYT1, CNR1, RGS4, ATP1A3, ADCY2*
	SLC6A3	RET (0), DDX3Y (0), SLC17A6 (0), TH (5), DDC (4), RAB3B (0), HTR2A (0), SLC18A2 (5), STX1A (1), KALRN (0), DRD2 (5), CACNB2 (0)	
	TH	DDN (0), RET (0), DDC (7), GCH1 (0), SLC18A2 (5), SLC17A7 (0), DRD2 (4), GRIN1 (3), SLC6A3 (5), SLC17A6 (0), ACHE (0)	
	DDC	MOXD1 (0), GCH1 (0), TH (7), SLC18A2 (5), SLC17A7 (0), DRD2 (4), SLC6A3 (4), SLC17A6 (0)	
	SLC18A2	DDC (5), TH (5), SLC17A6 (1), SLC6A3 (5), DRD2 (5), SLC17A7 (1), STX1A (2), SYT1 (0), CNR1 (0), RAB3B (0)	
	GCH1	RGS4 (0), DDC (1), ATP1A3 (0), ADCY2 (0), HTR2A (0)	
13	DRD2	PTK2B (0), SLC18A2 (4), TH (4), CACNB2 (0), GRIN2A (7), CCK (0), CHRM3 (1), CNR1 (1), CXCR4 (0), CXCL9 (0)	*PTK2B, CACNB2, GRIN2A, CCK, CHRM3, CNR1, CXCR4, CXCL9, SLC17A7, ACHE, DD3Y, HTR2A, STX1A, DDN, CHRNB3, GRIN1, CACNG3, MOXD1, CACNB2, ADCY2, GABRA1*
	SLC6A3	SLC17A7 (0), ACHE (0), GCH1 (0), DDX3Y (0), HTR2A (0), PTK2B (0), SLC18A2 (5), STX1A (1)	
	TH	DDN (0), CHRNB3 (0), SLC17A6 (0), ACHE (0), SLC17A7 (0), GCH1 (0), SLC18A2 (5), GRIN1 (3), DRD2 (4), CACNG3 (0), DDC (6)	
	DDC	MOXD1 (0), GRIN2A (4), CACNB2 (0), SLC18A2 (5), GRIN1 (3), GCH1 (0)	
	SLC18A2	SLC6A3 (5), GCH1 (0), TH (5), DDC (5), CACNB2 (0), PTK2B (0), DRD2 (5), STX1A (2)	
	GCH1	SLC17A6 (0), TH (1), CACNG3 (0), DDC (1), GRIN1 (0), ADCY2 (0), SLC18A2 (5), SLC17A7 (0), GABRA1 (0)	

Table 6.13 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
15	DRD2	PTK2B (0), CACNB2 (0), GRIN2A (7), CCK (0), CHRM3 (1), CNR1 (1), CXCR4 (0), CXCL9 (0)	*CACNB2, SLC17A7, ADCY2, STX1A, CXCL9, GRIN1, GRIN2A, GABRA2, RET, DDX3Y, SLC17A6,RAB3B, HTR2A, KALRN, DDN, GCH1, ACHE, MOXD1, SYT1, CNR1, RGS4, ATP1A3, ADCY2, CCK, CHRM3, CXCR4, CHRNB3, CACNG3*
	SLC6A3	SLC17A7 (0), ACHE (0), GCH1 (0), GRIN1 (3), HTR2A (0), PTK2B (0), SLC18A2 (5), STX1A (1), DDX3Y (0)	
	TH	DDN (0), CHRNB3 (0), SLC17A6 (0), ACHE (0), GCH1 (0), SLC18A2 (5), GRIN1 (3), DRD2 (4), CACNG3 (0), DDC (7)	
	DDC	MOXD1 (0), TH (7), GCH1 (0), GRIN1 (3), SLC18A2 (5), CACNB2 (0), GRIN2A (4)	
	SLC18A2	SLC6A3 (5), GCH1 (0), TH (5), DDC (5), CACNB2 (0), PTK2B (0), KALRN (0), SYT1 (2)	
	GCH1	SLC17A6 (0), TH (1), GRIN1 (0), ADCY2 (0), SLC18A2 (5), SLC6A3 (0), SLC17A7 (0)	
17	DRD2	PTK2B (0), TH (4), CACNB2 (0), GRIN2A (7), CCK (0), CHRM3 (1), CNR1 (1), CXCR4 (0), CXCL9 (0)	*CACNB2, SLC17A7, ADCY2, STX1A, CXCL9, GRIN1, GRIN2A, GABRA2, RET, DDX3Y, SLC17A6,RAB3B, HTR2A, KALRN, DDN, GCH1, ACHE, MOXD1, SYT1, CNR1, RGS4, ATP1A3, ADCY2, PTK2B, CCK, CHRM3, CXCR4, CHRNB3, CACNG3*
	SLC6A3	SLC17A7 (0), ACHE (0), GCH1 (0), GRIN1 (3), HTR2A (0), PTK2B (0), SLC18A2 (5), STX1A (1), DDX3Y (0)	
	TH	DDN (0), CHRNB3 (0), SLC17A6 (0), ACHE (0), SLC17A7 (0), GCH1 (0), SLC18A2 (5), GRIN1 (3), DRD2 (4), CACNG3 (0), DDC (6)	
	DDC	MOXD1 (0), TH (7), GRIN2A (4), CACNB2 (0), SLC18A2 (5), GCH1 (0)	
	SLC18A2	SLC6A3 (5), GCH1 (0), TH (5), DDC (5), CACNB2 (0), PTK2B (0), KALRN (0), SYT1 (2)	
	GCH1	SLC17A6 (0), TH (1), CACNG3 (0), GRIN1 (0), ADCY2 (0), SLC18A2 (5), SLC6A3 (0)	

Table 6.14 Causal genes along with their interacting partners among each control module

Module No.	**Causal gene**	**Interacting partners (No. of common pathways with the causal gene)**	**Suspecting genes**
18	DRD2	SLC18A2 (4), GRIN2A (7), STX1A (0), GNB5 (2), CNR1 (1), ADCY2 (3), SLC6A3 (4), CACNB2 (0)	*CACNB2, SLC17A7, ADCY2, STX1A, CXCL9, GRIN1, GRIN2A, GABRA2, RET, DDX3Y, SLC17A6,RAB3B, HTR2A, KALRN, DDN, GCH1, ACHE, MOXD1, SYT1, CNR1, RGS4, ATP1A3, ADCY2, GNB5, RASGRP1*
	SLC6A3	DDC (4), TH (5), RASGRP1 (0), GCH1 (0), CACNB2 (0), DRD2 (5), DDX3Y (0)	
	TH	SLC17A6 (0), SLC17A7 (0), RASGRP1 (0), GCH1 (0), SLC6A3 (5), DDC (6), ACHE (0), DDN (0), RET (0)	
	DDC	TH (7), SLC17A6 (0), RASGRP1 (0), SLC18A2 (5), GCH1 (0), SLC6A3 (4), MOXD1 (0)	
	SLC18A2	SLC17A7 (1), GRIN2A (3), STX1A (2), DRD2 (5), GCH1 (0), DDC (5), RASGRP1 (0), SLC17A6 (1)	
	GCH1	SLC18A2 (5), STX1A (0), HTR2A (0), CACNB2 (0), SLC6A3 (0), DDC (1), TH (1), RASGRP1 (0), ATP1A3 (0)	
21	DRD2	PTK2B (0), TH (4), CACNB2 (0), GRIN2A (7), CCK (0), CHRM3 (1), CNR1 (1), CXCR4 (0), CXCL9 (0)	*CACNB2, SLC17A7, ADCY2, STX1A, CXCL9, GRIN1, GRIN2A, GABRA2, RET, DDX3Y, SLC17A6,RAB3B, HTR2A, KALRN, DDN, GCH1, ACHE, MOXD1, SYT1, CNR1, RGS4, ATP1A3, ADCY2, CCK, CHRM3, CXCR4, CXCL9, CACNG3, GABRA1*
	SLC6A3	SLC17A7 (0), ACHE (0), GCH1 (0), GRIN1 (3), HTR2A (0), PTK2B (0), SLC18A2 (5), STX1A (1), DDX3Y (0)	
	TH	DDN (0), DDC (7), CACNG3 (0), DRD2 (4), GRIN1 (3), SLC18A2 (5), GCH1 (0), SLC17A6 (0), ACHE (0)	
	DDC	TH (7), MOXD1 (0), GRIN2A (4), CACNB2 (0), SLC18A2 (5), GRIN1 (3), GCH1 (0)	
	SLC18A2	SLC6A3 (5), GCH1 (0), TH (5), DDC (5), CACNB2 (0), PTK2B (0), KALRN (0), SYT1 (2)	
	GCH1	SLC17A6 (0), TH (1), CACNG3 (0), DDC (1), GRIN1 (0), ADCY2 (0), SLC18A2 (5), SLC6A3 (0), SLC17A7 (0), GABRA1 (0)	

son's Disease [125]. GB5, a β subunit of the GTP-binding proteins is present in the Central Nervous System. It is known to form complexes which control the transmission activity of neurons, thus affecting the behavioral consequences.

d) HTR2A: Impulsive behavior is one of the consequences seen in a Parkinson's patient undergoing treatment [149]. Serotonin pathways associated with serotonin 2A receptor gene (HTR2A) and dopamine are known to be causing such behavioral changes [69].

e) GRIN2A, GRIN1: People who are heavily addicted to coffee have shown chances of developing Parkinson's Disease via mutation in the glutamate receptor gene (GRIN2A) [42].

f) STX1A: The first symptoms of Parkinson's Disease is attributed to the loss of dopaminergic neurons [65]. A post-mortem experiment conducted in the brain tissue samples of Parkinson's Disease patients shows the association of STX1A with neurotransmitters [115], and hence can be said to be indirectly associated with the disease.

g) SLC17A7, SLC17A6: Gluatamate is one of the most important neurotransmitters associated with the brain's activity. Any mutation in the activity of two proteins- VGLUT1 (SLC17A7) and VGLUT2 (SLC17A6) affect the expression of glutaminergic neurons, which can cause an imbalance to the brain's functioning leading to depression and ultimately resulting in Parkinson's Disease [150].

h) SYT1: Mutations in the Parkin gene may be associated with juvenile Parkinson's Disease or late Parkinson's Disease. It monitors the expression of synaptotagmin1 (SYT1) in the brain, which is indirectly associated with synaptic vesicle release. A deficiency in the expression of the Parkin gene can cause an oxidative stress in dopaminergic neurons, thus affecting people's brain activity, ultimately leading to Parkinson's Disease [4].

i) FOS: Motor abnormality in Parkinson's Disease is caused by the degeneration of nigrostriatal pathway, which is often linked to changes in the pain perception capability of the living being. To see its role, certain experiments were done on the rat model of the disease. Rats with nigrostriatal abnormalities showed varying pain perceptions and hyperalgesic responses when they were injected with formalin drug. This kind of response to the injection lead to reduced expression levels of FOS in the hypothalamus, which is dircetly linked with the sensory stimulus of pain in the brain [142].

[4] grantome.com/grant/NIH/K01-NS047548-01AI

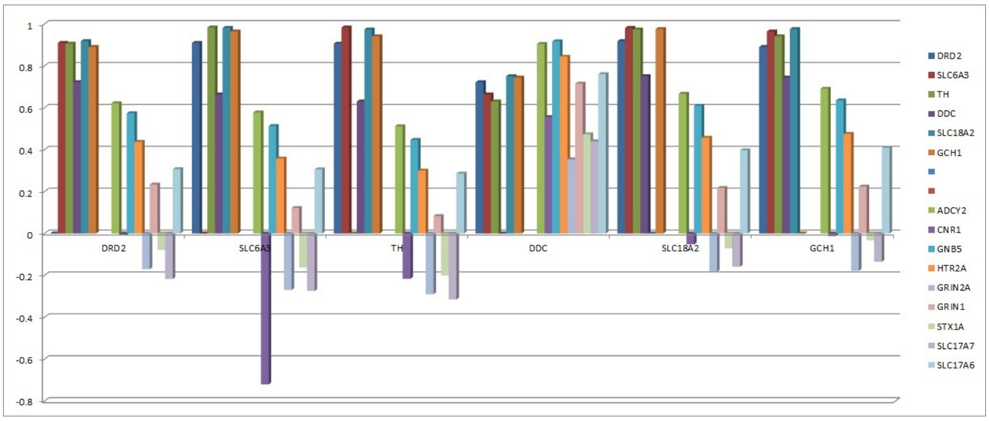


Figure 6.5. Expression patterns of six causal genes and new suspected genes

j) ATP1A3: Changes in the ATP1A3 gene is associated with both dystonia Parkinsonism and hemiplegia of childhood [46].

6.6.1 Expression pattern of new suspected genes

In order to further strengthen our findings, we plotted the expression patterns of suspected genes w.r.t. causal genes. Figure 6.5 depicts the expression level of new suspected genes w.r.t causal genes.

In Figure 6.5, we see that genes such as *ADCY2, GNB5, HTR2A, GRIN2, GRIN1* and *SLC17A6* have shown expression values at par or higher than DDC, SLC18A2 and GCH1, which are already established as causal genes in the database.

6.7 Hub gene analysis in modules based on centrality

In biological networks such as gene-gene or protein networks, the removal of a node may lead to functional changes besides structural changes. Hence, identification of such nodes is important. These nodes are often referred to as hub nodes or essential nodes. Earlier work [174] suggests that a node with a high degree, i.e., more number of inter-connections with other nodes tends to act as one of the central players in the network and removal of such a node would tend to cause structural deformities in the network. However, this measure does not consider the global structure of the network when deciding the significance of each node. To decide upon the essentiality of nodes in the network, a centrality measure can be used to rank nodes based on certain physical characteristics of the network. We used centrality measures [132], viz., betweenness, eigenvector, page rank, closeness and radiality measure to discover the role of nodes in a human protein protein interaction network and finally to pick up the best centrality measure(s) for our purpose. Experimental results show that radiality and pagerank measures are more most suitable while analyzing the importance of each node in such a network. We therefore use these two measures to determine the essential gene (hub gene) among the Parkinson's disease gene network. Tables 6.15 and 6.16 give the hub genes for each module using the radiality and page rank measures in both the stages. The interacting partners of the hub genes along with their association types are also given. Association type can be either direct or indirect. In direct association, a hub gene is found to be at one hop distance with the other gene as given by the network using the STRING tool whereas indirect association implies more than one hop distance in the same network.

In Table 6.15 and 6.16, we see that *RBFOX1, MAFF, MYH11, ACTR2, IGFIR, LY96, S100A9* and *MARCKS* were found as the hub genes among modules obtained in the two stages. However, these genes are not found to be associated with the causal genes in the STRING tool's repository. Taking a closer view of the role of these genes in Parkinson's Disease, we find that four of them (*RBFOX1, MAFF, MYH11, S100A9*) are associated with the disease. An experiment was conducted in vitro on mutations of the neurons in a person suffering from PD showed an increase in the expression level of *RBFOX1* which is associated with RNA processing activities resulting in phenotypic changes of the patient [83]. Another transciptomic study conducted on disease and control olfacatory neurosphere derived cells of a Parkinson's patient revealed that *MAFF* was induced only in PD cells [20]. Dementia is one of the early signs of PD and is diagnosed at least a year before the actual diagnosis of PD [5]. A protein called *S100A9* has been established as a biomarker of dementia progression and hence can be associated with the disease [49]. Another interesting finding is the association of a hub gene *MAFF* with *FOS* and *CNR1* gene in the control stage. The associated genes *FOS* and *CNR1* are among our suspecting genes. The association of these two genes with the disease is already discussed in Section 6.6. Hub gene, *MOXD1* in module 3 of the control stage is directly found to be associated to the causal gene, which makes it all the way more significant w.r.t. the disease. The other hub genes of the control stage do not form part of modules during the disease stage at the given threshold. This may be due to their low subspace overlap value with the seed node.

[5] http://www.alz.org/dementia/parkinsons-disease-symptoms.asp

Table 6.15 Hub genes based on radiality and pagerank measure for each module in control stage along with their associations with causal and suspected genes.

Associations with causal genes				
Measure	**Hubgene**	**Associated genes**		
		Module No	**Gene name(s)**	**Mode of association**
Radiality	MAFF	7	DRD2, SLC6A3, TH, DDC, SLC18A2, GCH1	Indirect
		1	DRD2, SLC6A3, TH, DDC, SLC18A2	Indirect
		14,20	DRD2, SLC6A3, TH, DDC,	Indirect
	MYH11	2, 5	DRD2, SLC6A3, TH, DDC, SLC18A2, GCH1	Indirect
		8	DRD2, SLC6A3, TH, DDC	Indirect
	RBFOX1	3,6,9,10,11,13, 15,17,18, 21	DRD2, SLC6A3, TH, DDC, SLC18A2, GCH1	Indirect
		4	DRD2, SLC6A3	Indirect
		16	DRD2, SLC6A3, TH	Indirect
		19, 22	DRD2, SLC6A3, TH, DDC, SLC18A2	Indirect
Associations with suspected genes				
	MAFF	7	FOS, CNR1	Indirect
		1	FOS, CNR1	Indirect
Associations with causal genes				
Pagerank	MOXD1	2,3,5,6,7,9,10, 11,12,13,15,17, 18,21	DRD2, SLC6A3, TH, DDC, SLC18A2, GCH1	Indirect
		4	DRD2, SLC6A3	Indirect
		8	DRD2, SLC6A3, TH, DDC	Indirect
		14,20	DRD2, SLC6A3, TH, DDC	Indirect
		16	DRD2, SLC6A3, TH	Indirect
		19, 22	DRD2, SLC6A3, TH, DDC, SLC18A2	Indirect
Associations with suspected genes				
		3	DDC	Direct

Table 6.16 Hub genes based on radiality and pagerank measure for each module in disease stage along with their associations with causal and suspected genes.

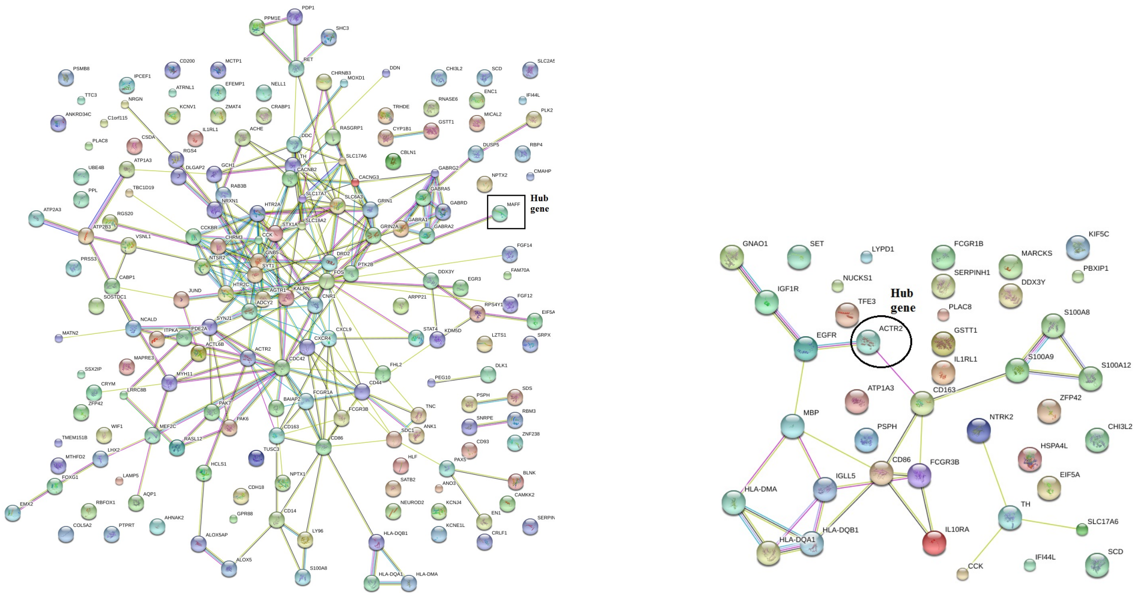
Associations with causal genes				
Measure	**Hubgene**	**Associated genes**		
		Module No	**Gene name(s)**	**Mode of association**
Radiality	ACTR2	1,2,3	SLC6A3	Indirect
	IGF1R	4	SLC6A3	Indirect
	DDX3Y	5	SLC6A3, TH	Indirect
	LY96	6	SLC6A3, DDC, GCH1	Indirect
	ACHE	7	SLC6A3, DDC, GCH1, TH	Indirect
Pagerank	S100A9			
	MARCKS	2	SLC6A3	Indirect
		3	SLC6A3	Indirect
		4	SLC6A3	Indirect
		5	SLC6A3, TH	Indirect
		6	SLC6A3, DDC, GCH1	Indirect
		7	SLC6A3, DDC, GCH1, TH	Indirect

A low subspace overlap value indicates fewer connections of this gene with the rest of the genes in the network. Although, we could not find grounded evidence for the association of a few hub genes with the disease, it can be a good starting point for the biologists to conduct experiments and analyze their roles in the genomic structure of patients suffering from the disease.

Two hub genes, MAFF in modules 1 and 7 of the control stage and ACTR2 for modules 1 and 2 in the disease stage are shown in Figure 6.6(a) and 6.6(b), respectively.

6.8 Discussion

In this chapter, we have explored the properties of genes expressed at different time points under different conditions and extended the discussions to define multi-edge gene-gene networks. We extract modules from the network at different stages. The use of multiple edges highlights the presence of even those edges which are eliminated due to the use of a higher threshold in Pearson correlation coefficient. The work proposed in Chapter 5 lead to information loss due to thresholding in the first step of network construction. In this work, it has been overcome by considering genes under subset of conditions. The obtained gene network is more reliable and informative and hence leads to better module extraction. The extracted modules in both the stages are then used to find consensus modules that show high



(a) Hub gene *MAFF* and its partners in control modules 1& 7.

(b) Hub gene *ACTR2* and its partners in disease modules 1 & 2.

Figure 6.6. Visual representation of control module and disease module along with their hub genes using STRING tool.

correspondence in terms of sharing of common genes and pathways. The concept of consensus modules encourages a detailed analysis of the differentially coexpressed genes across the stages. This analysis helps identify certain new genes such as *ADCY2, GNB5, HTR2A, GRIN2A, GRIN1* and *SLC17A6* which have been found strongly associated with the causal genes known apriori and, hence may also cause a critical disease like Parkinson. However, the use of the concept of multi-edge networks is limited by the ability of the hardware platform. Processing time is directly proportional to the number of genes times the condition set. Hence, we report the results for only one threshold. From the view point of centrality analysis, genes such as *RBFOX1, MAFF, S100A9* are found to be closely associated with the disease. Although the network construction process is slightly time consuming, a parallel implementation on CUDA platform is going on to overcome the issue. The publication associated with this chapter is listed as Publication No. 3 and 7 under the Publication section.