Chapter 5

SNMRS : An Effective Measure for Co-expression Network Analysis

5.1 Introduction

The challenge of identifying modules in a gene interaction network is important for a better understanding of the overall network architecture. In this work, a novel similarity measure called *Scaling-and-Shifting Normalized Mean Residue Similarity (SNMRS)* is developed based on the existing NMRS technique [7]. SNMRS yields correlation values in the range of 0 to +1 corresponding to negative and positive dependency. To study the performance of my measure, internal validation of extracted clusters resulting from different methods is carried out. Based on the performance, hierarchical clustering method is choosen and apply the same using the corresponding dissimilarity (distance) values of SNMRS scores, and utilize a dynamic tree cut method for extracting dense modules. The modules are validated using a literature search, KEGG pathway analysis, and Gene-Ontology (GO) analyses on the genes that make up the modules. Moreover, this measure can handle absolute, shifting, scaling, and shifting-and-scaling correlations and provides better performance than several other measures in terms of cluster-validity indices. Also, SNMRS based module detection method results in interesting biologically relevant patterns from gene microarray and RNA-seq dataset.

The Gene Co-expression Network (GCN) is a gene-interaction network that is frequently used to describe the complex functional organisation of biological

systems at the genome level. GCN is a square matrix the elements of which are derived from a preprocessed dataset. Each element of the square matrix is a co-expression score of a pair of genes greater than a user-defined threshold value. Analysis of GCN is carried out by extracting modules or clusters from a parent square matrix. A module or cluster is a group of co-expressed genes which are tightly connected. A co-expressed or correlated gene pair is similar in terms of its charcteristics or behaviour in most of the experimental conditions or time series. Correlation defines the interdependency between or among gene pairs. It is also reported that co-expressed genes can exhibit any type of correlation patterns such as, absolute, shifting, scaling, and shifting-and-scaling [265] [266].

In the literature, several measures have been proposed for GCN construction. The methods use a gene expression dataset as a primary input and then generate the corresponding co-expression networks using a correlation-based proximity measure. Frequently used correlation measures with linear relationship to construct GCN are: Pearson correlation coefficient (PCC) [84], Spearman rank correlation coefficient [88], Kendall rank correlations [267], Mutual information [88] [268] [89], Normalized mean residue similarity (NMRS) [7], and Negative Correlation aided Normalized Mean Residue Similarity (NCNMRS) [269]. Mahanta et. al [7] develope NMRS as an effective gene similarity measure. Both positive and negative correlation are handled by the NCNMRS correlation measure [269]. Spearman and Kendall can be used as alternatives to PCC but sometimes their performance are found weaker. Mallik et. al developed WeCoMXP which is a weighted connectivity measure to detect gene-modules for multi-omics dataset [270]. It integrates co-methylation, co-expression and protein- protein interactions.

A similarity measure for co-expression analysis should have the ability to detect all types of correlations to find the co-expressed modules/clusters. An effective measure is able to detect all types of correlations without scaling down all the genes to the same range of expression values. In a co-expression network, the absolute values of a co-expression measure are usually used to determine the associations between genes. The absolute values belong to the range 0 to 1 while the PCC range is -1 to +1 and an unsigned weighted network is obtained by transferring the negative values to positive ones. The PCC can detect all types of correlations but it can not address the issues of signed and unsigned network range. The range of NMRS values lie between 0 to +1 but it can detect only shifting patterns [7]. Further, the NCNMRS range can detect the co-expression values from 0.5 to +1 and can detect gene pairs with shifting patterns. It has been observed in [7] and [269] that the developed GCNs are unweighted. But, it is found

that the weighted networks are more robust and biologically more significant [78] than the unweighted GCNs. Following are the contributions from this work.

- An advanced correlation measure called SNMRS.
- Establishment of SNMRS as a metric.
- An approach to construct a weighted signed co-expression network and extraction of biologically significant modules using SNMRS.

5.2 Background

Module detection from co-expression network is a crucial task in computational biology. For module detection, an appropriate connectivity measure is needed. There are several such measures, and the oldest one is Pearsons correlation coefficient (PCC). PCC score between two genes g1: $g1_i = (g1_1, g1_2, ..., g1_m)$ and g2: $g2_i = (g2_1, g2_2, ..., g2_m)$ is defined as follows.

$$PCC = \frac{\sum_{i=1}^{m}(g1_i - \overline{g1}) \times (g2_i - \overline{g2})}{\sqrt{(\sum_{i=1}^{m}(g1_i - \overline{g1})^2 \times \sum_{i=1}^{m}(g2_i - \overline{g2})^2}} \tag{5.1}$$

Mahanta et. al [7] calculate the similarity between genes using NMRS and form a coexpression network using signum function and NMRS threshold. They construct GCN from a microarray gene expression data and extract network modules with the help of Topological Overlap Matrix and using a spanning tree-based method. The constructed coexpression network is unweighted. NMRS score between two genes g1 and g2 is defined as follows.

$$NMRS = 1 - \frac{\sum_{i=1}^{m}\left|g1_i - \overline{g1} - g2_i + \overline{g2}\right|}{2 \times max(\sum_{i=1}^{m}\left|g1_i - \overline{g1}\right|, \sum_{i=1}^{m}\left|g2_i - \overline{g2}\right|)} \tag{5.2}$$

where, $\bar{g1} = \frac{(g1_1, g1_2, ..., g1_m)}{m}, \bar{g2} = \frac{(g2_1, g2_2, ..., g2_m)}{m}$

Ahmed et. al [269] construct an unweighted co-expression network using NCNMRS correlation among genes with a correlation threshold. In practice, there is very little difference between NMRS and NCNMRS [269] both of which can identify correlation values both positive and negative between two gene expression

profiles. NCNMRS score between two genes g1 and g2 is defined as follows.

$$NCNMRS(g1, g2) = \begin{cases} NMRS(g1, g2); & NMRS(g1, g2) \geq 0.5 \\ 1 - NMRS(g1, g2); & NMRS(g1, g2) < 0.5 \end{cases} \tag{5.3}$$

Spearman rank correlation coefficient score between two genes g1 and g2 is defined as follows.

$$Spearman = \frac{\sum_{i=1}^{n}(rank(g1_i) - \overline{rank(g1)}) \times (rank(g2_i) - \overline{rank(g2)})}{(\sum_{i=1}^{n}(rank(g1_i) - \overline{rank(g1)})^2 \times \sum_{i=1}^{n}(rank(g2_i) - \overline{rank(g2)})^2} \tag{5.4}$$

Kendall rank correlation coefficient score between two genes g1 and g2 is defined as follows.

$$tau = \frac{n_c - n_d}{0.5 * n(n-1)} \tag{5.5}$$

Where, n_c: total number of concordant pairs, n_d: total number of discordant pairs and n: size of g1 and g2.

Steuer et al. [89] reports about Mutual Information which can be used as a similarity measure to form a GCN. MI score between two genes G1 and G2 is defined as follows.

$$MI(G1; G2) = \sum_{i,j} p(g1_i, g2_j) log \frac{p(g1_i, g2_j)}{p(g1_i)p(g2_j)} \tag{5.6}$$

$p(g1_i), p(g2_j)$= marginal probabilities of $G1 = g1_i$ and $G2 = g2_j$ for genes G1 and G2, respectively, $p(g1_i, g2_j)$ = joint probability of expression levels related to G1 and G2.

5.2.1 Materials and Method

This section presents an effective similarity measure called *Scaling-and-Shifting Normalized Mean Residue Similarity (SNMRS)*. The SNMRS of a gene $g1$ =

$(g1_1, g1_2, ..., g1_m)$ with respect to gene $g2 = (g2_1, g2_2, ..., g2_m)$ is defined by

$$SNMRS(g1, g2) = 1 - \frac{\sum_{i=1}^{m}|g1_i - \bar{g1} - g2_i + \bar{g2}| - \left|\sum_{i=1}^{m}|g1_i - \bar{g1}| - \sum_{i=1}^{m}|g2_i - \bar{g2}|\right|}{2 * min\left(\sum_{i=1}^{m}|g1_i - \bar{g1}|, \sum_{i=1}^{m}|g2_i - \bar{g2}|\right)} \quad (5.7)$$

where, $\bar{g1} = \frac{(g1_1, g1_2, ..., g1_m)}{m}$, $\bar{g2} = \frac{(g2_1, g2_2, ..., g2_m)}{m}$

5.2.2 Properties of SNMRS

SNMRS satisfies all the properties of a metric. It is established that proposed measure has non-negativity, symmetricity, and triangular inequality properties. The proofs of these properties are reported next.

(a) Non-negativity: To satisfy the non-negativity property, SNMRS of two genes must not be negative or it should be greater than or equal to zero i.e.,

$SNMRS(g1, g2) \geq 0$, where g1 and g2 are two gene profiles.

(b) Symmetricity: To satisfy the symmetricity property, for any two genes g1 and g2, SNMRS(g1,g2) should be equal to $SNMRS(g2, g1), i.e., SNMRS(g1, g2) = SNMRS(g2, g1)$.

(c) Subadditivity or Triangle Inequality: SNMRS satisfies triangular inequality property, for any three genes g1, g2 and g3. Mathematically, $SNMRS(g1, g2) + SNMRS(g2, g3) \geqslant SNMRS(g1, g3)$.

Besides these triangular properties, SNMRS also satisfies the following properties.

i. The score between a pair of gene expression profiles using SNMRS with shifting correlation is 1.

ii. The score between a pair of gene expression profiles using SNMRS with scaling correlation is 1.

iii. The score between a pair of gene expression profiles using SNMRS with shifting-and-scaling correlation is 1.

iv. All the diagonal elements of the correlation matrix are 1.

5.2.2.1 Proof-1: Non-negativity

To satisfy the non-negativity property, SNMRS of two genes should be always greater than or equal to zero, i.e., $SNMRS(g1, g2) \geq 0$, where g1 and g2 are two gene profiles.

Proof: $SNMRS = 1 - \frac{\left(\sum_{i=1}^{m}|g1_i - \bar{g1} - g2_i + \bar{g2}| - \left|\sum_{i=1}^{m}|g1_i - \bar{g1}| - \sum_{i=1}^{m}|g2_i - \bar{g2}|\right|\right)}{2 * min\left(\sum_{i=1}^{m}|g1_i - \bar{g1}|, \sum_{i=1}^{m}|g2_i - \bar{g2}|\right)}$

where, $g1_i = (g1_1, g1_2, g1_3, ..., g1_m)$, $g2_i = (g2_1, g2_2, g2_3, ..., g2_m)$; (1,2,...,m) are the indices of samples/conditions.

Let,

$sum = \sum_{i=1}^{m} |g1_i - \bar{g1} - g2_i + \bar{g2}| = \sum_{i=1}^{m} |(g1_i - \bar{g1}) - (g2_i - \bar{g2})|$,

$diff = |\sum_{i=1}^{m} |g1_i - \bar{g1}| - \sum_{i=1}^{m} |g2_i - \bar{g2}||$

$min = min\left(\sum_{i=1}^{m} |g1_i - \bar{g1}|, \sum_{i=1}^{m} |g2_i - \bar{g2}|\right)$

According to the reverse triangle inequality, $|g1 - g2| \geq ||g1| - |g2||$

$$\Rightarrow \sum_{i=1}^{m} |(g1_i - \bar{g1}) - (g2_i - \bar{g2})| \geq \left|\sum_{i=1}^{m} |g1_i - \bar{g1}| - \sum_{i=1}^{m} |g2_i - \bar{g2}|\right| \tag{5.8}$$

Therefore,

$$\left(\sum_{i=1}^{m} |g1_i - \bar{g1} - g2_i + \bar{g2}| - \left|\sum_{i=1}^{m} |g1_i - \bar{g1}| - \sum_{i=1}^{m} |g2_i - \bar{g2}|\right|\right) \geq 0 \tag{5.9}$$

Hence,

$$\frac{\left(\sum_{i=1}^{m} |g1_i - \bar{g1} - g2_i + \bar{g2}| - |\sum_{i=1}^{m} |g1_i - \bar{g1}| - \sum_{i=1}^{m} |g2_i - \bar{g2}||\right)}{2 * min\left(\sum_{i=1}^{m} |g1_i - \bar{g1}|, \sum_{i=1}^{m} |g2_i - \bar{g2}|\right)} \geq 0 \tag{5.10}$$

Again suppose,

$$\frac{\left(\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|-\left|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}|\right|\right)}{2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right)} > 1$$

$$\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|-\left|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}|\right| > 2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right) \tag{5.11}$$

From equation 5.8, we got two cases: case 1: When sum=diff, then (sum-diff)/(2*min) will be 0.

case 2: When sum>diff and diff=0, then minimum and maximum of

$\sum_{i=1}^{m}|g1_i-\bar{g1}|\, and \sum_{i=1}^{m}|g2_i-\bar{g2}|$ are same i.e.,

If $|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}||=0$

then

$$\frac{\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|}{2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right)} < 1, \tag{5.12}$$

Since difference between two positive absolute quantities can not be greater than the two times of maximum (both are equal so minimum is maximum one) absolute value of either, this contradicts the supposition. Hence, the supposition is false. So, the equation 5.12.

From equation 5.8 and 5.12,

$$0 \leq \frac{\left(\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|-|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}||\right)}{2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right)} < 1 \tag{5.13}$$

Multiplying by -1,

$$0 > -\frac{\left(\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|-|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}||\right)}{2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right)} \geq -1 \tag{5.14}$$

Adding 1 in both sides,

$$1+0 > 1-\frac{\left(\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|-\left|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}|\right|\right)}{2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right)} \geq 1-1$$

$$\Rightarrow 1 > 1-\frac{\left(\sum_{i=1}^{m}|g1_i-\bar{g1}-g2_i+\bar{g2}|-\left|\sum_{i=1}^{m}|g1_i-\bar{g1}|-\sum_{i=1}^{m}|g2_i-\bar{g2}|\right|\right)}{2*min\left(\sum_{i=1}^{m}|g1_i-\bar{g1}|,\sum_{i=1}^{m}|g2_i-\bar{g2}|\right)} \geq 0$$

$\Rightarrow 1 > SNMRS(g1, g2) \geq 0$

$$Hence, 1 > SNMRS(g1, g2) \geq 0 \tag{5.15}$$

Hence, we have proved that SNMRS satisfies the non-negativity property. Moreover, for any two genes, the value of SNMRS always lies between 0 and 1.

5.2.2.2 Proof-2: Symmetricity

To satisfy the symmetricity property, for any two genes g1 and g2, SNMRS(g1,g2) should be equal to $SNMRS(g2, g1), i.e., SNMRS(g1, g2) = SNMRS(g2, g1)$

Proof:

$$SNMRS = 1 - \frac{(\sum_{i=1}^{m} |g1_i - \bar{g1} - g2_i + \bar{g2}| - |\sum_{i=1}^{m} |g1_i - \bar{g1}| - \sum_{i=1}^{m} |g2_i - \bar{g2}||)}{2 * min\left(\sum_{i=1}^{m} |g1_i - \bar{g1}|, \sum_{i=1}^{m} |g2_i - \bar{g2}|\right)} \tag{5.16}$$

$$SNMRS = 1 - \frac{(\sum_{i=1}^{m} |g2_i - \bar{g2} - g1_i + \bar{g1}| - |\sum_{i=1}^{m} |g2_i - \bar{g2}| - \sum_{i=1}^{m} |g1_i - \bar{g1}||)}{2 * min\left(\sum_{i=1}^{m} |g2_i - \bar{g2}|, \sum_{i=1}^{m} |g1_i - \bar{g1}|\right)} \tag{5.17}$$

$\sum_{i=1}^{m} |g1_i - \bar{g1} - g2_i + \bar{g2}| - |\sum_{i=1}^{m} |g1_i - \bar{g1}| - \sum_{i=1}^{m} |g2_i - \bar{g2}||$

$\Rightarrow \sum_{i=1}^{m} |-(-g1_i + \bar{g1} + g2_i - \bar{g2})| - |-(\sum_{i=1}^{m} |g2_i - \bar{g2}| - \sum_{i=1}^{m} |g1_i - \bar{g1}|)|$

According to modulo property,

$|g1 - g2| = |-(g2 - g1)| = |g2 - g1|$ therefore,

$\Rightarrow \sum_{i=1}^{m} |(-g1_i + \bar{g1} + g2_i - \bar{g2})| - |(\sum_{i=1}^{m} |g2_i - \bar{g2}| - \sum_{i=1}^{m} |g1_i - \bar{g1}|)|$

$\Rightarrow \sum_{i=1}^{m} |-g1_i + \bar{g1} + g2_i - \bar{g2}| - |\sum_{i=1}^{m} |g2_i - \bar{g2}| - \sum_{i=1}^{m} |g1_i - \bar{g1}||$

$\Rightarrow \sum_{i=1}^{m} |g2_i - \bar{g2} - g1_i + \bar{g1}| - |\sum_{i=1}^{m} |g2_i - \bar{g2}| - \sum_{i=1}^{m} |g1_i - \bar{g1}||$

$\Rightarrow SNMRS(g2, g1)$

Hence,$SNMRS(g1, g2) = SNMRS(g2, g1)$(Denominator of both equation 5.16 and 5.17, are same) Hence, we have proved that SNMRS satisfies the symmetricity property.

5.2.2.3 Proof-3: Subadditivity or Triangle Inequality

To satisfy triangular inequality property, for any three genes x, y and z, the following condition should hold: $SNMRS(g1, g2) + SNMRS(g2, g3) \geq SNMRS(g1, g3)$.

Proof: From equation 5.15, we have

$$0 \leq SNMRS(g1, g2) < 1 \tag{5.18}$$

$$0 \leq SNMRS(g2, g3) < 1 \tag{5.19}$$

$$0 \leq SNMRS(g1, g3) < 1 \tag{5.20}$$

From equation 5.18 and 5.19,

$0 \leq SNMRS(g1, g2) + SNMRS(g2, g3) < 1 + 1$

$$\Rightarrow 0 \leq SNMRS(g1, g2) + SNMRS(g2, g3) < 2 \tag{5.21}$$

From equation 5.20 and 5.21,

$$SNMRS(g1, g2) + SNMRS(g2, g3) \geq SNMRS(g1, g3) \tag{5.22}$$

Hence, it is proved that SNMRS satisfies the triangle inequality property.

5.2.3 Proposed GCN construction and Module extraction method

Table 5.1: Dataset description

Dataset	Type	Details
D1	Synthetic data	9 genes x 9 conditions
D2	Synthetic data	8 genes X 20 conditions
D3	Synthetic data	8 genes x 10 conditions
Iris data	Real data - Multivariate	150 instances , 4 features
Yeast sporulation data	Real data - Microarray	6118 genes x 7 time points
Esophageal Squamous Cell Carcinoma	Real data - RNA-seq	58000 genes, (14 tumor, 15 normal samples

The workflow mechanism and the algorithm for gene co-expression network construction using proposed measure SNMRS and module extraction are presented in Figure 5-1 and Algorithm 1, respectively. A list of datasets used in proposed method is presented in Table 5.1. In Figure 5-1, time-series microarray and RNA-seq data with both normal (used as control) and disease conditions (test) are used as inputs. Time series data provides gene expressions over time points. The RNA-seq dataset provides genes and samples with normal and tumor samples. After preprocessing of the experimental raw data, normal versus diseased datasets are compared to identify statistically significant DEGs. Independent analysis is performed for these two types of data. The step for identifying DEGs is optional if precalculated DEGs list is used. The measure SNMRS is applied on preprocessed data to construct a weighted positive co-expression network using SNMRS co-expression measure with genes of higher SNMRS score. Each gene is considered as a vertex in the network and an edge exists between a pair of vertices if the SNMRS score of the corresponding genes is more than the user defined threshold value. To extract network modules from the co-expression network, a dissimilarity matrix is computed as (1-SM). Similarity matrix contains gene pairs with a SNMRS value greater than user defined threshold. Based on experimental study it has been observed that with correlation value 0.8 as cutoff, results are found significant. The diagonal of the similarity matrix is identical with a value 1 because of self-similarity between gene pairs and it is maximum. The resulting dissimilarity matrix is used as input for module extraction. In order to extract modules/clusters from the network, hierarchical clustering has been used. An average linkage hierarchical algorithm is applied and found a dendrogram for the effective extraction of biologically significant modules. The dynamic thresholding approach is used and specified the minimum module size as 30 for microarray data and RNA-seq data to identify the gene modules. These modules are assigned with a unique colour and extracted all the detected modules for subsequent downstream analysis. Modules are validated using GO analysis, pathway

analysis and literature search and from these modules, potential gene biomarkers are identified. A set of genes present in the non-preserved modules are recognized as hub genes using intra-modular connectivity analysis [78] and these genes are further studied to know their association in the progression of ESCC.

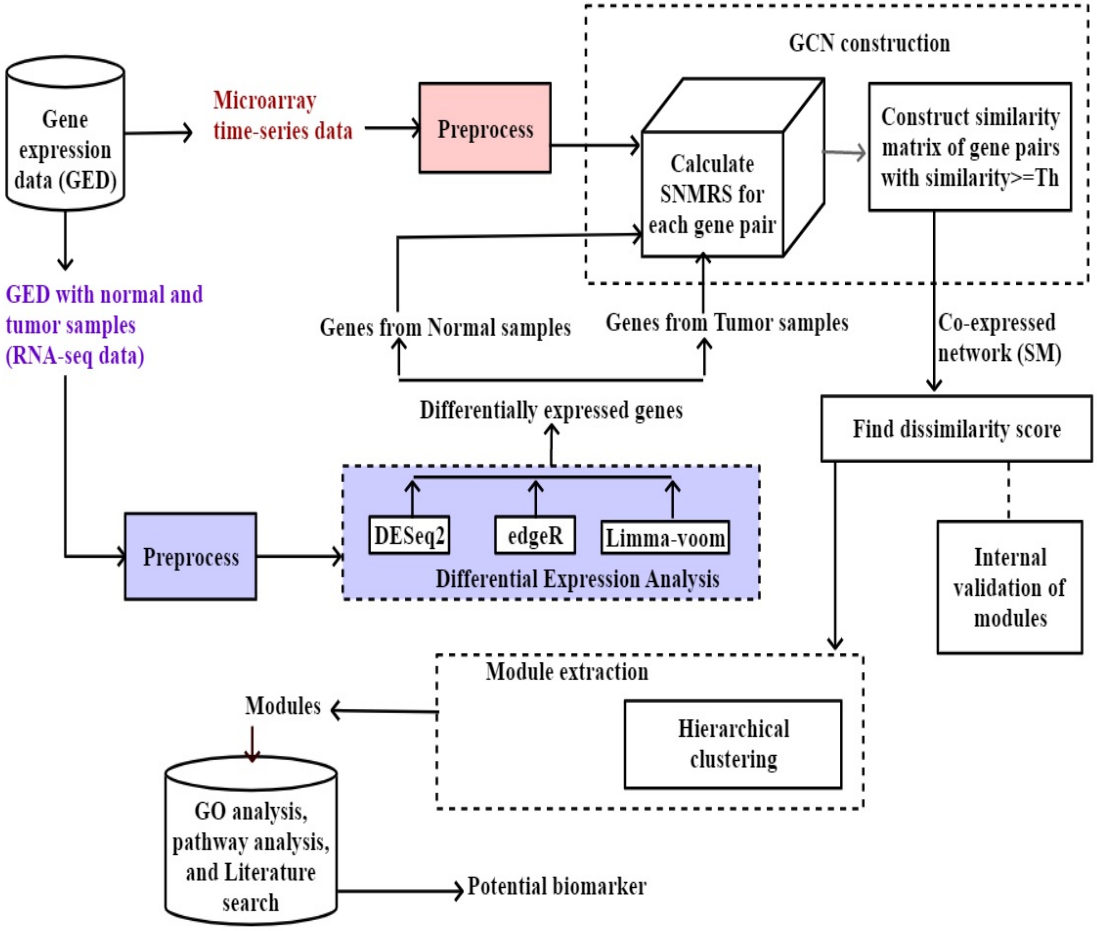


Figure 5-1: The workflow of module finding using SNMRS measure [DEA: Differential Expression Analysis]

Definition 1: *Two genes (g1, g2) are co-expressed and associated in forming the GCN iff their SNMRS score is* $\geq$ *a user-defined threshold value.*

Proposition 1: *If* M_i *is an enriched module extracted by this method and (g1,g2) is a pair of genes belonging to* M_i*, i.e. (g1,g2)*$\in M_i$*, then g1 and g2 are also functionally similar.*

Proof: If (g1,g2) are two genes and they belong to a module say M_i extracted by our method, i.e., (g1,g2)$\in$M$_i$, and assume that they are not similar functionally.

Algorithm 2: Construction of weighted signed GCN using SNMRS and extraction of modules

Input: Gene expression matrix (G), threshold
Output: Gene co-expression network, Gene Module

Preprocess the dataset.
Find SNMRS similarity matrix, SM by computing SNMRS for each pair of gene (gi, gj).
for $i \leftarrow 1$ to $nrow(dataset)$ **do**
  for $j \leftarrow 1$ to $ncol(dataset)$ **do**
    Compute SNMRS(gi,gj)
  end for
end for
Consider upper or lower triangle for the GCN.
if $(SNMRS(gi, gj) \leq threshold)$ **then**
  SM(i,j) = 0
end if
Gene co-expression network (Similarity matrix) ← SM
Dissimilarity matrix (DM) ← 1 - SM
Apply hierarchical clustering method on DM.
Set threshold to find minimum module size.
Identify different modules using dynamic cut tree technique
Assign different types of colors to the detected modules
A set of gene modules.

As per definition 1, two genes (g1,g2) are co-expressed and associated in forming the GCN if their co-expression similarity is ≥ 0.8. Again, our approach extracts a module say M_i from the co-expression network, if the co-expression similarity between any pair of genes, say (g1,g2)$\in$M$_i$ is very high i.e.$\geq$0.85. Therefore, the assumption is false and hence proved.

The construction of the similarity matrix involves a complexity of O(n(n-1))/2) and the finding of a module involves a complexity of O(nxn log n), where n is the number of genes.

5.3 Experimental Results

The proposed method is implemented in RStudio and use platform Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz 2.59 GHz processor with 16 GB memory running in Windows 10 operating system.

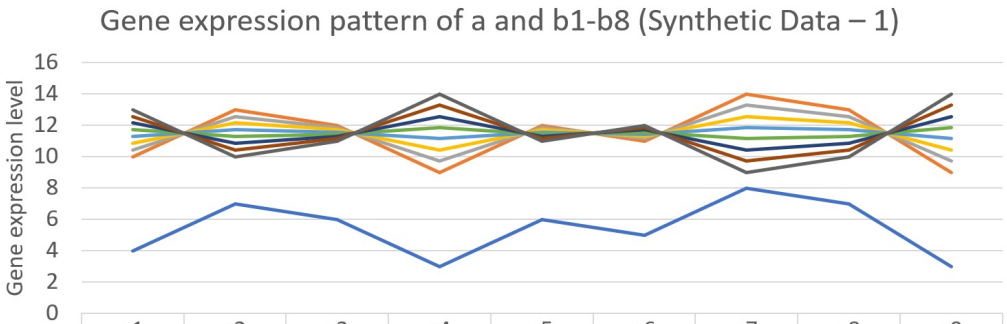
5.3.1 Benchmarking of the SNMRS measure using Synthetic and Multivariate data

5.3.1.1 Description of Datasets used

Three synthetic datasets and one Multivariate dataset called Iris from the UCI machine learning repository https://archive.ics.uci.edu/ml/datasets/iris are used to establish the effectiveness of the proposed measure SNMRS.

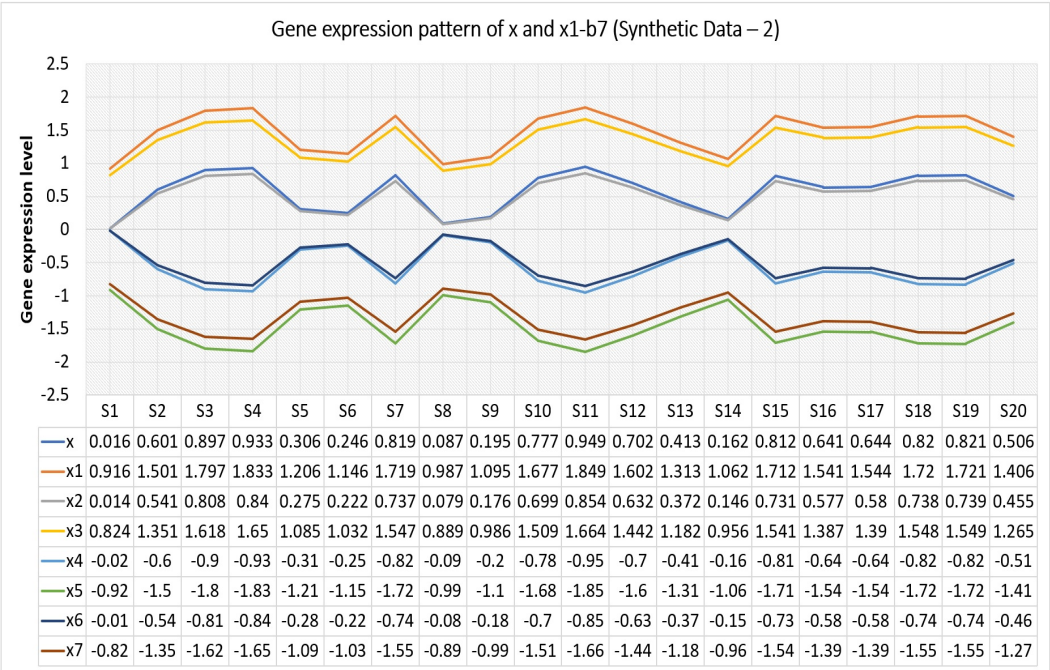
The *synthetic data - D1* shown in Figure 5-2 consists of 9 genes and 9 conditions; *synthetic data - D2* shown in Figure 5-3 contains 8 genes and 20 conditions; *synthetic data - D3* presented in Table 5.2 consists of 8 genes and 10 conditions. First *synthetic data - D1* is referred from [7]. *D1* contains 9 genes from 'a' to b1-b8' and 9 conditions from 'c1 to c9'. The variable 'b1' is perfectly shifted from variable 'a' and the variable 'b8' is negatively shifted from 'a'. The variable from 'b2 to b3' are obtained by distributing uniformly intermediate patterns between 'b1 and b8'. *D2* presents genes from 'x to x1-x7' and samples from 'S1 to S20'. Each gene 'x1 to x7' are associated with 'x' and thus all the genes from 'x to x7' are correlated with each other. 'x1, x2, and x3' have the *+shifting, +scaling, and +shifting-and-scaling* pattern, respectively. Again, 'x4, x5, x6, and x7' exhibit *-absolute, -shifting, -scaling, and -shifting-and-scaling* correlation patterns. *D3* contains genes from 'p to p1-p7' and samples from 'S1 to S10'. Each gene 'p1 to p1-p7' are associated with 'p' and thus all the genes are correlated with each other. 'p1, p2, and p3' have the +shifting, +scaling, and +shifting-and-scaling pattern. Again, 'p4, p5, p6, and p7' exhibit *-absolute, -shifting, -scaling, and -shifting-and-scaling* pattern.

The iris dataset includes samples from three different Iris species (Iris versicolor, iris virginica, and iris setosa). The length and width of the sepals and petals, both in centimetres, are measured for each sample. The dataset is composed of three classes, each with 50 instances, each referring to a different iris plant kind.



	c1	c2	c3	c4	c5	c6	c7	c8	c9
a	4	7	6	3	6	5	8	7	3
b1	10	13	12	9	12	11	14	13	9
b2	10.4286	12.5714	11.8571	9.7143	11.8571	11.1429	13.2857	12.5714	9.7143
b3	10.8571	12.1429	11.7143	10.4286	11.7143	11.2857	12.5714	12.1429	10.4286
b4	11.2857	11.7143	11.5714	11.1429	11.5714	11.4286	11.8571	11.7143	11.1429
b5	11.7143	11.2857	11.4286	11.8571	11.4286	11.5714	11.1429	11.2857	11.8571
b6	12.1429	10.8571	11.2857	12.5714	11.2857	11.7143	10.4286	10.8571	12.5714
b7	12.5714	10.4286	11.1429	13.2857	11.1429	11.8571	9.7143	10.4286	13.2857
b8	13	10	11	14	11	12	9	10	14

Figure 5-2: *Synthetic data - D1*: The artificial gene patterns for analysis of different similarity measures.



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
x	0.016	0.601	0.897	0.933	0.306	0.246	0.819	0.087	0.195	0.777	0.949	0.702	0.413	0.162	0.812	0.641	0.644	0.82	0.821	0.506
x1	0.916	1.501	1.797	1.833	1.206	1.146	1.719	0.987	1.095	1.677	1.849	1.602	1.313	1.062	1.712	1.541	1.544	1.72	1.721	1.406
x2	0.014	0.541	0.808	0.84	0.275	0.222	0.737	0.079	0.176	0.699	0.854	0.632	0.372	0.146	0.731	0.577	0.58	0.738	0.739	0.455
x3	0.824	1.351	1.618	1.65	1.085	1.032	1.547	0.889	0.986	1.509	1.664	1.442	1.182	0.956	1.541	1.387	1.39	1.548	1.549	1.265
x4	-0.02	-0.6	-0.9	-0.93	-0.31	-0.25	-0.82	-0.09	-0.2	-0.78	-0.95	-0.7	-0.41	-0.16	-0.81	-0.64	-0.64	-0.82	-0.82	-0.51
x5	-0.92	-1.5	-1.8	-1.83	-1.21	-1.15	-1.72	-0.99	-1.1	-1.68	-1.85	-1.6	-1.31	-1.06	-1.71	-1.54	-1.54	-1.72	-1.72	-1.41
x6	-0.01	-0.54	-0.81	-0.84	-0.28	-0.22	-0.74	-0.08	-0.18	-0.7	-0.85	-0.63	-0.37	-0.15	-0.73	-0.58	-0.58	-0.74	-0.74	-0.46
x7	-0.82	-1.35	-1.62	-1.65	-1.09	-1.03	-1.55	-0.89	-0.99	-1.51	-1.66	-1.44	-1.18	-0.96	-1.54	-1.39	-1.39	-1.55	-1.55	-1.27

Figure 5-3: *Synthetic data - D2*: The artificial gene patterns for analysis of different similarity measures.

5.3.1.2 SNMRS vs other measures: A comparison

Seven similarity measures PCC, NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS are applied on synthetic datasets *D1, D2,* and *D3* and evaluated

Table 5.2: *Synthetic data - D3*

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p	1	3	1	3	1	3	1	3	1	3
p1	3	5	3	5	3	5	3	5	3	5
p2	3	9	3	9	3	9	3	9	3	9
p3	9	15	9	15	9	15	9	15	9	15
p4	-1	-3	-1	-3	-1	-3	-1	-3	-1	-3
p5	-3	-5	-3	-5	-3	-5	-3	-5	-3	-5
p6	-3	-9	-3	-9	-3	-9	-3	-9	-3	-9
p7	-9	-15	-9	-15	-9	-15	-9	-15	-9	-15

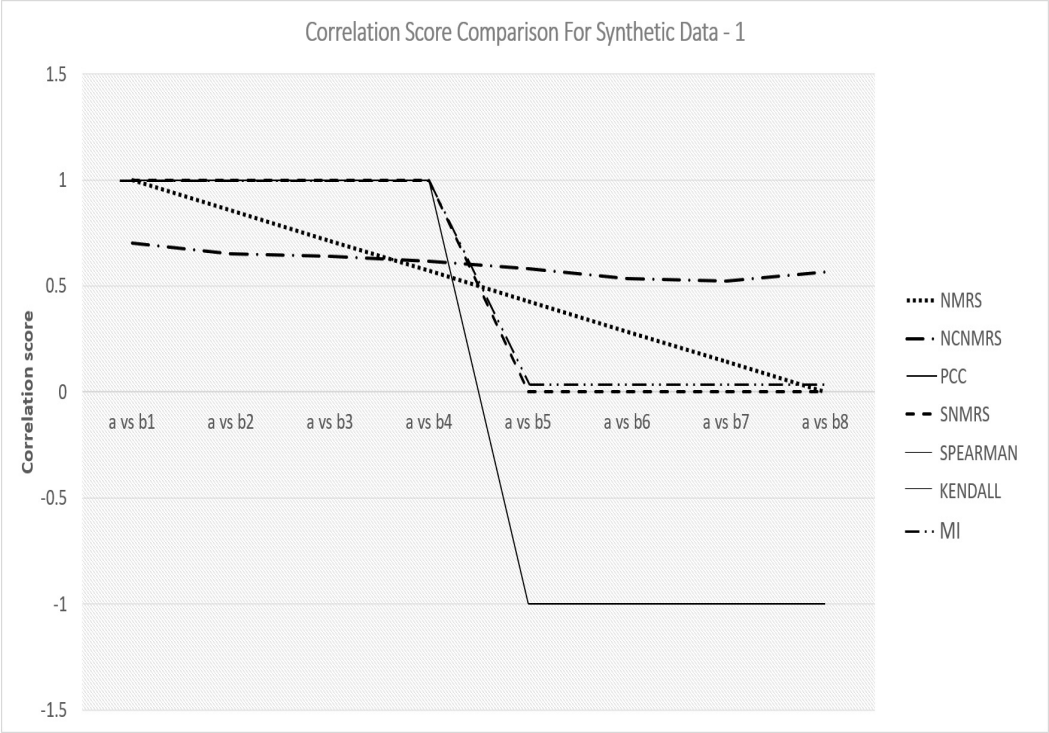


Figure 5-4: Correlation values obtained from NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS while applied on example patterns 'b1-b8' with that of 'a'

based on their co-expression results.

Figure 5-4 is the output for given synthetic data *D1*. This figure presents correlation scores given by each measure for gene pattern 'a' with 'b1-b8'. Figure 5-5 shows the correlation values resulting from NMRS, NCNMRS, PCC, SPEARMAN, KENDALL, MI, and SNMRS for each gene pair (x, x1)-(x, x7) for *D2*. For *D3*, the resultant correlation scores are shown by plotting a graph in Figure 5-6. In Figure 5-4 to 5-6, correlation scores given by NMRS, NCNMRS, PCC,

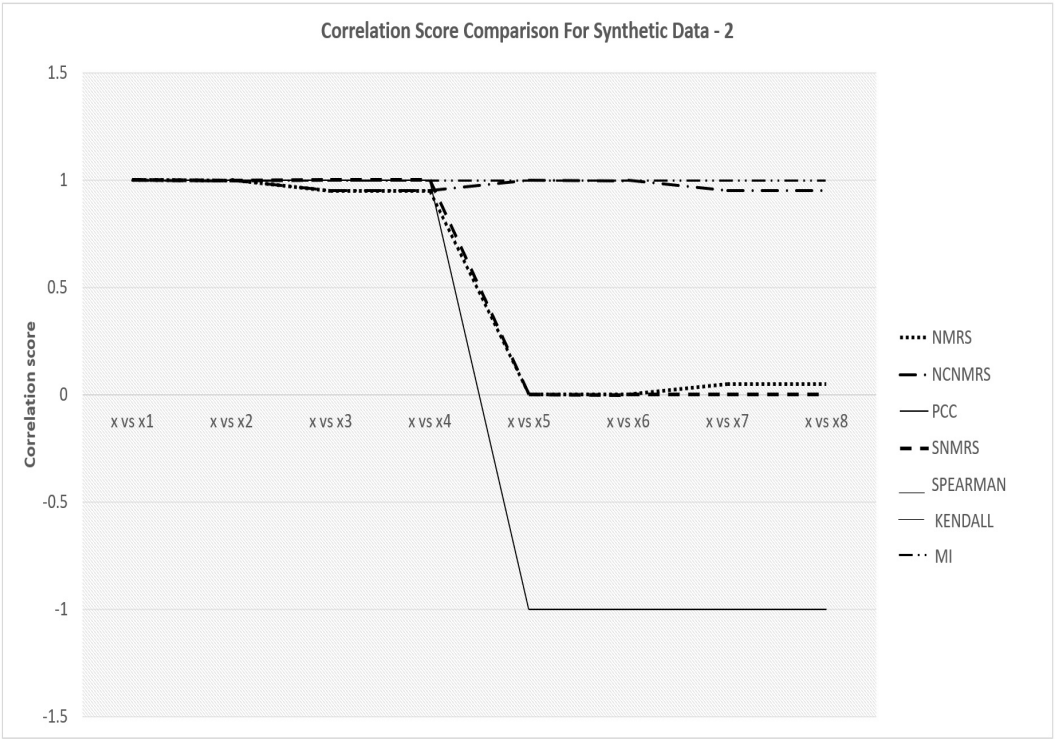


Figure 5-5: Correlation values obtained from NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS while applied on example patterns 'x1-x7' with that of 'x'

SPEARMAN, KENDALL, MI, and SNMRS are plotted individually. It has been observed that NMRS, NCNMRS, and SNMRS give output in the range between 0 to 1. Also, SNMRS, SPEARMAN, KENDALL, and PCC can identify different types of correlation patterns for each gene pair whereas NMRS, MI and NCNMRS give some undesired correlation values for gene pairs (a, b1)-(a, b8), (x, x1)-(x, x8) and (p, p2)-(p, p8). SNMRS, SPEARMAN, KENDALL, and PCC are found capable to distinguish patterns with shifting, scaling, and shifting-and-scaling associations. Further, it is observed that the negative correlation is -1 to 0 for PCC, SPEARMAN, and KENDALL, but for SNMRS it is from 0 to 0.5.

5.3.1.3 Internal cluster validation of SNMRS with Iris data

Iris (Fisher 1936) dataset is considered and NbClust [271] is used on Iris data for internal cluster validity indices implemented in the package. One of the most useful features of this package is that it includes a comprehensive list of indices

Table 5.3: Comparisons among several measures including SNMRS in terms of internal cluster validation results for Iris data for average linkage hierarchical clustering.

	KL	**CH**	**CCC**	**Scott**	**TrCovW**	**TraceW**	**Friedman**	**Rubin**	**Cindex**	**DB**	**Silhouette**	**Duda**	**Pseudot2**
Euclidean	5.67	**502.8216**	**35.7286**	**1089.296**	948.8461	154.947	**761.2242**	61.5649	0.2718	**0.436**	0.6867	**0.4599**	115.0825
NMRS	**15.2096**	**502.8216**	**35.7286**	**1089.296**	948.8461	154.947	**761.2242**	61.5649	0.3435	**0.436**	0.7233	0.6711	48.0304
NCNMRS	**15.2096**	**502.8216**	**35.7286**	**1089.296**	948.8461	154.947	**761.2242**	61.5649	0.3435	**0.436**	0.7233	0.6711	48.0304
PCC	1.3062	0.026	16.9695	732.464	**59030.17**	**681.2507**	642.2833	**14.0026**	0.7685	6.1556	0.0834	1.0135	-1.9566
SPEARMAN	1.3384	0.4757	17.0079	734.9767	58634.81	679.1878	651.0814	14.0451	0.5511	1.4403	0.0258	1.008	-1.1613
KENDALL	1.3384	0.4757	17.0079	734.9767	58634.81	679.1878	651.0814	14.0451	0.5511	1.4403	0.0336	1.008	-1.1613
MI	1.3152	0.5314	17.0126	734.4479	58537.72	678.9329	645.2954	14.0504	0.3497	1.3626	-0.0168	1.0176	**-2.5444**
SNMRS	13.6194	**502.8216**	**35.7286**	**1089.296**	948.8461	154.947	**761.2242**	61.5649	**0.1082**	**0.436**	**0.9202**	0.9368	6.6137

Table 5.4: Continuation of Table 5.3.

	Ratkowsky	Ball	Ptbiserial	Gap	McClain	Gamma	Gplus	Dunn	Hubert	SDindex	Dindex	SDbw
Euclidean	**0.5535**	77.4735	0.8358	0.1282	0.2622	0.9587	57.0657	0.3389	0.0019	**1.2114**	0.8535	**0.1578**
NMRS	**0.5535**	77.4735	0.9291	0.1282	0.2349	**0.9996**	**0.6048**	0.628	0.0298	1.6272	0.8535	**0.1578**
NCNMRS	**0.5535**	77.4735	0.9291	0.1282	0.2349	**0.9996**	**0.6048**	0.628	0.0298	1.6272	0.8535	**0.1578**
PCC	0.0127	**340.6254**	0.0508	**-1.3527**	0.0125	0.0475	69.9756	**0.7679**	**0.0924**	16.3001	**1.9417**	1.4598
SPEARMAN	0.0254	339.5939	0.0261	-1.3496	0.0129	0.2346	27.1857	0.6667	0.0115	9.03	1.9359	2.3765
KENDALL	0.0254	339.5939	0.0261	-1.3496	0.0129	0.2346	27.1857	0.6667	0.0115	9.03	1.9359	2.3765
MI	0.0371	339.4664	0.0275	-1.3493	**0.0118**	0.1715	44.4658	0.1303	0.0079	8.7445	1.9356	2.4181
SNMRS	**0.5535**	77.4735	**0.9412**	0.1282	0.0705	0.9985	2.0505	0.4787	0.0272	3.3145	0.8535	**0.1578**

Table 5.5: Comparisons among several measures including SNMRS in terms of internal cluster validation results for Iris data and for K-means clustering.

	KL	**CH**	**CCC**	**Scott**	**TrCovW**	**TraceW**	**Friedman**	**Rubin**	**Cindex**	**DB**	**Silhouette**	**Duda**	**Pseudot2**
Euclidean	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	0.2728	**0.4744**	0.681	**1.9253**	**-52.8667**
NMRS	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	0.2918	**0.4744**	0.6903	**1.9253**	**-52.8667**
NCNMRS	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	**0.0623**	**0.4744**	**0.9098**	**1.9253**	**-52.8667**
PCC	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	0.9377	**0.4744**	-0.286	**1.9253**	**-52.8667**
SPEARMAN	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	0.9751	**0.4744**	-0.1817	**1.9253**	**-52.8667**
KENDALL	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	0.9751	**0.4744**	-0.3037	**1.9253**	**-52.8667**
MI	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	0.4678	**0.4744**	-0.7097	**1.9253**	**-52.8667**
SNMRS	**5.9068**	**513.9245**	**35.9428**	**1044.605**	**1045.97**	**152.348**	**732.8086**	**62.6152**	***0.1266***	**0.4744**	***0.8677***	**1.9253**	**-52.8667**

Table 5.6: Continuation of Table 5.5

	Ratkowsky	**Ball**	**Ptbiserial**	**Gap**	**McClain**	**Gamma**	**Gplus**	**Dunn**	**Hubert**	**SDindex**	**Dindex**	**SDbw**
Euclidean	**0.5462**	**76.174**	0.8345	**0.1448**	0.2723	0.9563	60.6542	0.0765	0.0019	**1.6173**	**0.8556**	**0.1618**
NMRS	**0.5462**	**76.174**	**0.8902**	**0.1448**	0.2584	**0.9689**	**43.1258**	***0.1245***	**0.0282**	**1.6173**	**0.8556**	**0.1618**
NCNMRS	**0.5462**	**76.174**	0.8727	**0.1448**	**0.0509**	***0.9632***	***51.0459***	3.00E-04	0.0206	**1.6173**	**0.8556**	**0.1618**
PCC	**0.5462**	**76.174**	-0.8727	**0.1448**	1.1958	-0.9632	2724.503	**0.3574**	-0.0177	**1.6173**	**0.8556**	**0.1618**
SPEARMAN	0.5462	76.174	-0.9207	0.1448	1.0449	-0.9969	2553.682	0.8	-0.0236	**1.6173**	**0.8556**	**0.1618**
KENDALL	0.5462	76.174	-0.9207	0.1448	1.2325	-0.9969	2553.682	0.6667	-0.0192	**1.6173**	**0.8556**	**0.1618**
MI	0.5462	76.174	-0.8905	0.1448	3.1041	-0.9772	2686.793	0.1044	-0.0129	**1.6173**	**0.8556**	**0.1618**
SNMRS	**0.5462**	**76.174**	***0.8844***	**0.1448**	***0.0906***	0.9383	84.0561	0.1051	***0.0252***	**1.6173**	**0.8556**	**0.1618**

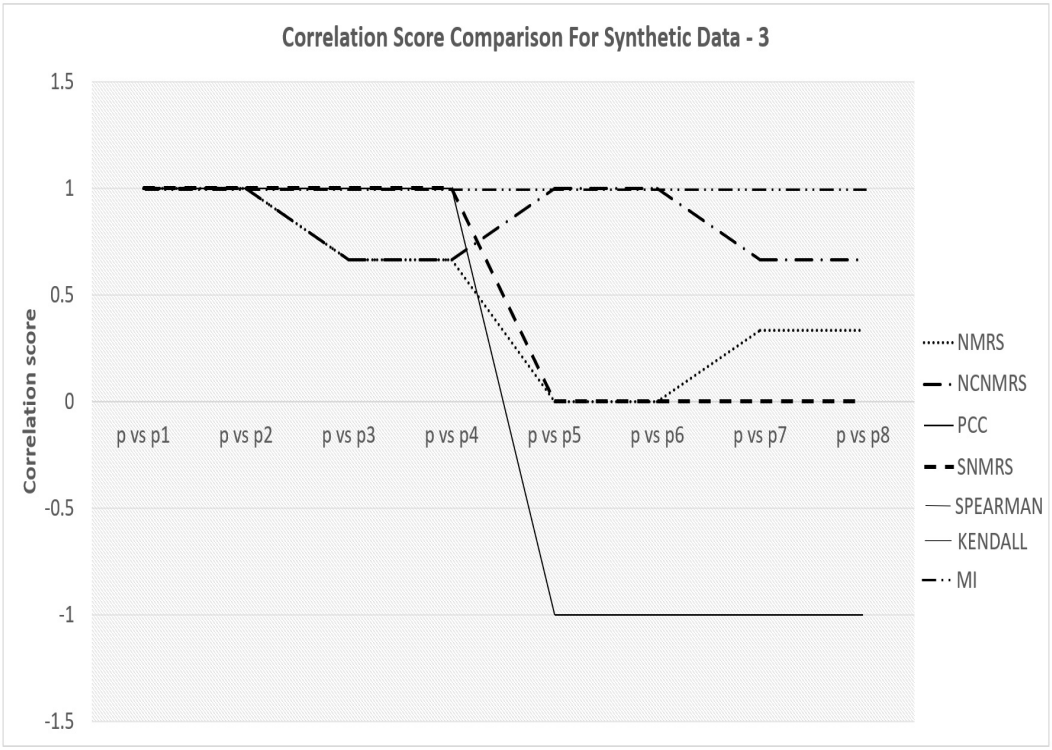


Figure 5-6: Correlation values obtained from NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS while applied on example patterns 'p1-p7' with that of 'p'

in the R package. It allows the user to change the number of clusters, the clustering algorithm, and the indices all at the same time to determine the optimal way to group the observations in a dataset. NbClust also recommends the ideal number of clusters for each index. Validity indices are found by applying the clustering method on dissimilarity matrices obtained by computing SNMRS, NMRS, NCNMRS, PCC, SPEARMAN, KENDALL, MI, and Euclidean measures individually so that all these measures can be compared and know the performance of each measure while clustering. According to the majority rule, the Iris dataset's optimal number of clusters for each case is two i.e. individual dissimilarity matrix obtained by SNMRS, NMRS, NCNMRS, PCC, SPEARMAN, KENDALL, MI, and Euclidean measure. A performance comparison table is shown in Table 5.3- 5.6. Here, 25 cluster validity indices [271] are considered and found the performance of SNMRS is better in most indices as compared to NMRS, NCNMRS, PCC, SPEARMAN, KENDALL, MI, and euclidean. This experiment is

performed for average linkage hierarchical and K-means clustering algorithms. It is found that for Iris data, average hierarchical clustering method is effective than k-means while applying SNMRS. The values with the bold font (in Table 5.3- 5.6) signifies the better output for the corresponding measure and the italic font indicates the second better index score among each other. In order to examine the degree of goodness of a clustering structure without referring to external data, internal cluster validation with the help of internal knowledge of the clustering process is carried out. Further, the clustering procedure and the total number of clusters can also be estimated with the help of this technique without any use of external data.

5.3.2 Benchmarking of the SNMRS method using independent RNA-seq and microarray dataset

5.3.2.1 Datasets used and preprocessing

A time-series microarray dataset and an RNA-seq dataset are used which are associated with Esophageal Squamous Cell Carcinoma (ESCC) disease. Microarray dataset is Yeast Sporulation with 6118 gene profiles measured across 7 different time points downloaded from the website [1] and the RNA-seq data used here is ESCC - accession number RP064894 with 58000 genes and 29 samples downloaded from Recount2 [2]. ESCC dataset consists of 14 tumor and 15 normal samples. The yeast Sporulation dataset is log-transformed and among the 6118 genes, the genes whose expression levels are not changing significantly have been ignored from further analysis. After preprocessing Yeast Sporulation dataset consists of 474 genes. Differentially expressed genes are found from ESCC dataset using DEseq2 [39], edgeR [37], and Limma-Voom [44] using the method reported in [272]. Preprocessing of the ESCC dataset is done by discarding low read count instances with a user-defined threshold (here, it is 5). TMM normalization method available in

[1] http://cmgm.stanford.edu/pbrown/sporulation
[2] https://jhubiostatistics.shinyapps.io/recount/

edgeR is used to obtain the normalized expression values of the RNA-seq dataset. Transformation of the data is done by using "vst" transformation method available in the R package. At last 5165 genes from the ESCC dataset are considered for further analysis. Ensemble id of each gene is mapped to Official Symbol gene name.

5.3.3 Threshold selection

Our experimental study reveals that for effective GCN analysis the appropriate SNMRS threshold value is 0.8. Figure Figure 5-7 evidenced the decision. Figure 5-7 presents a comparative study which depicts the change of node and change of a number of edges or connectivity with different values of SNMRS respectively. It is observed that from a threshold value of 0.8 a drastic change can be seen in the number of nodes. A number of nodes are constant up to correlation value 0.7 for Yeast Sporulation and ESCC datasets and after that gradually it decreases. Also for connectivity, it is noticed that the detected number of edges decrease with the increase of the SNMRS score.

5.3.4 Time Series data

The yeast sporulation dataset is preprocessed and the SNMRS score is computed for each gene pair. To understand the performance of NMRS, PCC, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS, they are applied separately on Yeast Sporulation dataset with different clustering algorithms. Internal validation of clusters for each case is examined using the Dunn index, Silhouette width, cindex, and Mcclain index. The validation results are reported in Table 5.7. Silhouette width [273] and Dunn index [274] combine measures of compactness and separation of the clusters. From the comparison, it can be observed that for measure SNMRS the compactness and separation scores are better as compared to other measures in most cases. Since the hierarchical clustering method gives the best result for SNMRS, this clustering method is used in the subsequent downstream analysis.

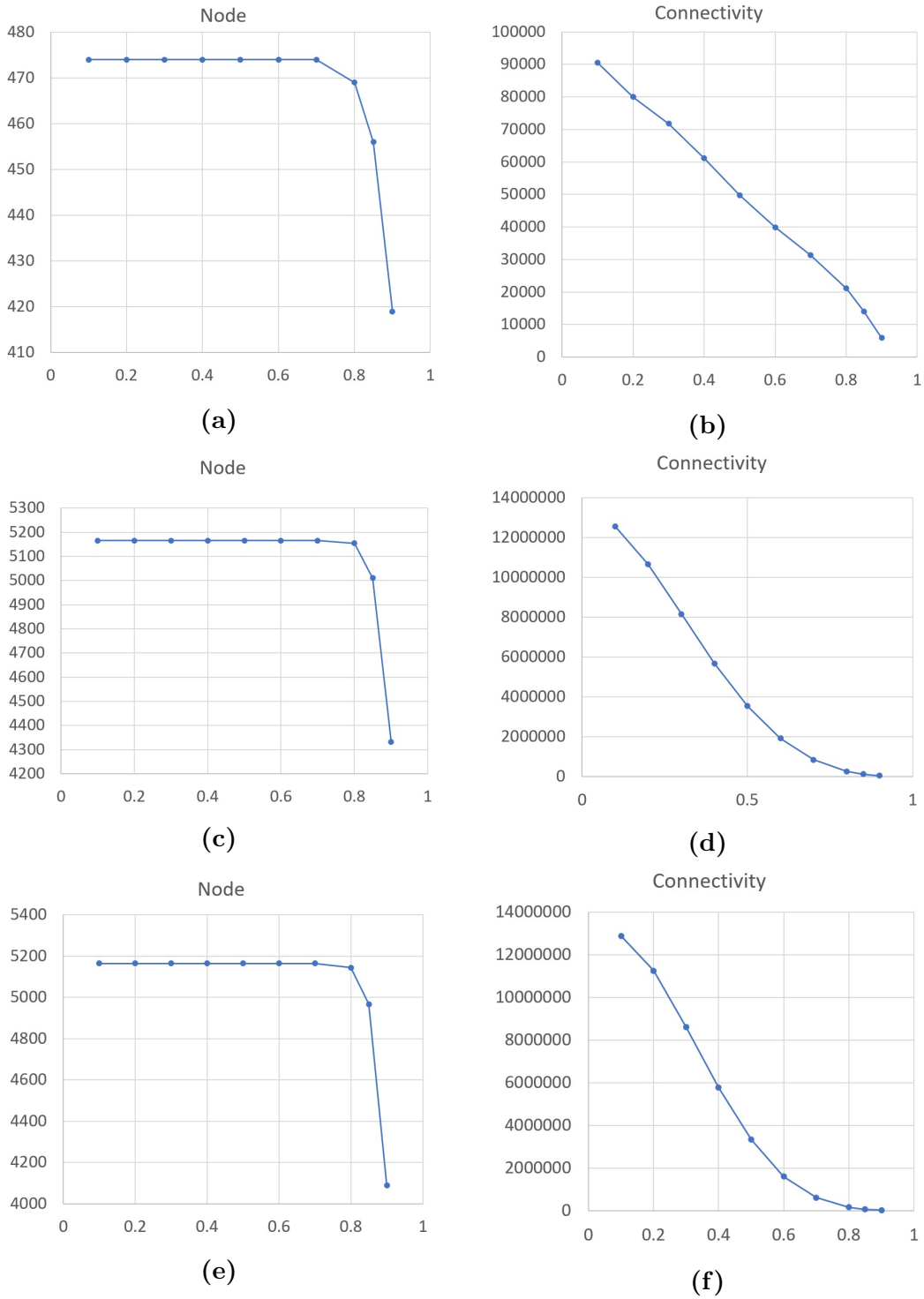


Figure 5-7: Comparative study of node (Y-axis) and connectivity (Y-axis) at different correlation scores (X-axis) to analyse the threshold value. Effect diagram of the node after threshold 0.8 is increased. Here, (a) Node vs SNMRS for Yeast sporulation dataset, (b) Connectivity vs SNMRS for Yeast sporulation dataset, (c) Node vs SNMRS for ESCC normal dataset, (d) Node vs SNMRS for ESCC normal dataset, (e) Node vs SNMRS for ESCC tumor dataset, and (f) Node vs SNMRS for ESCC tumor dataset.

Using the corresponding dissimilarity scores of SNMRS values, the average linkage clustering method is applied. Finally, the dynamic tree cut approach is used, which allows us to find a total of five gene modules. It extracts a total of 6 modules as green, turquoise, blue, yellow, brown, and red. The number of genes detected in each module is reported in Table 5.8.

Each module in this experiment contains a different number of genes. After finding the modules, validation is done in terms of the p-value. Gene enrichment analysis is performed for each module using DAVID [3] and through the gene-ontology (GO) and KEGG pathway analyses, it is observed that the percentage of enrichment of genes belonging to a module in Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) is more than 90% in most cases, as reported in Table 5.8. The table 5.9 reports the GO enrichment term and KEGG pathway with the lowest p-values for each module.

5.3.5 Gene Expression data with normal and tumor conditions

Preprocessed ESCC dataset is vertically partitioned into two subsets based on the type of samples. One subset is for normal samples and the other is for tumor samples. Experiments have been performed for these two. For each pair of genes from both datasets, the SNMRS score is computed and found two separate similarity matrices (SMN and SMT). The GCN is constructed using the threshold 0.8 as this threshold signifies the highly correlated genes. The dissimilarity score is calculated as 1-SMN for the normal sampled dataset and 1-SMT for tumor sampled dataset. The average linkage clustering method is applied using the corresponding SNMRS dissimilarity scores. Finally, the dynamic tree cut technique is used and a total of 16 gene modules are identified for the ESCC normal dataset and 17 gene-modules for the ESCC tumor dataset through the dynamic tree cut method. The Figure 5-8-5-9 present the pictorial view of expression patterns of each gene present in a

[3]https://david.ncifcrf.gov/

Table 5.7: Comparisons among measures including SNMRS in terms of internal cluster validation results for Yeast Sporulation Gene expression data for average linkage hierarchical clustering.

Validity	**Euclidean**		**PCC**		**NMRS**		**NCNMRS**		**SPEARMAN**		**KENDALL**		**MI**		**SNMRS**	
indices	**Best**	**nc**	**index**	**Best**	**nc**	**index**	**Best**	**nc**	**index**	**Best**	**nc**	**index**	**Best**	**nc**	**index**	**Best**	**nc**	**index**
silhouette	2	0.569	2	0.5985	10	0.1664	3	0.41	10	0.602	10	0.4408	9	0.6211	10	0.3294		
dunn	2	0.1934	2	0.0168	2	1	2	0.084	8	0.0455	2	0.1	2	-0.3802	2	1		
cindex	2	0.2932	10	0.1122	10	0.8503	10	0.334	7	0.1673	10	0.2777	10	0.399	10	0.6581		
mcclain	2	0.4064	2	0.1911	2	0.0036	2	0.003	2	0.2253	2	0.4907	10	0.2352	2	0.0035		

Table 5.8: Comparative enrichment analysis results obtained for Yeast sporulation dataset. Different types of colours are assigned to module names. Abbreviations: BP-Biological Process, MF-Molecular Function, CC-Cellular Component.

Module	**Gene**	**BP**	**CC**	**MF**	**KEGG pathway**
green	38	97.00%	100.00%	78.80%	33.30%
turquoise	190	83.40%	88.6%	65.7%	32.6%
blue	109	89.5%	97.1%	83.80%	41.00%
yellow	41	94.60%	94.60%	73.00%	24.30%
brown	56	98.00%	96.00%	96.00%	94.00%
red	32	93.5%	96.8%	80.60%	38.70%

Table 5.9: Comparative biological analysis results obtained by Yeast sporulation dataset

Module	CC	PValue	BP	PValue	MF	PValue	PATHWAY	PValue
green	GO:0005730 nucleolus	7.71E-23	GO:0042254 ribosome biogenesis	5.84E-19	GO:0030515 snoRNA binding	1.48E-04	sce03008: Ribosome biogenesis in eukaryotes	2.34E-10
turquoise	GO:0005628 prospore membrane	3.98E-31	GO:0030435 sporulation resulting in formation of a cellular spore	1.44E-42	GO:0004842 ubiquitin-protein transferase activity	3.15E-05	sce04113:Meiosis - yeast	1.85E-16
blue	GO:0009277 fungal-type cell wall	4.56E-07	GO:0006094 gluconeogenesis	8.60E-14	GO:0016829 lyase activity	2.91E-04	sce00010:Glycolysis / Gluconeogenesis	6.59E-16
yellow	GO:0000794 condensed nuclear chromosome	4.32E-07	GO:0051321 meiotic cell cycle	3.22E-17	GO:0008569 ATP-dependent microtubule motor activity, minus-end -directed	0.0211	sce04113:Meiosis - yeast	8.36E-07
brown	GO:0005840 ribosome	9.18E-46	GO:0002181 cytoplasmic translation	8.44E-55	GO:0003735 structural constituent of ribosome	1.47E-44	sce03010:Ribosome	4.60E-39
red	GO:0000794 condensed nuclear chromosome	3.28E-04	GO:0051321 meiotic cell cycle	2.73E-09	GO:0004092 carnitine O-acetyltransferase activity	0.0146	sce04113:Meiosis - yeast	0.002539

module of ESCC normal dataset and tumor subsets of dataset, respectively.

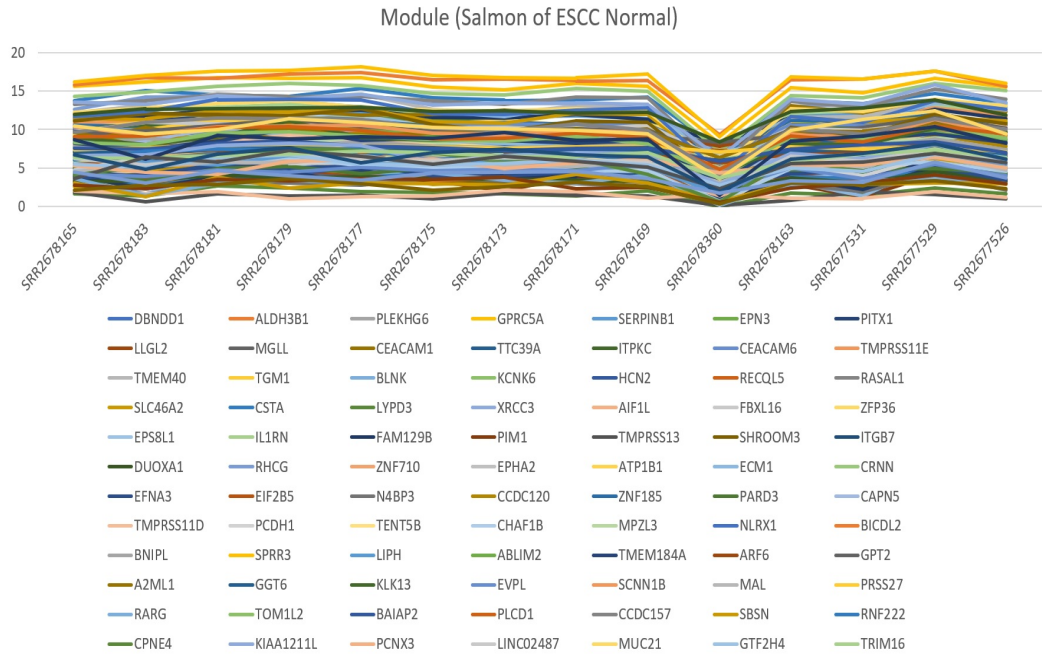


Figure 5-8: Visualization of extracted module (salmon) for ESCC normal

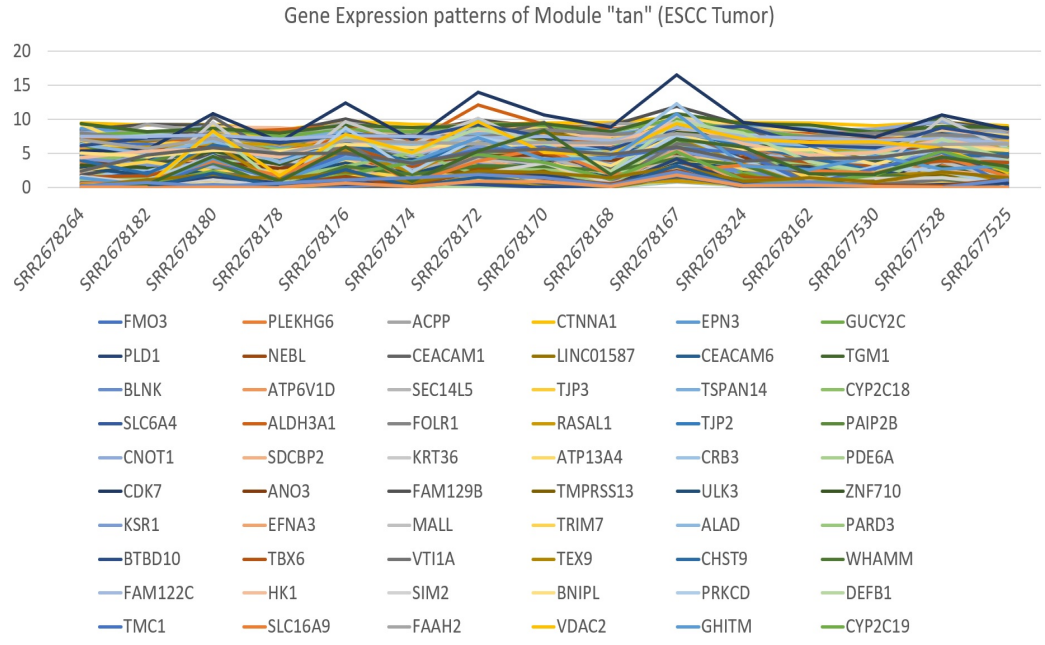


Figure 5-9: Visualization of extracted module (tan) for ESCC tumor

These clusters or modules are found highly co-expressed. Each module is assigned a colour name. GO enrichment analysis and pathway analysis are carried out to establish the performance by identifying biologically associated genes with

ESCC cancer for the proposed measure. The number of genes detected in each module and percentage of enriched GO terms are presented in Table 5.10 for ESCC normal dataset and in Table 5.11 for ESCC tumor dataset. Details of enriched GO terms and pathways with the lowest p-value for each module for ESCC normal and tumor are reported in Table 5.12 and 5.13.

Table 5.10: Comparative biological analysis results obtained using ESCC normal dataset. Different types of colours are assigned to module names. Abbreviations: BP-Biological Process, MF-Molecular Function, CC-Cellular Component.

Module Name	Gene	BP	CC	MF	KEGG pathway
black	285	82.10%	89.1%	82.5%	38.00%
blue	583	85.00%	88.30%	85.2%	36.00%
brown	457	82.10%	88.70%	82.10%	38.8%
cyan	196	85.50%	89.2%	87.10%	41.9%
green	397	81.80%	89.40%	81.6%	35.8%
greenyellow	251	80.40%	88.60%	80.80%	34.30%
grey	200	80.00%	82.60%	82.10%	38.4%
magenta	275	77.60%	87.60%	78.80%	31.70%
midnightblue	187	92.2%	96.1%	93.90%	43.3%
pink	283	87.70%	91.0%	86.6%	41.9%
purple	268	73.60%	78.50%	74.40%	27.60%
red	302	79.2%	84.20%	81.70%	29.00%
salmon	199	77.40%	88.70%	79.00%	27.4%
tan	246	87.1%	91.6%	89.80%	38.7%
turquoise	605	87.30%	91.80%	85.7%	34.90%
yellow	431	86.7%	92.00%	85.50%	37.9%

5.4 Discussion

As reported by Mahanta et. al [7], a line graph is plotted of the data shown in Figure 5-2 and visually observe that the genes (variables) are correlated or co-expressed with each other prominently. Patterns of 'a' and 'b1-b8' are matched at each condition. If one is going down, the other is also going down and vice versa. Mahanta et. al [7] reported that PCC results in some undesired output when applied to the artificial data having no shifting or scaling correlation pattern. However, NMRS [7] can distinguish patterns throughout this uniform distribution from a shifted pattern to a shifted and negatively correlated pattern of a given

Table 5.11: Comparative biological analysis results obtained using ESCC tumor dataset. Different types of colours are assigned to module names. Abbreviations: BP-Biological Process, MF-Molecular Function, CC-Cellular Component.

Module Name	Gene	BP	CC	MF	KEGG pathway
turquoise	580	84.2%	88.00%	82.5%	37.60%
purple	268	88.40%	96.50%	88.40%	38.80%
blue	457	86.60%	90.90%	85.60%	39.2%
brown	405	84.7%	88.40%	85.2%	34.70%
cyan	200	66.7%	79.70%	69.5%	24.3%
green	364	76.5%	82.8%	78.5%	31.4%
greenyellow	238	82.0%	87.70%	83.30%	39.5%
grey	136	73.4%	78.10%	73.40%	25.0%
magenta	319	86.8%	91.80%	87.2%	41.10%
midnightblue	340	83.90%	87.7%	85.20%	31.60%
pink	340	81.70%	90.10%	82.6%	30.1%
red	356	89.20%	92.40%	87.2%	38.7%
salmon	229	81.9%	89.80%	81.9%	31.20%
tan	230	79.3%	86.20%	83.9%	38.7%
black	347	83.50%	89.00%	83.2%	37.50%
yellow	369	84.9%	90.3%	87.10%	38.90%
lightcyan	163	89.70%	92.90%	87.10%	40.00%

pattern by giving different correlation values of patterns 'b1-b8' with that of 'a'. It is true that from the 'b1 to b7' variable, there is no perfect pattern of neither shifting nor scaling but PCC can detect the pattern and give results as perfectly correlated. It is also plotted each pair of variables and scatter random plot for pair 'a' with 'b1-b8' shown in Figure 5-10a- 5-10h. It is known that PCC is a linear measure and it checks the perfect fitting of data in the linear regression line; based on the regression line PCC gives value with a sign positive or negative. The data of pair of genes with different types of correlation patterns are perfectly fitted in the straight line or regression line, that's why the PCC value 1 or -1 results for positively shifting patterns or negatively shifting patterns. Here, from Figure 5-10a- 5-10h, It is observed that the scatter plot behaviour of fitting is in the straight lines of each gene pairs 'a' with 'b1-b8', and found all are perfectly fitted. PCC is giving more reliable output than NMRS. NMRS gives correlation values within the range of 0 to 1 and PCC gives values in the range of -1 to +1. NMRS gives a perfectly negative correlation value of 0. The NMRS method fails while applied to synthetic data as mentioned in Table-1 for the detection of shifting and scaling

Table 5.12: P-value biological analysis results obtained by ESCC Normal dataset

Module Name	CC	PValue	BP	PValue	MF	PValue	KEGG_PATHWAY	PValue
turquoise	GO:0005654 nucleoplasm	2.09E-26	GO:0051301 cell division	2.81E-30	GO:0005524 ATP binding	8.67E-15	hsa04110 Cell cycle	7.73E-18
blue	GO:0005737 cytoplasm	8.68E-10	GO:0098609 cell-cell adhesion	1.25E-08	GO:0098641 cadherin binding involved in cell-cell adhesion	1.25E-08	hsa04114 Oocyte meiosis	3.40E-04
brown	GO:0070062 extracellular exosome	1.13E-21	GO:0006914 autophagy	4.55E-04	GO:0008417 fucosyltransferase activity	6.65E-04	hsa01100 Metabolic pathways	1.63E-04
cyan	GO:0005576 extracellular region	9.36E-08	GO:0030198 extracellular matrix organization	4.22E-08	GO:0008083 growth factor activity	2.54E-05	hsa05146 Amoebiasis	1.68E-04
green	GO:0005788 endoplasmic reticulum lumen	1.32E-08	GO:0034341 response to interferon-gamma	3.24E-06	GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds	1.70E-04	hsa04142 Lysosome	5.06E-15
greenyellow	GO:0005578 proteinaceous extracellular matrix	3.67E-04	GO:0030198 extracellular matrix organization	5.20E-07	GO:0003725 double-stranded RNA binding	7.21E-04	hsa04512 ECM-receptor interaction	8.33E-07
grey	GO:0043235 receptor complex	7.88E-04	GO:0048007 antigen processing and presentation, exogenous lipid antigen via MHC class Ib	0.00163	GO:0030884 exogenous lipid antigen binding	8.23E-04	hsa04640 Hematopoietic cell lineage	0.00207
magenta	GO:0005615 extracellular space	7.00E-08	GO:0001895 retina homeostasis	0.00127	GO:0008201 heparin binding	7.13E-04	hsa00260 Glycine, serine and threonine metabolism	0.01059
midnightblue	GO:0005887 integral component of plasma membrane	3.47E-07	GO:0006954 inflammatory response	6.24E-09	GO:0004872 receptor activity	6.43E-05	hsa04380 Osteoclast differentiation	9.75E-05
pink	GO:0031090 organelle membrane	0.0070	GO:0034220 ion transmembrane transport	9.62E-04	GO:0070330 aromatase activity	5.32E-04	hsa01100 Metabolic pathways	4.68E-05
purple	GO:0005578 proteinaceous extracellular matrix	2.83E-05	GO:0009952 anterior/posterior pattern specification	1.84E-07	GO:0004222 metalloendopeptidase activity	3.05E-05	hsa04620 Toll-like receptor signaling pathway	0.019438
red	GO:0005622 intracellular	9.30E-05	GO:0007264 small GTPase mediated signal transduction	0.00154	GO:0005149 interleukin-1 receptor binding	0.01273	hsa01100 Metabolic pathways	0.03585
salmon	GO:0070062 extracellular exosome	1.24E-07	GO:0030216 keratinocyte differentiation	4.60E-05	GO:0003810 protein-glutamine gamma-glutamyl-transferase activity	0.00257	hsa04960 Aldosterone-regulated sodium reabsorption	0.03233
tan	GO:0031012 extracellular matrix	1.09E-09	GO:0030199 collagen fibril organization	5.38E-06	GO:0048407 platelet-derived growth factor binding	6.03E-06	hsa04512:ECM receptor interaction	5.51E-10
black	GO:0005925 focal adhesion	0.002357549	GO:0071559 response to transforming growth factor beta	8.72E-05	GO:0015301 anion:anion antiporter activity	0.00339	hsa04670 Leukocyte transendothelial migration	0.00735
yellow	GO:0005654 nucleoplasm	3.30E-09	GO:0000398 mRNA splicing, via spliceosome	8.72E-05	GO:0044822 poly(A) RNA binding	1.89E-12	hsa00510 N-Glycan biosynthesis	7.72E-04

patterns. Further, it is observed that while testing for shifting and scaling for the given dataset, PCC, SPEARMAN, and KENDALL met it in a perfect match.

Therefore, it is concluded that NMRS, NCNMRS, and MI are not able to provide the results of scatter plot which in turn leads to their conclusion, but while testing for the same with the PCC it is very prominent even from Figure 5-10a- 5-10h. For the data, it achieves the output 0 or 1 because the standard deviation is the same for that pair. Moreover, it is found from this analysis that NMRS could not detect perfect shifting, scaling, and shifting-and-scaling correlations which have motivated us to enhance this measure and consequently, SNMRS has been introduced.

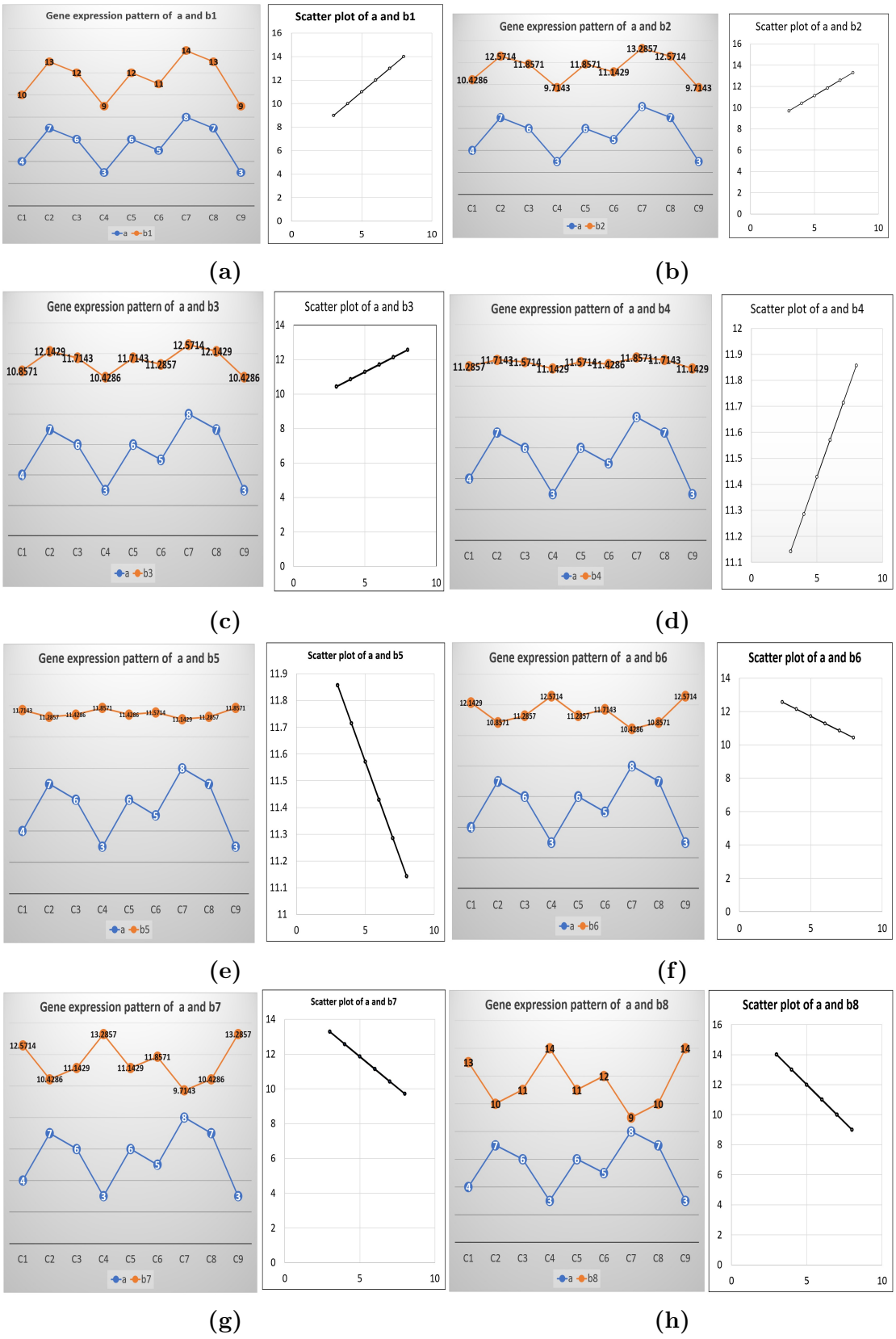


Figure 5-10: The artificial gene patterns and their scatter plot of Synthetic data - D1: (a) a vs b1, (b) a vs b2, (c) a vs b3, (d) a vs b4, (e) a vs b5, (f) a vs b6, (g) a vs b7, (h) a vs b8

Table 5.13: P-value biological analysis results obtained by ESCC Tumor dataset

Module Name	CC	PValue	BP	PValue	MF	PValue	KEGG_PATHWAY	PValue
turquoise	GO:0070062 extracellular exosome	2.17E-07	GO:0006066 alcohol metabolic process	3.89E-04	GO:0004029 aldehyde dehydrogenase (NAD) activity	9.22E-04	hsa00071 Fatty acid degradation	3.07E-06
purple	GO:0005578 proteinaceous extracellular matrix	4.70E-18	GO:0030198 extracellular matrix organization	2.30E-15	GO:0005201 extracellular matrix structural constituent	8.17E-11	hsa04510 Focal adhesion	2.95E-08
blue	GO:0005813 centrosome	7.03E-08	GO:0007067 mitotic nuclear division	7.88E-11	GO:0005524 ATP binding	4.62E-08	hsa05323 Rheumatoid arthritis	2.08E-07
brown	GO:0005654 nucleoplasm	1.60E-10	GO:0006281 DNA repair	5.96E-07	GO:0044822 poly(A) RNA binding	4.70E-08	hsa00240 Pyrimidine metabolism	0.01190
cyan	GO:0005615 extracellular space	3.18E-05	GO:0070244 negative regulation of thymocyte apoptotic process	0.0544	GO:0005198 structural molecule activity	0.002122	hsa04512 ECM-receptor interaction	0.01572
green	GO:0030672 synaptic vesicle membrane	0.001633	GO:0050905 neuromuscular process	0.00148	GO:0042826 histone deacetylase binding	0.023621	hsa04727 GABAergic synapse	0.04292
greenyellow	GO:0070062 extracellular exosome	6.28E-05	GO:0016310 phosphorylation	0.005094	GO:0019902 phosphatase binding	0.0138858	hsa01100 Metabolic pathways	2.83E-05
grey	GO:0016021 integral component of membrane	0.031715	GO:0060037 pharyngeal system development	0.004382	GO:0030276 clathrin binding	0.0333156	-	-
magenta	GO:0031012 extracellular matrix	1.55E-07	GO:0030198 extracellular matrix organization	1.90E-09	GO:0004222 metalloendopeptidase activity	7.07E-05	hsa04512 ECM-receptor interaction	3.12E-06
midnightblue	GO:0005737 cytoplasm	6.42E-05	GO:0018149 peptide cross-linking	3.98E-05	GO:0005198 structural molecule activity	0.012666	hsa04973 Carbohydrate digestion and absorption	0.03438
pink	GO:0070062 extracellular exosome	2.26E-06	GO:0008544 epidermis development	5.38E-05	GO:0098641 cadherin binding involved in cell-cell adhesion	5.69E-05	hsa01100 Metabolic pathways	0.00750
red	GO:0005829 cytosol	4.79E-06	GO:0098609 cell-cell adhesion	0.001465	GO:0004842 ubiquitin-protein transferase activity	2.70E-04	hsa04114 Oocyte meiosis	0.00125
salmon	GO:0000502 proteasome complex	3.95E-05	GO:0002479 antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	4.88E-05	GO:0005515 protein binding	4.64E-05	hsa03050 Proteasome	5.67E-05
tan	GO:0005923 bicellular tight junction	2.41E-07	GO:0006805 xenobiotic metabolic process	0.001197	GO:0015280 ligand-gated sodium channel activity	0.003068	hsa04530 Tight junction	8.61E-06
black	GO:0070062 extracellular exosome	8.05E-08	GO:0034372 very-low-density lipoprotein particle remodeling	1.42E-04	GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds	0.012182	hsa04142 Lysosome	2.52E-09
yellow	GO:0005654 nucleoplasm	2.04E-18	GO:0051301 cell division	2.63E-21	GO:0005524 ATP binding	2.39E-07	hsa04110 Cell cycle	1.61E-14
lightcyan	GO:0043025 neuronal cell body	1.23E-06	GO:0006954 inflammatory response	2.93E-06	GO:1902282 voltage-gated potassium channel activity involved in ventricular cardiac muscle cell action potential repolarization	0.001697	hsa00531 Glycosaminoglycan degradation	5.81E-04

SNMRS has been tested on publicly available datasets. The network modules determined by our method have been biologically validated in terms of the p-value. The measure SNMRS can detect absolute, scaling, shifting, and shifting-and-scaling correlation patterns, and it has the ability to discover biologically significant network modules from GCN, according to our findings. P-values are used to evaluate the biological significance of the gene sets contained in the derived network modules. [275]. The p-value indicates how well these genes match various

GO categories.

DAVID, a web-based tool is used to calculate the p-value [4]. Based on Molecular Function, Cellular Component, and Biological Process annotations, DAVID computes the hyper-geometric functional enrichment score. Tables 5.8-5.13 show the enriched functional categories for some of the network modules produced using the proposed method on the datasets. The GCN modules produced by the method for yeast Sporulation dataset include highly enriched GO terms such as nucleolus, prospore membrane, fungal-type cell wall, ribosome, chromosome, ribosome biogenesis, sporulation resulting in the formation of a cellular spore, gluconeogenesis, cytoplasmic translation, meiotic cell cycle, snoRNA binding, ubiquitin-protein transferase activity, structural constituent of ribosome, and carnitine O- acetyltransferase activity. KEGG pathway analysis results involved genes in Ribosome biogenesis in eukaryotes, Meiosis - yeast, Glycolysis / Gluconeogenesis, Ribosome, and Meiosis - yeast.

In Tables 5.10- 5.13 biological validation results are shown for ESCC dataset. The co-expression network modules for ESCC normal dataset produced by our method contains the highly enriched cellular components, Biological process and Molecular Function such as cell division, extracellular matrix organization, poly(A) RNA binding. We observe that the genes of modules follow either an ESCC related significant pathway or a GO annotation term using KEGG pathway and gene-ontology analysis such as Focal adhesion reported in [276], Fatty acid degradation reported in [277], Tight junction reported in [278], Oocyte meiosis reported in [248], Proteasome reported in [279], Lysosome reported in [280], Metabolic pathways [281], extracellular matrix reported in [282] [283], mitotic nuclear division reported in [284], ATP binding reported in [285]. Based on the reported p-values, we may conclude that our suggested technique using SNMRS provides good enrichment functional category. As a result, it projects positive biological importance.

[4]https://david.ncifcrf.gov/

A PPI network has been generated using STRING web tool for the module (Yellow) of ESCC normal presented in the Figure 5-11 which reveals existence of connectivity among most of the genes from biological databases and known web resources.

Module preservation analysis is carried out and two non-preserved modules [78] are found with zsummary score [78] 3.2 and 3.3 and median rank [78] 62 and 61, respectively. From these modules we find the hub genes using the intramodular connectivity method available in WGCNA [78]. From these hub gene lists IP6K3, EMP1, PIK3C2B, FMO2, FREM2, and AJAP1 potential biomarkers are identified. Details are given in Table 5.14.

SNMRS is an advanced version of NMRS and it outperforms many co-expression measures such as NMRS, NCNMRS, PCC, Spearman, MI, and Kendall. It is a suitable method to construct a co-expression network from microarray and RNA-seq data. The main limitation of this method that we have not yet tested in with Single-cell RNA sequencing (scRNA-seq).

GCN is used to extract biologically relevant information, such as for the identification of novel genes not yet associated with explicit biological function, processes and phenomena. The assumption is that tightly co-expressed genes (modules) are associated with similar types of biological processes. Therefore, new functional associations can be derived for causality. This can be used in identifying novel biomarkers across domains such as basic biology, biotechnology, medicine (identifying disease-causing genes, diagnostic and therapeutic targets), agriculture (identifying stress tolerant genes in crops, genes for better traits), microbiology, environmental science etc. These methods are benchmarked using two independent datasets from human cancer and Yeast Sporulation.

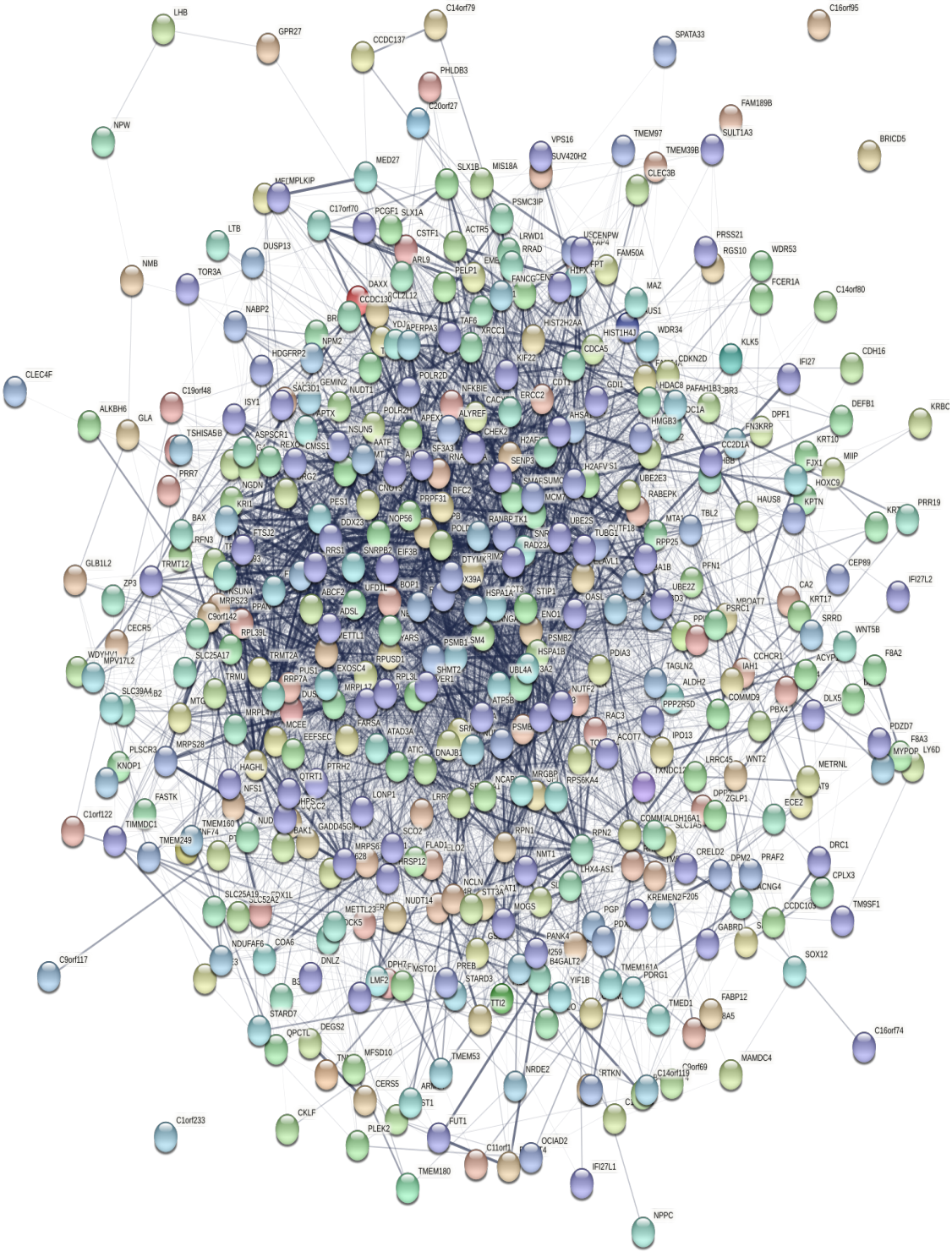


Figure 5-11: Network obtained from STRING tool for module (Yellow) from ESCC normal

Table 5.14: Potential biomarker list for ESCC with biological processes and literature evidences

Gene ID	Literature Evidence	Enriched GO Terms
IP6K3	Significantly associated with risk of renal cell carcinoma (RCC)	Protein phosphorylation, inositol phosphate biosynthetic process
EMP1	Highly associated with ESCC development [259], Downregulated mRNA in ESCC [286], tumour suppressor gene [287]	Regulation of the cell cycle and proliferation [288], multicellular organism development, cell growth
PIK3C2B	Associated with ESCC - tumour metastasis [289]	Phosphatidylinositol biosynthetic process, Akt signalling pathway [289]
FMO2	Down-regulated in ESCC, Associated with ESCC progression [290]	Organic acid metabolic process, toxin metabolic process
FREM2	Overexpressed in ESCC tissue samples, contributed to ESCC recurrence [291].	Cell adhesion, morphogenesis of an epithelium
AJAP1	It acts as a putative tumor suppressor in ESCC, a tumor biomarker to predict recurrence of ESCC after esophagectomy [292].	Cell adhesion

5.5 Conclusion

An effective co-expression/similarity measure called SNMRS has been introduced. A method has been presented to construct a co-expression network using SNMRS that handles all types of correlations followed by extraction of network modules from the network applying average linkage clustering algorithm. SNMRS is able to find highly similar patterns containing genes with high biological relevance. Experiments using real-world datasets show that proposed method is capable of extracting clusters that are much better than other similar methods on a variety of quality measures.

Identification of potential biomarkers from the ESCC single cell RNA sequencing dataset is another important task which can be performed using the differential expression analysis method. The next chapter presents a framework to identify potential biomarkers using an ensemble of differential expression analysis methods and SNMRS based module detection method on scRNA-seq data.