Journal of Signal and Image Processing
ISSN: 0976-8882 & E-ISSN: 0976-8890, Volume 3, Issue 2, 2012, pp.-85-87.
Available online at http://www.bioinfo.in/contents.php?id=48

INFORMATION THEORY TO FIND CO-EXPRESSED GENE NETWORK FOR MICROARRAY GENE EXPRESSION

DAMLE T.[1*] AND KSHIRSAGAR M.[2]

Department of Computer Technology, Yeshwantrao Chavan College of Enggineering, Nagpur- 441110, MS, India.
*Corresponding Author: Email- damle.tejashree@gmail.com

Received: March 15, 2012; Accepted: April 12, 2012

Abstract- Information theory is useful for finding the information content of the data which is referred to as Entropy. Microarray chip gives the data for gene expression. For finding co-expression between different genes generated from microarray, we have applied mutual information theory by finding entropy of each gene expression. This mutual information is converted into adjacency and dissimilarity matrix. Co-expressed networks are formed by applying cut- tree algorithm to dissimilarity matrix.
Our work finds the pair wise relatedness of different microarray genes. This relatedness finds the different gene networks. We used the diabetes Mellitus Type II as a disease model. This paper describes our approach of finding co-expressed gene network .
Keywords- Entropy, tree cut, co-expressed gene network.

Citation: Damle T. and Kshirsagar M. (2012) Information Theory to find Co-Expressed Gene Network for Microarray Gene Expression. Journal of Signal and Image Processing, ISSN: 0976-8882 & E-ISSN: 0976-8890, Volume 3, Issue 2, pp.-85-87.

Introduction

In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two random variables. The most common unit of measurement of mutual information is the bit, when logarithms to the base 2 are used. From Microarray chip thousand of gene expression produced simultaneously. Mutual Information theory calculates the pair wise relatedness of gene sets and then finds matrices. This information is useful for analyzing the dependency between gene set.
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. The average linkage clustering is a method of calculating distance between clusters in hierarchical cluster analysis. The linkage function specifying the distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster.
Our approach uses average linkage method for finding the clusters of correlated and similar genes. Co- expressed networks are formed from this clusters. For finding Co-expressed networks we used the gene set after elaborating Significance Analysis of Microarray (SAM) algorithm. This set contains total 238 genes.

Data Source

In this we take data from Mootha VK *et al.* (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes. *Nature Genetics*; Vol 34(3); 267-273. The disease model is Diabetes mellitus (Type II). The study involved 34 males, 17 with normal glucose tolerance (NGT), and 17 with Diabetes Mellitus (DM Type II). [7]

Process of finding Co-expressed Gene Network

Figure 1 shows the flowchart that explains our approach for finding co-expressed gene network. We used different R packages for finding gene network. The Co-expression networks are formed

separately for Diabetes and Normal microarray Data. The steps given below explains the process of finding Co-expressed gene network.

Step 1- mutual Info uses the binning approach. The distance between entropies of each gene set is calculated. The output of this step is mutual information matrix. Table 1 shows this matrix.

Table 1- Mutual Information Matrix

	1	2	3	4	5	6	7	8	9	10
1	0	0.022473	0.136505	0.097582	0.396008	0.145702	0.115864	0.132433	0.11823	0.039042
2	0.022473	0	0.223065	0.126655	0.118883	0.201483	0.155729	0.30054	0.072476	0.149662
3	0.136505	0.223065	0	0.303952	0.100601	0.183201	0.265689	0.169932	0.16652	0.133086
4	0.097582	0.126655	0.303952	0	0.097582	0.164266	0.246754	0.170985	0.151657	0.14835
5	0.396008	0.118883	0.100601	0.097582	0	0.227249	0.130727	0.087444	0.117177	0.126367
6	0.145702	0.201483	0.183201	0.164266	0.227249	0	0.131781	0.137558	0.083379	0.105726
7	0.115864	0.155729	0.265689	0.246754	0.130727	0.131781	0	0.20413	0.053541	0.07996
8	0.132433	0.30054	0.169932	0.170985	0.087444	0.137558	0.20413	0	0.206496	0.092457
9	0.11823	0.072476	0.16652	0.151657	0.117177	0.083379	0.053541	0.206496	0	0.078254
10	0.039042	0.149662	0.133086	0.14835	0.126367	0.105726	0.07996	0.092457	0.078254	0

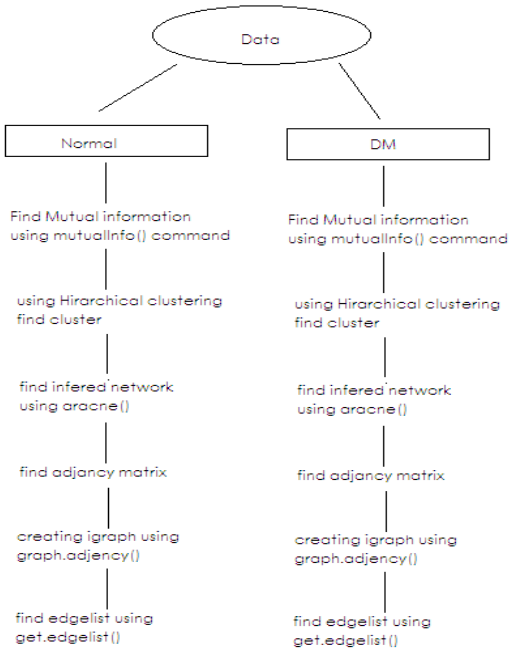

Fig. 1- Flowchart for finding Co-expressed Gene Network

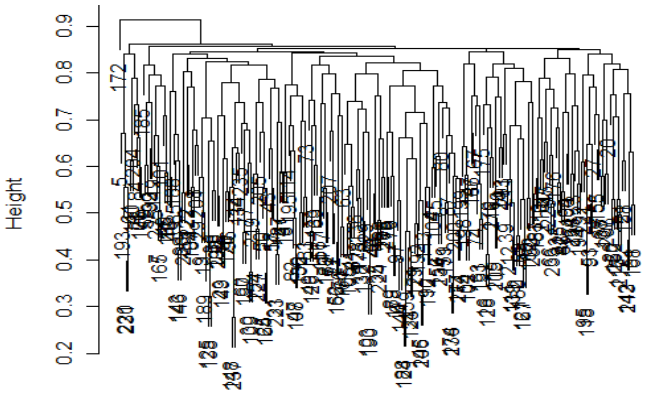

Fig. 2- Cluster Dendrogram for Normal Data

Step 2- Hierarchical clustering is takes place using hclust() method in R and average linkage analysis for the mutual information matrix. Figure 2 shows the cluster dendrogram. Figure 3 shows the dendrogram using cut tree method.

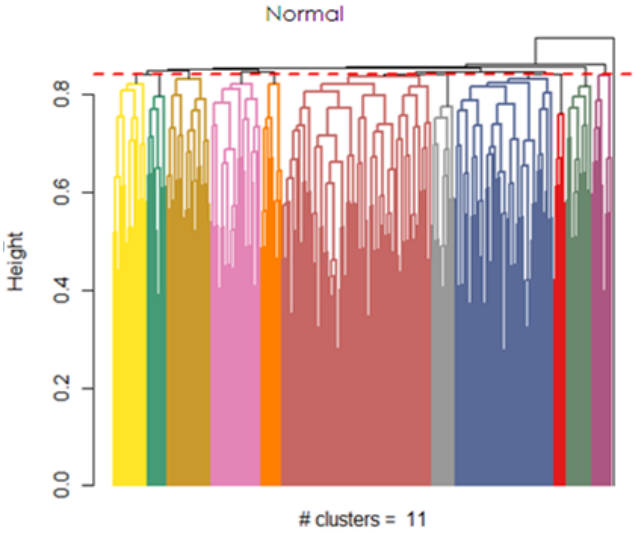

Fig. 3- Hierarchical clustering for adjacency matrix

Step 3- ARACNE is Algorithm for the Reconstruction of Accurate Cellular Networks which is a method for reconstructing biological networks from microarray data.[9, 10]

Step 4- The adjacency matrix formed is given in Table 2.

Table 2- adjacency matrix for Normal Data

		1	2	3	4	5	6	7	8	9	10
2	1	0	0	0	0	0.396008	0	0	0	0	0
3	2	0	0	0	0	0	0	0	0	0	0
4	3	0	0	0	0	0	0	0	0	0	0
5	4	0	0	0	0	0	0	0	0	0	0
6	5	0.396008	0	0	0	0	0	0	0	0	0
7	6	0	0	0	0	0	0	0	0	0	0
8	7	0	0	0	0	0	0	0	0	0	0
9	8	0	0	0	0	0	0	0	0	0	0
10	9	0	0	0	0	0	0	0	0	0	0

Step 5- igraph is plotted using graph.adjency() method and is shown in figure 4.

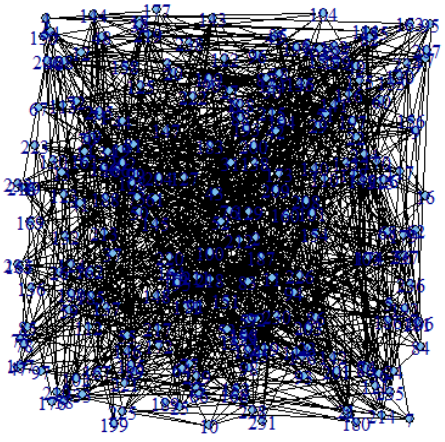

Fig. 4- Plot for Edge list

Step 6- Table 3 shows the edge list obtained after edge.getList()

Method. We obtained total 937 edges.

Table 3- Edge list

V1	V2
1	5
1	72
1	82
1	94
1	101
1	112
1	130
1	173
1	183
1	227
2	12
2	33
2	49
2	53
2	64
2	110
2	116
2	150
2	180
2	229
3	11
3	21
3	30

Figure 5 shows the Co-Expression network of normal data with cluster frequency 10,11 and 12.

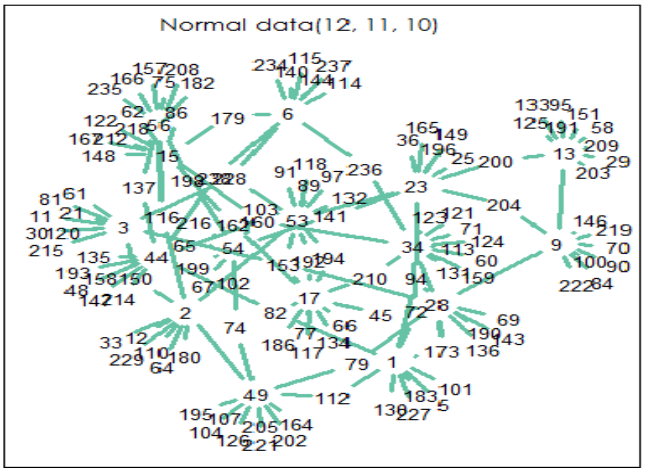

Fig. 5- Co-expression network for frequency 10,11,12

Conclusion

In our work Microarray gene set is used for finding Co-Expressed gene network. This set consists of 238 genes. We found the hub of genes which are related with each other. In our further analysis we will explore this data for finding regression analysis of data with different biochemical parameters.

References

[1] Tusher V.G., Tibshirani R. and Chu G. (2001) *PNAS*, 98 (9).

[2] Chu G., Li J., Balasubramanian Narasimhan, Tibshirani R. and Tusher V. (2002) *Department of Biochemistry*.

[3] Selvaraj S. and Natarajan J. (2011) *Bioinformation*, 6 (3), 95-99.

[4] Sivriver J., Habib N. and Friedman N. (2011) *An integrative clustering and modeling algorithm for dynamical gene expression data*.

[5] Langfelder P., Zhang B. and Horvath S. (2008) *Data Mining of Microarray Databases for the Analysis of Environmental Factors on Plants Using Cluster Analysis and Predictive Regression*, 24 (5), 719-720.

[6] Hovatta I., Saharinen J., Kimppa K., Laine M.M. and Antti Lehmussola (2005) *DNA microarray data analysis*.

[7] Mootha V.K., Lindgren C.M., Eriksson K.F. and Aravind Subramanian (2003) *Nature Genetics*, *Nature Genetics,* 34 (3), 267-273.

[8] *http://spotfire.tibco.com*.

[9] Meyer P.E., Lafitte F. and Bontempi G. (2008) *BMC Bioinformatics.*

[10] Margolin A.A., Nemenman I., Basso K, Wiggins C., Stolovitzky G., Favera R.D., Califano A. (2006) *BMC Bioinformatics*.