Reconstruction of Gene Co-expression Networks to Examine Yeast Microarray Data

Sabyasachi Patra[1] Dinesh Maharana[2]
[1,2]DST-FIST Bio-Informatics Laboratory
[1,2]Department of Computer Science and Engineering
[1,2]International Institute of Information Technology, Bhubaneswar, India

Abstract— A biological network such as Protein-protein interaction (PPI), Gene regulatory network, Gene Co-expression Network (GCN) which describes complex activities of genes, proteins, and products of them participate in different chemical reactions and also identifies the biological functionality. These networks constructed from different types of large biological datasets like DNA microarray data. In this paper, we have discussed the reconstruction of GCN with different Co-expression measures like Euclidean distance, Pearson's correlation coefficient (PCC) and Gene CO-expression Network called GeCON. We have discussed the behaviour of GeCON by comparing it with global similarities measures like Euclidean distance and correlation coefficient.

Key words: Biological network, GCN, PCC, EDM, GeCON, Microarray Data

I. INTRODUCTION

Cell forms constitute complex frameworks and can't be depicted utilizing an oversimplified version. To completely comprehend the working of cell procedures, it is insufficient to just assign capacities to individual cell molecules, proteins and genes. Cell forms constitute complex frameworks and can't be depicted utilizing an oversimplified version. To completely comprehend the working of cell procedures, it is insufficient to just assign capacities to individual cell molecules, proteins and genes. Biological system portraying co-operation among segments exhibit an incorporated at the dynamic conduct of the cell framework. The GCN is an undirected network, where every hub relationship among them [1]. Gene expression profiles of various genes for a few examples or test conditions, a GCN can be developed by searching for pair wise genes which demonstrate a comparative expression levels over all samples, therefore the profile of a pair of similar genes such as the gene expression values are parallel increase and decrease for both of them. Transcriptional regulation and protein complex controls co-expressed genes. And DNA microarray data is a single microarray experiment analyzes expression levels of a large number of genes under a test condition. Most reviews include different microarray tests covering a scope of conditions [3]. For instance, microarray studies frequently look at the expression levels of genes across various development stages, or in tissue specimen with and without a specific disease. Microarrays produce large volumes of data, making the requirement for advanced techniques for data mining, translation, and representation.is represented as gene and a couple of hubs are associated by an interaction if there is a significant

II. GENE CO-EXPRESSION NETWORKS

One well-known data structure that has come into the picture is the network. Networks offer a characteristic approach to model interactions between genes with hubs, is represents genes and edges are represents different interactions deduced from various information sources. Significant favourable circumstances to network data structure incorporate their strength to a lot of information. A GCN interfaces set genes that are altogether co-expressed crosswise over states of microarray data. All things considered, the initial phase in making a GCN is to score all sets of gene vectors. The second step is to pick a scoring edge and interface all genes matches whose scores surpass this esteem.

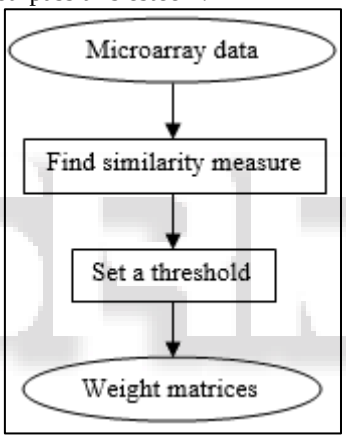

Fig. 1: classical model to construct GCN

Above flowchart describes the workflow for the construction of gene co-expression network and after getting the weight matrices we can easily visualize the co-expression network by using Cytoscape [7].

A. Euclidean Distance Measure

The Euclidean distance measure (EDM) finds the physical distance between two gene microarray expressions or their geometric separation, considering both the magnitude and the direction of the vectors. For instance, the EDM won't be powerful at identifying a transcription factor along with its objectives if overall levels of the transcription component are similarly lowered across tested conditions. Besides, if a pair of genes has smaller expression values, but they are haphazardly correlated then both of the genes may even now seem close in Euclidean space. However, the EDM can be helpful for recognizing genes in a common pathway that react uniquely from the rest of the genome [3]. For instance, if a cell is subjected to a natural anxiety, for example, heat shock, a sensational transcriptional increment may be normal for genes.

Let us consider two n-dimensional gene vectors *A* and *B*. And the Euclidian distance between *A* and *B* is ED_{AB}.

$$ED_{AB} = \sqrt{\sum_{j=1}^{n}(A_j - B_j)^2} \qquad (1)$$

Where 'n' is the number of conditions of gene expression.

B. Pearson's Correlation Coefficient

The PCC measures the propensity of two gene vectors to go up against genes above and down each of their normal levels in a planned manner. As the measure is with respect to the genes claim normal level, it is helpful for distinguishing likeness between genes that may have distinctive supreme levels of expression. For instance, PCC can distinguish situations where a transcriptional activator and its objective genes expression levels fall and rise in synchrony regardless of the possibility that the objectives have a more outrageous level of expression. Since the first experience with the field of microarray analysis [4], PCC has apparently been the most prominent strategy for the correlation of gene vectors.

The correlation *'p'* between two gene vectors *S* and *T* is calculated as:

$$p = \frac{n(\sum ST) - (\sum S)(\sum T)}{\sqrt{[n\sum S^2 - (\sum T)^2][n\sum S^2 - (\sum T)^2]}} \qquad (2)$$

Where *'n'* is the number of conditions of gene expression microarray data.

C. GeCON

To construct gene co-expression network GeCON is another algorithm in which we can identify both positive and negative similarity between two gene expressions from microarray data. An expression profile pattern can be identified by GeCON. The line joining two contiguous expression levels is considered as a regulation. Thus, for an expression profile there are '$(n-1)$' regulations are either up-regulation or down-regulation represented by 1 and -1 respectably for 'n' conditions. The J^{th} edge regulation value of a gene A_l, $A_l(p_j)$, based on two successive gene conditions C_{j-1} and C_j can be determined as.

$$A_l(P_j) = \begin{cases} 1 \; if \; C_{j-1} < C_j \\ -1 \; if \; C_{j-1} > C_j \end{cases} \qquad (3)$$

Then it finds the degree of fluctuation between two gene expression level at same condition or time point and calculates positive similarity(*Pos_sim*) if they have same regulation else they have opposite regulation so it gives negative similarity (*Neg_sim*)[3]. Where tanarc() find the angle of deflection from condition to present condition.

$$A_l(D_j) = \begin{cases} 180 - abs\left(tanarc(C_j, C_{j-1})\right) \; if \; C_j < C_{j-1} \\ abs\left(tanarc\left(C_j, C_{j-1}\right)\right) \qquad else \end{cases} \qquad (4)$$

Here $A_l(d_j)$ is angle at j^{th} condition of A_l gene

$$Pos\_sim(A_{lk}, A_{mk}) = \begin{cases} 1, if \; A_l(p_j) = A_m(p_j) \; and \; A_l(d_j) = A_m(d_j) < \mu \\ 0, \; otherwise \end{cases} \qquad (5)$$

$$Neg\_sim(A_{lk}, A_{mk}) = \begin{cases} 1, if \; A_l(p_j) = -A_m(p_j) \; and \; |180 - A_l(d_j) + A_m(d_j)| < \mu \\ 0, \; otherwise \end{cases} \qquad (6)$$

Here A_{lk} and A_{mk} are two different genes at same condition 'k'. Now we can find how many regulations are positively and negatively similar and then calculate positive support and negative support by finding average similarity. If the support is greater than some threshold, where θ = 0.5 and another threshold μ = 25[4].

III. RESULT AND DISCUSSION

The association between gene pairs in a network, with global similarity measures such as EDM and PCC measure where they identify transcriptional factor and non-linear statistical relationships or in the same pathway which responds differently from other genomes[3]. The most straightforward threshold technique is to pick a score cut-off. This strategy can be valuable when a clear elucidation of the scoring plan is accessible. For instance, a few reviews utilize a PCC cut-off of 0.50 as a threshold for average correlation, and 0.80 for good correlation [6]. In case of lesser Euclidean distance the distance between genes is higher similar among them. And here Euclidean space is ranging from 0 to 20. To compare with other algorithm we drop down to 0 to 1 and reverse it by subtracting 1 from each value. Gene co-expression network construction for Yeast microarray database (http://faculty.washington.edu/kayee/cluster/) so we get interactions generated by PCC is 2250 numbers and EDM is 2491 numbers and also we analyze with GeCON In order to find co-expression between two gene expression profiles in terms of degree of deflection and reach good results with is from 15 to 25 and θ is 0.5 and we get 2129 number of interactions [5].

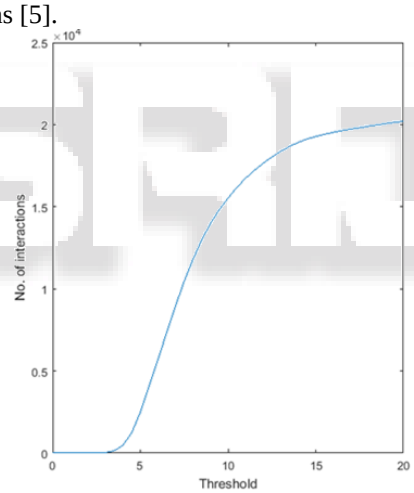

Fig. 2: Number of gene-gene interactions at different threshold of EDM

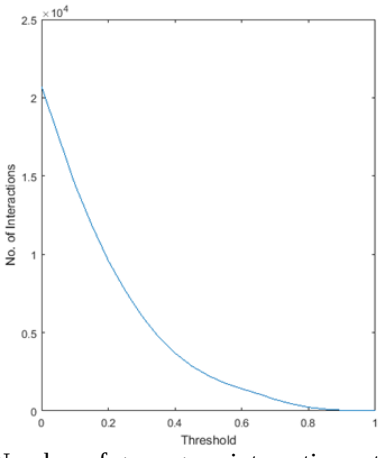

Fig. 3: Number of gene-gene interactions at different threshold of PCC.

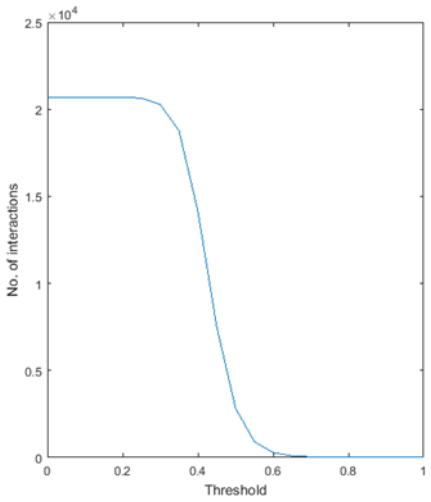

Fig. 4: Number of gene-gene interactions at different threshold ( ) of GeCON.

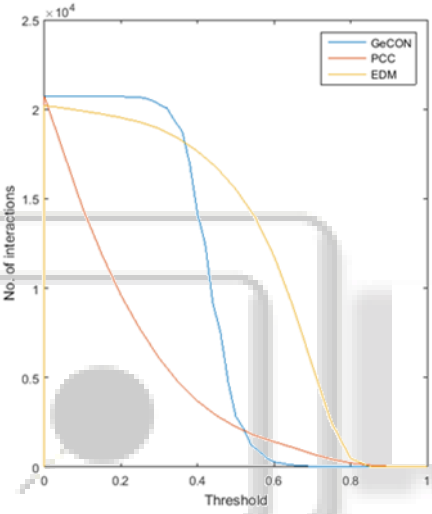

Fig. 5: Comparison of EDM, PCC with GeCON according to the number of interaction.

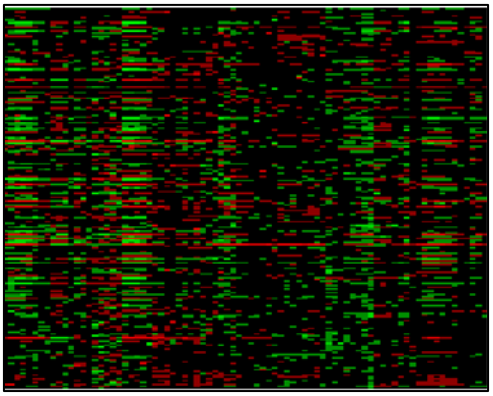

Fig. 6: Heatmap for Yeast microarray database in which rows are genes and columns are conditions.

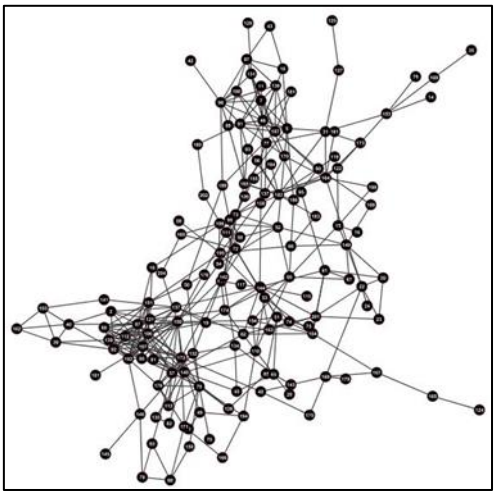

Fig. 7: Visualization of GCN for Yeast microarray data using Cytoscape.

IV. Conclusion

In this work, we perform an analysis of comparing co-expression networks from GeCON with other famous co-expression networks algorithms i.e. PCC measure and EDM based. From total interactions of GeCON 30.6% matches with PCC measure co-expression network and 34.4% matches with EDM co-expression network. This concludes that GeCON co-expression network carries some property of EDM as well as PCC measure co-expression network. We reviewed that gene expression profiles are shared local similar expression rather global similar expression because we observed that in GeCON co-expression network and find local co-expression patterns between gene expression profiles.

Acknowledgement

The authors acknowledge the support by DST-FIST Bioinformatics Lab under FIST project, Govt. of India, the Department of Computer Science and Engineering, IIIT, Bhubaneswar for providing all the computational resources and facilities for carrying out the research work.

References

[1] Stuart, Joshua M., Eran Segal, Daphne Koller, and Stuart K. Kim. "A gene-coexpression network for global discovery of conserved genetic modules." science 302, no. 5643 (2003): 249-255.

[2] Vandenbon, Alexis, Viet H. Dinh, Norihisa Mikami, Yohko Kitagawa, Shunsuke Teraguchi, Naganari Ohkura, and Shimon Sakaguchi. "Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system." Proceedings of the National Academy of Sciences 113, no. 17 (2016): E2393-E2402.

[3] Weirauch, Matthew T. "Gene coexpression networks for the analysis of DNA microarray data." Applied statistics for network biology: methods in systems biology (2011): 215-250.

[4] Berriz, Gabriel F., Oliver D. King, Barbara Bryant, Chris Sander, and Frederick P. Roth. "Characterizing gene sets with FuncAssociate." Bioinformatics 19, no. 18 (2003): 2502-2504.
[5] Roy, Swarup, Dhruba K. Bhattacharyya, and Jugal K. Kalita. "Reconstruction of gene co-expression network from microarray data using local expression patterns." BMC bioinformatics 15, no. 7 (2014): S10.
[6] Roy, S., and D. K. Bhattacharyya. "Reconstruction of genetic networks in yeast using support based approach." In Trendz in Information Sciences & Computing (TISC), 2010, pp. 116-121. IEEE, 2010.
[7] Margolin, Adam A., Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC bioinformatics 7, no. 1 (2006): S7.
[8] van Verk, Marcel C., John F. Bol, and Huub JM Linthorst. "Prospecting for genes involved in transcriptional regulation of plant defenses, a bioinformatics approach." BMC plant biology 11, no. 1 (2011): 88.

