20-67

Final report of the project entitled

BEYOND CLUSTERING GENE EXPRESSION DATA

Department of Computer Science and Engineering
Tezpur University

Principal Investigator

Prof. Dhruba Kr. Bhattacharyya

Co investigators

Prof. J. K. Kalita
Prof. Malay Dutta
Dr. Rosy Sharmah
Dr. SauravjyotiSarmah

1) PROJECT PROFILE

Title of the project:	**Beyond Clustering gene expression data**
Name of the implementing institute:	**Tezpur University, Napaam, India**
Principal Investigator(PI):	**Prof. Dhruba Kr. Bhattacharyya**
Co-Investigators(CIs):	**Prof. J. K. Kalita** **Prof. Malay Dutta** **Dr. Rosy Sharmah** **Dr. SauravjyotiSarmah**
Research Associate:	**PriyakshiMahanta**
Project Duration :	**36 months**

Amount Received:

Sl. No.	Budget Head	Amount (in Rs.)
(i)	Project Manpower	2,88,000
(ii)	TA/DA Accommodation Expenses	75,000
(iii)	Permanent Equipment	3,85,000
(iv)	Contingency/Consumables	90,000
(v)	Institutional Overheads	1,50,000
	Total	9,88,000

Principal Investigator

(Dr. Dhruba Kr. Bhattacharyya)

2. WORK DONE

The work accomplished under the project can be categorized into four different category as given below:

A. Gene expression data analysis

DNA microarray technology has revolutionized biological and medical research by enabling biologists to measure expression levels of thousands of genes in a single experiment. Different computational techniques have been proposed to extract important biological information from the massive amount of gene expression data generated by DNA microarray technology. A first step toward addressing this challenge is the use of clustering[2,3,6,10,17] techniques, which is essential in the data mining process to reveal natural structures and identify interesting patterns in the underlying data. Analysis is done using the following techniques:

a) Discretization techniques[7] are widely used as preprocessing task in different classification techniques especially in the area of machine learning. These techniques have also been used as a preprocessing task for computational construction of regulatory networks in gene expression data analysis.
b) The existence of various types of correlations among the expressions of a group of biologically significant genesposes challenges in developing effective methods of gene expression data analysis. The initial focus of computational biologistswas to work with only absolute and shifting correlations. However, researchers have found that the ability to handle shifting-and-scaling correlation[11] enables them to extract more biologically relevant and interesting patterns from gene microarray data.
c) Triclustering techniques[1] extract genes that have similar expression patterns in a set of samples across a set of time points. A challenge in triclustering is to account for both inter-temporal and intra-temporal gene coherence. Other challenges include avoidance of time-dominated and sample-dominated results and detection of time latent triclusters.

B. Network module extraction from gene co-expression network

The development of high-throughput microarray technologies has provided various opportunities to systematically characterize diverse types of computational biological networks. Co-expression network[4,5,8,15,16] have become popular in the analysis of microarray data, such as for detecting functional gene modules.

C. Feature selection

Feature selection[9,14] is a multi-step process particularlyuseful in analyzing high-dimensional data from many real-life applications. The filter approaches are typically faster while wrapper approaches are more reliable though computationally expensive. Feature selection techniques often strive to achieve performance similar to wrapper approaches employing various computational approaches. Feature selection techniques typically depend on ways how they compute feature–feature correlation and feature–class

correlation[13]. These two computations are highly governed by the correlation measure being used. Selection of an optimal subset of features that correctly distinguishes objects with different classes in thelearning process is a major issue in knowledge discovery.

D. Protein complex detection in PPI network

With the growing development in the field of genetics and gene expression, biologists have been compelled to think as to what is the driving force that brings about variations in biological activities among the organisms. This variation is basically a result of biological phenomenon, viz., expression of genes and encoding of proteins by the expressed genes. Regulatory relationships among genes also play a crucial role in this process. To have further insight into this entire pathway, it is very important to understand the functional characteristics of the proteins encoded by various expressed genes. Different functional properties exhibited by the proteins are dependent on the interactions that take place among them. There are a number of computational methods to identify such groups of proteins that work together to perform specific functions. Protein complexes[12,18] play key role in many biological processes and hence considered to be an important task in post-genomic era.

LIST OF PUBLICATION

1. H. A. Ahmed, P.Mahanta, D. K. Bhattacharyya, J. K. Kalita, and A. Ghosh "Intersected coexpressedsubcube miner: An effective triclustering algorithm." *Information and Communication Technologies (WICT), 2011 World Congress on*. IEEE, 2011.

2. H. A. Ahmed, P.Mahanta, D. K. Bhattacharyya, and J. K. Kalita "Gerc: tree based clustering for gene expression data." *Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on*. IEEE, 2011.
3. H. A. Ahmed, P.Mahanta, D. K. Bhattacharyya, and J. K. Kalita "Autotuned Multilevel Clustering of Gene Expression Data." *American Journal of Bioinformatics Research* 2.5 (2012): 68-78.
4. H. A. Ahmed, P.Mahanta, D. K. Bhattacharyya, and J. K. Kalita "Module extraction from subspace co-expression networks." *Network Modeling Analysis in Health Informatics and Bioinformatics* 1.4 (2012): 183-195.
5. P.Mahanta, Hasin A Ahmed,Dhruba K Bhattacharyya,Jugal K Kalita, "An effective method for network module extraction from microarray data", in BMC Bioinformatics, 13(Suppl 13):S4, 2012 (http://www.biomedcentral.com/1471-2105/13/S13/S4)
6. H. A. Ahmed, P.Mahanta, and D. K. Bhattacharyya. "Finding gene coherent patterns using PATSUB+." *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*. ACM, 2012.
7. P.Mahanta, H. A. Ahmed, D K Bhattacharyya and J K Kalita, "Discretization in Gene Expression Data Analysis: A Survey", in *the ACM Proc of CCSEIT 2012, pp 69-75, 2012 [DOI>10.1145/2393216.2393229].*
8. H. A. Ahmed, P.Mahanta, and D. K. Bhattacharyya. "Negative Correlation Aided Network Module Extraction." *Procedia Technology* 6 (2012): 658-665.
9. P. Borah, H. A. Ahmed, and D. K. Bhattacharyya. "A statistical feature selection technique." *Network Modeling Analysis in Health Informatics and Bioinformatics* 3.1 (2014): 1-13.

10. A. Goyal, H. A. Ahmed, and D. K. Bhattacharyya. "PNCSIM: An Effective Measure to identify Gene Coexpressed Patterns" *Proceedings of the fifth international joint conference on CNC and CCPE*, ELSEVIER, 2014: 155-162.
11. H. A. Ahmed,, P.Mahanta, D. K. Bhattacharyya, and J. K. Kalita "Shifting-and-scaling Correlation based Biclustering Algorithm", accepted for publication in *IEEE/ACM Transactions on Computational Biology and Bioinformatics.*
12. Sharma, Pooja, Hasin A. Ahmed, and Dhruba K. Bhattacharyya "Unsupervised Complex Finding from PPI Networks", *communicated toBioData Mining, Springer.*
13. Ahmed, H. A., D. K. Bhattacharyya, and J. K. Kalita "Strew Index: An effective feature class correlation measure", *communicated to Computational Statistics and Data Analysis, Elsevier.*
14. Ahmed, H. A., D. K. Bhattacharyya, and J. K. Kalita "Feature Selection : Approaches, Methods and Tools", *communicated to ACM Computing Surveys.*
15. S Deb, P Mahanta, D K Bhattacharyya and M Dutta, "Subspace Module Extraction from MI-Based Co-expression Network" *submitted to International Journal of Bioinformatics Research and Applications (IJBRA)*
16. P Mahanta, H A Ahmed, D K Bhattacharyya and A Ghosh, "FUMET: A Fuzzy Network Module Extraction Technique for Gene Expression Data" Journal of Bioscience, vol 39, no 2, June, 2014, Springer.
17. R. C.Baishya, R.Sarmah, D. K. Bhattacharyya, M. A. Dutta,"*A Similarity Measure for Clustering Gene Expression Data"Lecture Notes in Computer Science Volume 8321, 2014, pp 245-256*
18. P Mahanta, D K Bhattacharyya and A Ghosh " PDComp: An Effective PPI complex Finding Method", communicated to Computational Biology and Chemistry, Elsevier