Name of Scholar: **Khalid Raza**

Name of Supervisor: Dr. Mansaf Alam

Name of Co-supervisor: Dr. Rafat Parveen

Department: Computer Science, Jamia Millia Islamia, New Delhi

Title of PhD Thesis: **Soft Computing Approach for Modeling Biological Networks**

Abstract

With the rapid advancement in high-throughput techniques for the measurement of biological data, the attention of the research community has shifted from a reductionist view to a more complex understanding of biological system. The enriched understanding about genomes of various organisms, together with advancement in microarray technology, has fuelled the researchers for the development of computational and mathematical model of biological networks. The major objectives for modeling biological networks are: to present a synthetic view of the available biological knowledge in the form of network to better understand interactions and relationships on a holistic level; allow researchers to make predictions about gene function that can then be tested at the bench; allow study of network's dynamical behavior; complexity of molecular and cellular interactions requires modeling tools that can be used to design and interpret biological experiments; essential for understanding the cell behavior, which in-turn leads to better diagnosis, predicts interactions between biological macromolecules not known so far; and also allows for drug effect simulations.

The discovery of biological pathways or regulatory networks leads to a wide range of applications, such as pathways related to a disease can unveil in what way the disease acts and provide novel tentative drug targets. In addition, the development of biological models from discovered networks or pathways can help to predict the responses to disease and can be much useful for the novel drug development and treatments. The high-throughput measurement techniques and plenty of data produced have brought the hope that we would be able to discover entire regulatory networks from these data. Unfortunately, the data and the biological systems that produce the data are often noisy and also biological processes are not well understood. Due to promising results of soft computing techniques, it has gradually opening up several opportunities in bioinformatics by producing low-cost, low-precision (approximate) and better solutions.

We constructed colon-cancer network using one of the popular information theoretic approach called mutual information. The proposed algorithm demonstrates how we can reveal novel gene regulatory interactions in case of cancer. We constructed ten different networks by varying the number of interactions ranging from 30 to 500.

The identified signature in the first network captures regulated interactions among 22 differentially expressed genes (DEGs). In case of tenth network consisting of 500 interactions, it shows regulated interaction among 79 DEGs. Most of the identified DEGs have been validated and found to participate in colon cancer but interaction between them needs further biological validation for its reliability. For the validity of information theoretic based proposed algorithm, we used simulated benchmark network of yeast taken from DREAM3 challenge and AUROC is found to be approximately 0.93, which is better than other techniques such as LP, LASSO and PCA-PCC.

We proposed another technique for modeling cancer-specific network using popularly used statistical approaches – Pearson correlation coefficient, t-test and fold-change. Here, genes relevant to a specific cancer have been identified using a two-stage filtering technique and then pair-wise correlation among gene-pairs were calculated. The technique was applied on prostate cancer data and three network modules have been identified. We validated obtained results with biological databases and literature. We also performed GO-based enrichment analysis for all the extracted modules. Although statistical approaches are simple and performing better in our case but unfortunately these techniques are directly affected by noises in the data which are prevalent in microarray.

Further, we proposed a recurrent neural network (RNN) based GRN model hybridized with extended Kalman filter for weight update in backpropagation through time training algorithm. The RNN is a complex neural network that gives a better settlement between the biological closeness and mathematical flexibility to model GRN. The advantage of RNN is that it is able to capture complex, non-linear and dynamic relationship among variables. Since, GRN is so complex and relationships between genes are highly non-linear and dynamic in nature so it motivated us to develop RNN-based model of GRN. Gene expression data are inherently noisy and Kalman filter performs well for estimation even in noisy data. Hence, non-linear version of Kalman filter, i.e., extended Kalman filter has been applied for weight update during network training. The proposed model has been tested on four benchmark datasets, two real and two simulated networks, and results show that model performs well on all four datasets. We also compared our results with other state-of-the-art techniques that indicate the superiority of our proposed hybrid model. To test the robustness of the model, we added 5% Gaussian noise in the dataset and performance of the prediction is assessed. The result shows the added noise has negligible effect on the accuracy of predicted results.

Finally, topological properties of the inferred networks of colon cancer as well as prostate cancer have been analyzed. Our observation shows that inferred network modules are scale-free, modular, having few highly connected genes and mostly sparsely connected genes. These network modules also hold ‘small world’ property.

---: :---