Inference of Gene Regulatory Networks for Prostate Cancer using Bayesian Networks with Feedback and Feed Forward Loops

Nimrita Koul, Sunil kumar S Manvi

Abstract: . The solution to any problem depends on the depth of our understanding of it. Cancer is a disease that is being investigated at multiple levels and from multiple perspectives to understand the details of its origins and expansions in order to be able to figure a cure for it. We can now computationally analyze the biological data produced by genome analysis techniques like genomics, proteomics, and transcriptomics. DNA micro array technology has made available large gene expression datasets for entire genomes. It has been clinically observed that inside a human cell, activity of a gene often turns on or turns off one or more other genes. Such relationships in the co-regulation of genes is captured by gene regulatory network models which are computationally constructed from gene expression datasets. It has been observed that healthy and diseased states of a human cell show different regulatory interrelations between genes. In this paper, we have proposed to use a stochastic approach called Bayesian Networks with Feedback and Feedforward loops for inference of inter dependence in the regulation of genes in case of Prostate Cancer. It was observed that 4 of the networks revealed by the proposed approach matched the ones observed in clinical studies.

Index Terms: Bayesian Networks , Computational Genomics, Gene Expression Data, Gene Regulatory Network, Reverse Engineering

I. INTRODUCTION

The expression of our genes controls the proteins that are manufactured in our body and ultimately the working of it. Availability of biological data from the technologies like DNA Microarray, DNA sequencers etc. has provided us with comprehensive data about the working of human cell. But this data is compartmentalized and needs to be analyzed by suitable system biological and computational methods to derive meaningful insights [1] about the state of health of a human being from this data. Analysis of the biological data using computational techniques can help a scientist to integrate the insights from various types of input data and come upon a diagnosis or prognosis about a subject.

The process of gene expression involves three steps – combination of transcription of DNA sequences, processing of the primary RNA transcripts, and translation of the mature messenger RNA (mRNA) to proteins in ribosomes[1]. From the repositories of gene expression data available to us we can computationally infer the dependency or co-regulation relation between some genes in a cell. It is a well-known fact from clinical research and systems biology that many genes in our cells work in tandem and are regulated by one another through the output of expression of one gene the other gene is turned on. A Gene Regulatory Network (GRN) [2] is a set of genes, their product molecules and chemicals produced by interactions of the products of genes that affect the expression of other genes located in the DNA of a human cell.

The state of these regulatory networks in our genes and the regulatory networks that get activated in response to an environmental stimulus can be a strong marker of the state of our health. Fig 1. Shows the process of gene regulation in a cell.

Rest of this paper is organized as follows –Section II contains a brief walk through the back ground and related works, Section III contains the introduction into various methods of computational construction of gene regulatory networks, Section IV describes the proposed method, Section V discusses the results obtained and analysis thereof, Section VI concludes the paper and is followed by Acknowledgements and References.

Revised Manuscript Received on April 25, 2019.

Nimrita Koul, School of Computing & Information Technology, REVA University, Bangalore, India.

Dr. Sunilkumar S Manvi, Professor, School of Computing & Information Technology, REVA University, Bangalore, India.

The authors thankfully acknowledge the financial grant received from Department of Science & Technology (DST), Government of India, under the ICPS Scheme 2019

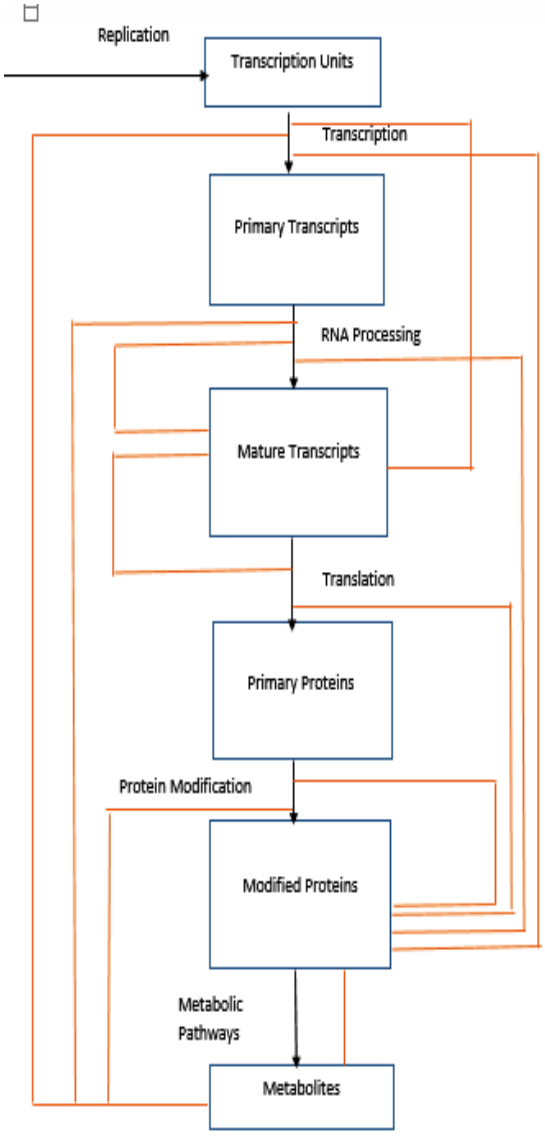

Fig.1 Interplay of Genetic Regulation Boxes represent cellular molecules, Black Arrows represent Molecular Processes, and red lines represent feedback loops

II. RELATED WORK

While inferring gene regulatory networks we use the concept of reverse engineering in order to estimate the biochemical reactions and interactions inside a cell from the gene expression data. From these interactions we can identify the genes responsible for down or up regulation of other genes. Identification of such interactions with the cells of an organism can help establish the knowledge about the state of health of such individual.

In the literature, the GRN construction methods are categorized as model based and non-model based, supervised and semi-supervised, those based on observational data and those based on perturbation data. [3]

Supervised techniques like Weighted Gene Correlation Network (WGCNA) [4] Context Likelihood of Relatedness (CLR) algorithm[5], Trustful Inference of Gene Regulation using Stability Selection (TIGRESS)[6] are used in literature to predict GRNS from gene expression data. In [6] the authors have presented a comparison of supervised and unsupervised methods. Supervised methods are proven to work better however the performance of unsupervised methods can be improved by using knockout. [7,11] have shown that prediction accuracy of Gaussian SVM Kernel methods on a 200 nodes graph is highest while CLR method gives highest accuracy for a graph with more than 500 nodes. Basic modeling techniques used for GRN construction include –Bayesian Networks, Logical Networks, Neural Networks , Space Models , Relevance Networks and Differential Equations[6].

III. PROPOSED METHOD

In a cell, if gene X is regulated in its expression by the some proteins say, Y and Z the we can model the expression of gene X as a stochastic function of activity of proteins Y and Z. The relation between the regulators and the regulated gene has to be modelled in terms of probabilistic models[7] due to variability and noise in the recording of biological data. A Bayesian approach can be used to model the regulatory influences in genetic pathways by using the expression levels of genes as a measure of activity level of proteins encoded by the corresponding genes. However, certain protein modifications after transcription can activate or silence certain genes.

In the proposed scheme, we have used Bayesian Networks [7]to identify regulatory relationships between genes. This involves identification of a Bayesian Network that gives as output an array ‘Ax’ of regulatory genes for each input gene ‘X’, array Ax contains all the genes whose expression is influenced by expression of gene X. The influential relations are identified in form of a network architecture by applying Structural learning to Ax. We have considered 10 different architectures and compared their likelihood of generating the observations of gene expression levels. The high dimensionality and low sample size which is characteristic of gene expression datasets poses a major problem in this method of inferring the regulatory influences. We have employed bootstrap algorithm to make the scheme resistant to perturbations in the observation. We aim at inferring both direct and indirect influences in regulation.

The data set we used is the prostate cancer data set [5]. This data set contains 6033 genes and 102 samples with two classes. 52 samples are of prostate cancer and 50 are normal.

The expression values are normalised and thresholded. The genes with more than five- fold variability have been used to avoid redundancy.

We begin with the assumption of no inter-regulation among gene expressions and with no directionality in the regulations. The inferred networks are assigned with weights based on their confidence levels. The sub networks with higher weights are chosen from all the networks inferred and the output of the scheme is the sub networks in which minimum number of genes regulate the expression of

maximum number of other genes. Such genes are most influential regulators. At this stage, the feedback and feedforward loops [8] are used to identify next level of regulators i.e. regulators of regulators and so on. Thus, giving us a list of coordinated genes. While we are trying to figure out the co-regulation in proteins produced by genes but the input type being gene expression data sets allows us to indirectly find these relations between proteins by figuring out the coordinated changes in expression values of genes responsible for production of corresponding proteins. i.e. coordinated expression values hint at coordinated protein production in a cell. Fig2. Shows a building unit of a probabilistic network. Fig.3 shows a sample Bayesian Network showing dependencies among 5 sample genes. Table 1 presents a brief comparison of the network based methods for GRN construction.

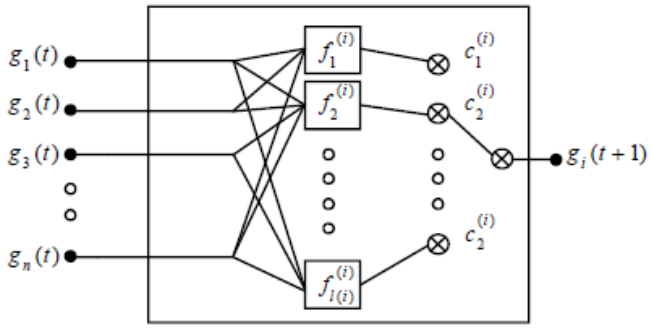

Fig. 2 – A Fundamental Unit of a Probabilistic Network

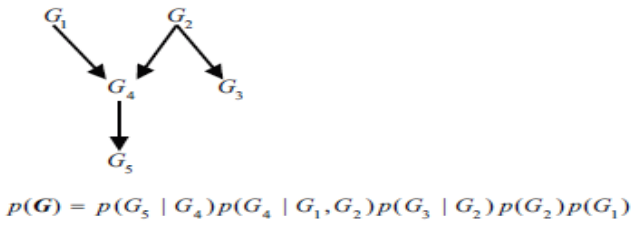

Fig 3. A sample Bayesian Network showing the conditional dependencies among nodes.

	Dynamism	Stochasticity	Discrete State Space
Boolean Networks	Yes	No	Yes
Probabilistic Boolean Networks	Yes	Yes	Yes
Bayesian Networks	No	Yes	No
Deep Belief Networks	Yes	Yes	Yes
Recurrent Neural Networks	Yes	No	No
Differential Equations	Yes	Yes	No
Relevance Networks	No	No	Yes

Table 1. A Brief Comparison of Network based methods for construction of GRNs.

IV. RESULTS AND DISCUSSION

Fig. 4 below shows one of the regulatory network obtained for the gene ABC2 which has been clinically [9] shown to be upregulated for prostate cancer.

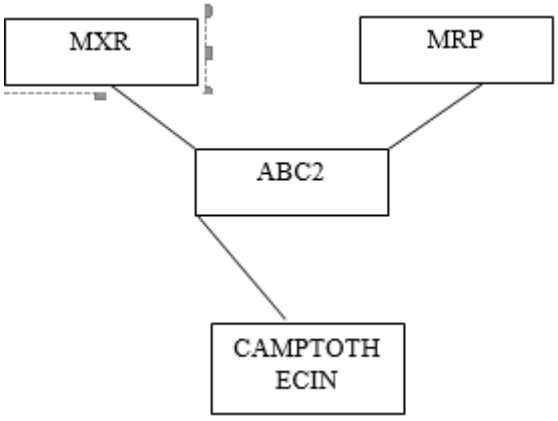

Fig. 5 below shows the associations between ABC2 and other cellular compounds

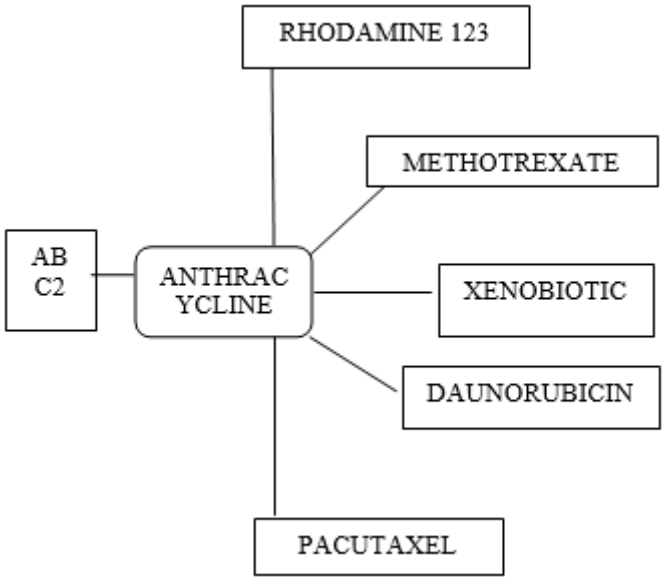

Fig 5. A Regulatory Network produced by the proposed method showing regulatory relation between the gene ABC2 and the cellular compounds.

Fig. 6 shows the representation of markers of Prostate Cancer in the form of a heat map

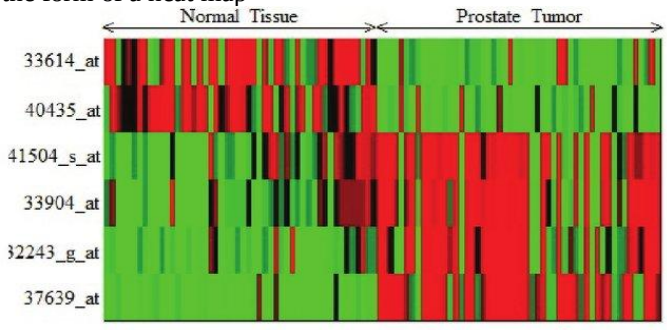

Fig. 6 – Heat Map of Prostate Cancer Gene Markers

Fig. 7 shows the signaling pathways in Prostate Cancer

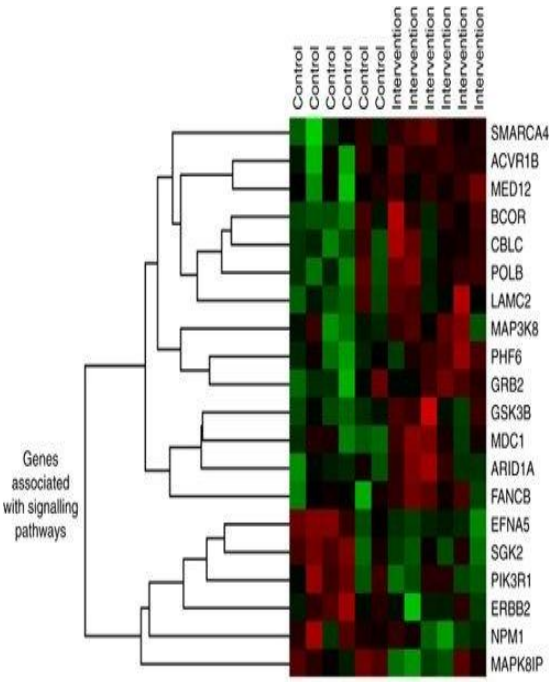

Fig. 7 Signaling Pathways in Prostate Cancer

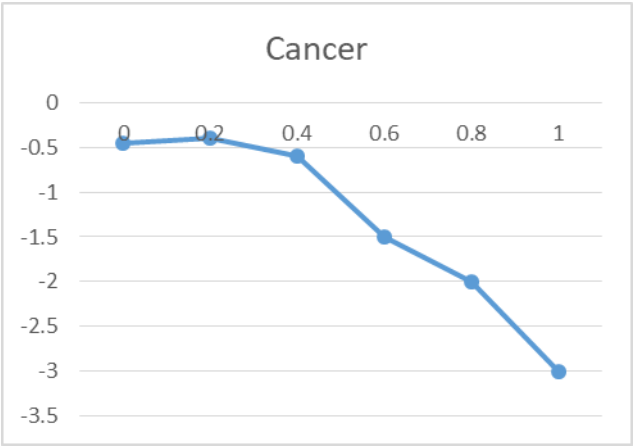

Fig. 8 Log(K) plot of the networks for Cancer samples

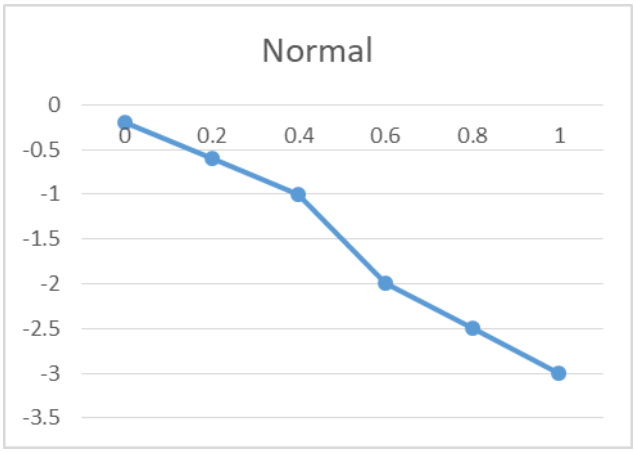

Fig. 9. Log(K) plot of Normal Samples

Fig. 8 and Fig. 9 show the distributions of the gene networks of cancer and normal gene networks on a log scale. On X axis we have the log of k-links in the gene network and y axis is log of degree disribution. We can see that the degree of distribution is less steep in case of cancer samples.

V. CONCLUSION AND FUTURE SCOPE

In conclusion, we can say that correct inference of gene regulatory networks will sure help in understanding the state of health of a cell and ultimately that of an individual. It will reveal molecular interactions that can cause diseases like cancer. We have worked with only the observational gene expression data in this paper, in future work we can include datasets from proteomics [10] data in order to cross verify the relations obtained from gene expression data network analysis.

ACKNOWLEDGEMENT

The authors thankfully acknowledge the financial grant received from Department of Science & Technology (DST), Government of India, under the ICPS Scheme 2019

REFERENCES

1. T. R. Golub, D. K. Slonim1, P. Tamayo, C. Huard, M. Gaasenbeek J.P.Mesirov, H.Coller, M.L. Loh, and J. R. Downi, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531–538, 1999.
2. Maetschke,S.R.,Madhamshettiwar,P.B.,Davis,M.J.,andRagan,M.A.(2014). "Supervised, semi-supervised and unsupervised inference of gene regulatory networks", *Brief.Bioinformatics*, 15,195–211.doi:10.1093/bib/bbt034
3. Marbach,D.,Costello,J.C.,Kuffner,R.,Vega,N.M.,Prill,R.J., Camacho,D. M.,etal.(2012), "Wisdom of crowds for robust gene network inference. *Nat. Methods"* 9,796–804.doi:10.1038/nmeth.2016
4. Yaghoobi, *et al.*: " GRN modeling techniques", *Journal of Medical Signals & Sensors*, Vol 2 , Issue 1, Jan-Mar 2012.
5. S. D, F. P, R. K, K. D, Manola, L. C, and T. P, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203–209, 2002.
6. J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Transactions*

*on Computational Biology and bioinformatic*s, vol. 13, no. 5, pp. 971–989, 2016.
7. Nir Friedman, "Inferring Cellular Networks Using Probabilistic Graphical Models", *Science*, Vol 303 6 Feb, 2004 803
8. Penfold,C.A.,and Wild,D.L.(2011), "How to infer gene networks from expression profiles, revisited", I*nterface Focus* 1,857–870.doi:10.1098/rsfs.2011.0053
9. Mordelet, F., and Vert,J.-P.(2008), "SIRENE: supervised inference of regulatory networks", *Bioinformatics* 24,i76–i82.doi:10.1093/bioinformatics/btn273
10 . Omranian, N.,Eloundou- Mbebi, J.M.,Mueller-Roeber, B.,and Nikoloski,Z. (2016), "Gene regulatory network inference using fused LASSO on multiple data sets", *Sci.Rep.* 6:20533.doi:10.1038/srep20533
11 . Patel, N., and Wang,J.T .(2015), "Semi-supervised prediction of gene regulatory networks using machine learning algorithms", *J.Biosci.* 40,731–740.doi:10.1007/s12038-015-9558-9
12. Y. Lin, H. Yeh, S. Cheng and V. Soo, "Comparing Cancer and Normal Gene Regulatory Networks Based on Microarray Data and Transcription Factor Analysis," *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, Boston, MA, 2007, pp. 151-157. doi: 10.1109/BIBE.2007.4375558

AUTHORS PROFILE

Nimrita Koul is a B.Tech and M.Tech in Computer Science & Engineering and pursuing PhD in Computer Science from REVA University.

Dr. Sunilkumar S Manvi is Professor in the School of Computing & Information Technology, REVA University. Dr. Manvi is an eminent researcher and academician with more than 300 publications in International and National Journals