May-June 2020
ISSN: 0193-4120 Page No. 3863-3865

Algorithm for Inference of Gene Regulatory Networks in Breast Cancer

[1]Nimrita Koul, [2]Sunilkumar S Manvi

[1]Research Scholar, School of CIT, REVA University
[2]Professor, School of CIT, REVA University

Article Info
Volume 83
Page Number: 3863-3865
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract
Interactions among genes regulate the physiology of a human cell. The influence that genes exercise on activation or deactivation of other genes is actuated through signaling pathways which involve synthesis of many molecular compounds. Such a group of genes which have a directed relation of either activation or deactivation on one another is known as a gene regulatory network. Computational understanding of regulatory networks by analysis of genomic data can help in early detection and diagnosis of many diseases. This paper presents an algorithm based on clustering and conditional mutual information for inference of gene regulatory networks from breast cancer gene expression data set. Our algorithm has been compared with conventional approaches in terms of number of true positive nodes and edges, true negative nodes and edges. We have done clinical validation of the networks obtained by gene ontology analysis and gene pair enrichment analysis. The results show that our approach works better.

Keywords: *Conditional Mutual Information, Gene Expression Data, Gene Regulatory Networks, Breast Cancer, Clustering, Mutual Information*

1. Introduction

Application of DNA micro array technology has resulted in large amounts of genomic data. This data can be computationally analyzed to discover the patterns in cellular activity and interactions among cells. A gene regulatory network is group of genes that work in tandem through their outputs and regulate each other's activity. Normal interactions result in normal body health while any dis-regulation of these regulatory networks results in diseases like cancer, Alzheimer's etc.By analyzing the topological structure of these networks, we can understand disease and its progression. We can design appropriate drugs to stop the progression of a disease like cancer by targeting various points of the regulatory networks that lead to it and break the network.

Related Works

The methods for inference of regulatory network are based on information theoretic concepts like mutual information and correlation, machine learning methods from both supervised and unsupervised learning have been applied for this task as well. Supervised learning methods need prior data about regulatory factors, their relations and order of the factors, where as unsupervised learning establishes these relations based on information theoretic or probabilistic approaches. Supervised methods are model based or statistical inference based. Model based methods highlight the structure of network, inference based ones highlight the strength of influence. Unsupervised learning methods include differential equations, partial differential equations and variants of Boolean networks. Bayesian network methods use joint probability distribution to describe the influence of one gene on another through a directed acyclic graph which incorporates prior knowledge required for the task. The authors in [1] have proposed a path consistency algorithm that uses conditional mutual information for inference of gene regulatory networks. They have considered non-linear dependence and topological structure of GRNs. The authors in [2] have used an algorithm called Context Based Dependency Network. This method uses gene expression data to infer GRNs by calculating the changes in expression dependencies between target and other genes and its conditional influence on source genes. The authors in [3] have used both human B cell gene expression data and synthetic data for inference of GRNS. It is a method based on

May-June 2020
ISSN: 0193-4120 Page No. 3863-3865

mutual information. They have also presented the effect of mis-estimation of mutual information on GRNS. They have used Gaussian kernel for estimation mutual information of gene expression values. This method is not able to find regulatory relations in two genes. The authors in [4] have proposed a method that uses stability criterion to select the interactions that are long lasting and reliable. In [5] authors have used random forest method to identify the most highly ranked interactions among the genes. In [6] authors have used a bi-clustering approach to identify interaction networks in entire genome at a global level. In [9] authors have proposed a Bayesian method for estimation of missing gene expression values for improving the accuracy of inference of gene expression networks. 1.2 Problem Statement The problem that this paper is that of discovering the regulatory interactions among genes in breast cancer samples. We have used conditional mutual information among gene expression values to find groups of correlated genes and their complex interrelationships. The input data is the gene expression data sets with accession number GSE2109 and GSE42568 for breast cancer downloaded from NCBI GEO portal [8]. We have applied K-NN clustering and conditional mutual information for identification of groups of correlated and similar genes.

2. Proposed Algorithm

Data pre-processing

Data pre-processing was done by imputation of missing values through averages, log2 normalization and standardization of the resulting data set. The experiment was performed on i7 Windows 10 Pro machine using following tools Bioconductor R package, Python 3.7, Cytoscape 3.2. Biological process enrichment analysis from Gene Ontology was carried out for the groups of co-expresses genes using DAVID [10] tool. Data set consists of 104 samples from patients between 31 and 89 years of age. There were 17 normal breast samples, 82 invasive ductal carcinomas, 17 invasive lobular , 2 tubular, 3 mucinous tumors. Rest of the paper is organized as follows - Section 2 presents the proposed algorithm for inference of gene regulatory networks, section 3 presents the results and a discussion there of. Section 4 concludes the paper

Proposed Algorithm

A total of 18000 genes were run through the algorithm presented in Table 1.–

Table 1: Proposed Algorithm

Algorithm 1: Algorithm for Inference of Gene Regulatory Network In Breast Cancer
Data: Breast Cancer Gene Expression Dataset with accession number GSE2109 and GSE42568 from NCBI GEO Portal
Result: A set of nodes and a set of edges with highest mutual information comprising the network
Find the set of differentially expressed gene D
while there is a unconsidered gene g in D do
For each gene g compute the clusters of most similar genes using K-NN clustering algorithm
For each cluster C
Compute the conditional mutual information for each pair of genes in C
Sort the mutual information values for each pair in increasing order
Take top 10 percent of genes and edges as the network
Visualize the networks
End

3. Results and Discussion

We obtained multiple networks ranging from 3 nodes to maximum 200 nodes. The validation of the networks was carried out by gene pair enrichment. We compared our algorithm with two standard algorithms ARACNE and B3CNET with respect to number of true positives, true negatives, false positives and false negatives. The results show that our algorithm performs better in all the parameters. Figure 1 shows the networks around the genes BRCA 1 taking into account only the physical interaction among the genes. Figure 2 shows the interaction among genes with shared proteins and co-localization. Figure 3 shows the predictions of physical interaction among the genes. Figure 4shows a comparison of true positives and true negatives in terms of number of nodes and edges in predicted networks by our algorithm, ARACNE and B3CNET.

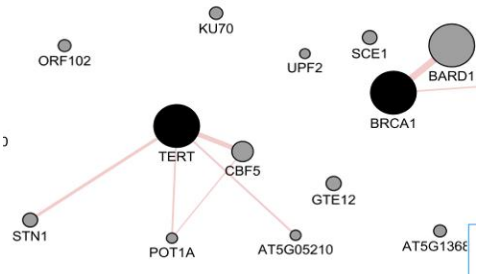

Figure 1: Networks around BRCA 1 with physical interaction

May-June 2020
ISSN: 0193-4120 Page No. 3863-3865

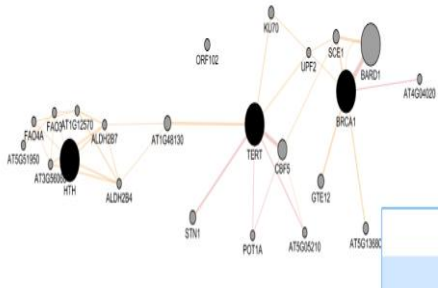

Figure 2: Interaction among genes with shared proteins and co-localization

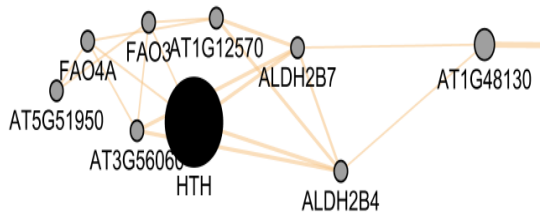

Figure 3: Predicted physical interaction among the genes

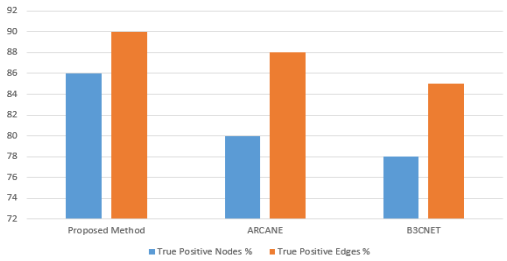

Figure 4: Comparison of proposed method with ARACNE and B3CNET

4. Conclusion

The algorithms for GRN inference are based on the concepts of information theory, Boolean networks, Bayesian Networks, differential equations etc. The algorithm proposed here is from information theory and uses differential expression, clustering, and mutual information for inferring the regulatory relationships among the genes in breast cancer gene expression dataset. The comparison with respect to true positives, true negative, false positives and false negatives with ARACNE and B3CNET shows that our algorithm performs better on the said dataset.

Acknowledgment

The authors would like to thank the Department of Science and Technology(DST), Government of India, for financial support to this work under the scheme DSTICPS, 2018.

References

[1] Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, Luonan Chen, ”Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information”, Bioinformatics, Volume 28, Issue 1, 1 January 2012, Pages 98104, https://doi.org/10.1093/bioinformatics/btr626

[2] Lu Zhang, Xi Kang Feng, Yen Kaow Ng and Shuai Cheng Li, ”Reconstructing directed gene regulatory network by only gene expression data”, In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine 2015, Washington, DC, USA. 9-12 November 2015

[3] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, DallaFavera R, Califano A., ”ARACNE: an algorithm for the reconstructionof gene 5 regulatory networks in a mammalian cellular context”, BMC Bioinformatics 2006;7 Suppl 1:7.

[4] Haury AC, Mordelet F, Vera-Licona P, Vert JP.,”TIGRESS: trustful inference of gene regulation using stability selection”, BMC Syst Biol. 2012;6:145.

[5] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P., ”Inferring regulatory networks from expression data using tree-based methods”, PLoS ONE 2010;5: 9(e12776): 110.

[6] Reiss DJ, Baliga NS, BonneauR., ”Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks”, BMC Bioinformatics 2006;7:280.

[7] Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V., ”The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.’, Genome Biol. 2006;7(5):36.

[8] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A., ”NCBI GEO:archive for functional genomics data setsupdate’, Nucleic Acids Res. 2013;41 (Database issue):9915

[9] Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S., ”A Bayesian missing value estimation method for gene expression profile data”, Bioinformatics 2003; 19(16): 208896.

[10] http://david.abcc.ncifcrf.gov