ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Reconstruction of Gene Regulatory Network to Identify Prognostic Molecular Markers of the Reactive Stroma of Breast and Prostate Cancer Using Information Theoretic Approach

Prof. Shanthi Mahesh[1], Dr. Neha Mangla[2], Pooja V[3], Suhas A Bhyratae[4]

Department of ISE, Atria Institute of Technology, Bengaluru, Karnataka, India. [1,2,3,4]

ABSTRACT: Gene regulation refers to a number of sequential processes, the most well-known and understood being translation and transcription, which control the level of a gene's expression and ultimately result with specific quantity of a target protein. Reconstruction of gene regulatory networks is a process of analyzing the steps involved in gene regulation using computational techniques. In this paper, cancer-specific gene regulatory network has been reconstructed using information theoretic approach-Mutual Information. The microarray database used contains 12 Gene samples each of breast cancer and prostate cancer having both normal and tumor cell information. This data has been preprocessed, normalized and filtered using the t-test; the MI value is applied on the filtered genes to determine the Gene-Gene Interaction. Based on the interactions obtained, 10 different networks have been constructed and the statistical analysis has been performed on that network. Finally, validation of the inferred results has been done with available biological databases and literature**.**

KEYWORDS: *gene* regulatory network, microarray, reactive stroma of breast and prostate cancer, mutual information

I. INTRODUCTION

Malignant cancer is one of the most widespread diseases in today's world that affects the mortality rate of human beings. The cancerous cells divide and grow in an uncontrollable manner forming tumors and infest the nearby part of body. Various significant genes are responsible for the genesis of different tumors. Radiotherapy, chemotherapy and surgery are the possible ways of treating cancer. Therefore, identification of genes that lead to cancer can typically solve the uncontrollable growth of cancer at an early stage.

Reconstruction of gene regulatory networks (GRNs) explicitly represents the causality of developmental or regulatory process. It has become a challenging computational problem for understanding the complex regulatory mechanisms in cellular systems. An important problem in molecular biology is to identify and understand the gene regulatory networks (GRNs). Microarray technologies have produced tremendous amounts of gene expression data, which provide opportunity for understanding the underlying regulatory mechanism.

Recently, information theoretic approaches are increasingly being used for reconstructing GRNs. Several mutual information based methods have been successfully applied to infer GRNs and minet. In general, these approaches start by computing the pair-wise MIs between all possible pairs of genes, resulting in an MI matrix. The MI matrix is then manipulated to identify the regulatory relationships. MI provides a natural generalization of the correlation since it measures non-linear dependency and therefore attracts much attention. Another advantage of these methods is their ability to deal with thousands of variables (genes) in the presence of a limited number of samples. With these advantages, MI-based methods only work when investigating pair-wise regulations in a GRN. The inference of gene networks from high-throughput data is a very complex and vastly expanding; triggered by the invention of measurement technologies. In order to provide a systematic discussion of the underlying principles we limit this review to observational steady-state gene expression data and consider correlation-and mutual information-based inference methods. These methods are representative of linear and non-linear methods. Principally, there are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

three fundamental levels of a molecular system as given by the central dogma of molecular biology (Crick, 1970), namely, the DNA, mRNA and the protein level. Figure 1 shows the overview of Central Dogma. The central dogma of molecular biology describes the two-step process, transcription and translation by which the information in genes flows into proteins:
DNA->RNA->Protein

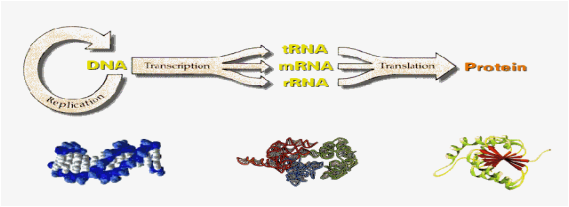

Figure 1: Central Dogma

In this work, we propose a relevance network model for gene regulatory network inference which employs mutual information to determine the interactions between genes. For this purpose, we propose a mutual information estimator based on adaptive partitioning which allows us to condition on both discrete and continuous random variables. We provide experimental results that demonstrate that the proposed regulatory network inference algorithm finds the high degree genes and predicts the gene responsible for both breast cancer and prostate cancer. The results are validated using biological database.

II. LITERATURE SURVEY

The reconstruction or 'reverse engineering' of GRNs, which aims to find the underlying network of gene–gene interactions from the measurement of gene expression is considered one of most important goals in systems biology [2,3]. For this, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) program was established to encourage researchers to develop new efficient computation methods to infer robust GRNs [4]. A variety of approaches have been proposed to infer GRNs from gene expression data [5,7], such as discrete models of Boolean networks and Bayesian networks[8], differential equations [9-12], regression method[13,14] and linear programming [15]. Although many popular network inference algorithms have been investigated [16, 5], there are still a large space for current models to be improved [20]. Recently, information-theoretic approaches are increasingly being used for reconstructing GRNs. Several mutual information (MI)- based methods have been successfully applied to infer GRNs, such as ARACNE, CLR [23] and minet [21]. In general, these approaches start by computing the pair-wise MIs between all possible pairs of genes, resulting in an MI matrix. The MI matrix is then manipulated to identify the regulatory relationships. MI provides a natural generalization of the correlation since it measures non-linear dependency (which is common in biology) and therefore attracts much attention. Another advantage of these methods is their ability to deal with thousands of variables (genes) in the presence of a limited number of samples. Despite these advantages, MI- based methods only work when investigating pair-wise regulations in a GRN.

III. ABBREVIATIONS

DNA: Deoxyribo nucleic acid RNA: Ribonucleic acid NCBI: National center for Biotechnology Information GEO: Gene expression omnibus TMI: Threshold Mutual Information Microarray: Collection of microscopic DNA spots attached to solid surface. MI: Mutual Information. GRN: Gene Regulatory Network.

IV. DATASET DESCRIPTION

Evaluation of the performance of our approach is experimentally tested on the Reactive stroma of breast and prostate cancer dataset. The full data set can be downloaded from the Gene Expression Omnibus website: http://www.ncbi.nlm.nih.gov/geo/GSE26910. The dataset has information on 54675 genes under 24

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

different experimental conditions.

V. METHODOLOGY

The algorithm presented in this approach is shown in Figure 2.

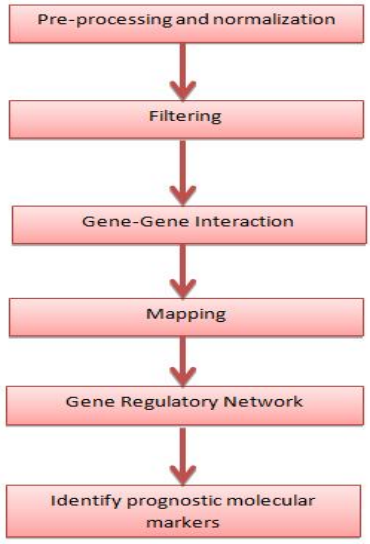

Figure 2: System Flow chart

A. Preprocessing and Normalization

The dataset is quite large with 54675 genes and a lot of information corresponds to genes that do not show any interesting changes during the experiment. During the pre- processing, genes that do not show any changes during the experiment are removed which reduces the size of the dataset. If we look through the gene list, we have several spots marked as 'EMPTY'. These are empty spots on the array and these spots can be noise. The function isnan() is used to identify the genes with missing data and indexing commands are used to remove the genes with missing data.

B. Filtering

T-test is applied between the normal and tumor cell data to obtain the most significant genes. The t-test for unpaired data and both for equal and unequal variance can be computed as

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} \quad (1)$$

where x_1 and x_2 are the means S_1 and S_2 are variances and N_1 and N_2 are sizes of two groups of samples-tumor and normal. The threshold p-value is set at 0.01. This further reduces the size of the dataset to 1017 genes.

C. Gene-Gene Interaction

MI is generally used as a powerful criterion for measuring the dependence between two variables (genes) *X* and *Y*. For gene expression data, variable *X* is a vector, in which the elements denote its expression values in different conditions (samples). For a discrete variable (gene) *X*, the entropy *H*(*X*) is the measure of average uncertainty of variable *X*.

$$H(X) = -\sum_{x \in X} p(x) \log\, p(x), \quad (2)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Where p(x) is the probability of each discrete value x in X. The joint entropy H(X, Y) of X and Y can be denoted by

$$H(X,Y) = -\sum_{x \in X, y \in Y} p(x,y) \log\, p(x,y), \tag{3}$$

Where p(x, y) is the joint probability of x in X and y in Y. MI measures the dependency between two variables. For discrete variables X and Y, MI is defined as

$$I(X,Y) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{4}$$

MI can also be defined in terms of entropies as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{5}$$

Where H(X, Y) is joint entropy of X and Y. High MI value indicates that there may be a close relationship between the variables (genes), while low MI values imply their independence.

Mutual information therefore measures dependence in the following sense: $I(X; Y) = 0$ if and only if X and Y are independent random variables. This is easy to see in one direction: if X and Y are independent, then $p(x,y) = p(x)\, p(y)$, and therefore:

$$\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) = \log 1 = 0. \tag{6}$$

Moreover, mutual information is nonnegative (i.e. $I(X;Y) \geq 0$) and symmetric (i.e. $I(X;Y) = I(Y;X)$). Mutual Information approach is used to obtain regulatory interactions between the selected significant gene pairs. The TMI value is set at 3.4. The gene relationships with F value greater than TMI are said to interact with each other. The gene numbers which interact with each other are obtained.

D. Mapping

The gene interaction matrix obtained in the previous step is mapped onto the gene names.

E. Gene Regulatory Network

The result of gene-gene interaction matrix is imported into the network visualization and analysis tool, Cytoscape. Cytoscape is more powerful when used in conjunction with large databases of protein-protein, protein-DNA and genetic interactions that are increasingly available for humans and model organisms. It allows the visual integration of the network with expression profiles, phenotypes and other molecular state information and links the network to databases of functional annotations. The interacting genes are selected to obtain networks of interacting genes. This helps us in easily identifying the genes with highest degree. Such genes, called as highly connected genes, are said to have a higher impact in causing cancer.

F. Identify prognostic molecular markers

The highly connected genes are used in the identification of the prognostic molecular markers. This analysis is done using. This analysis is done using G2SBC (Genes-to-Systems Breast Cancer Database) .The G2SBC is a bioinformatics resource that collects and integrates data about genes. From this analysis it is found that the genes GOLM1, CSMD2, MICAL2, TMEM167A, TBC1D2, POSTN, AEBP1, ZNF668, ZFAND3, TXNL1, VOPP1, TRIP13 are common for causing both breast and prostate cancer [22].

VI. EXPERIMENTS AND RESULTS

Experiments were conducted on the reactive stroma of breast and prostate cancer with 54675 genes under 24 different experimental conditions. Regulatory network for 30 genes with 306 interactions is shown in the Figure 3. Statistical

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

analysis of GRN for 30 genes is shown in table 1.

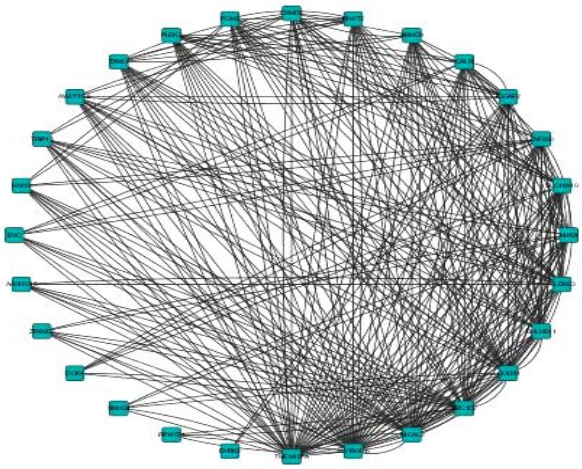

Figure 3: Network for 30 genes with 306 interactions

Clustering Coefficient	0.7
Connected Components	1
Network Diameter	3
Network Radius	2
Network Centralization	0.481
Number of Nodes	28
Number of Interactions	153
Network Density	0.4
Network Heterogenity	0.56
Isolated Nodes	0
Number of Self-Loops	0
Analysis Time(sec)	0.672

Table 1: Network statistics

1. Neighborhood Connectivity: The connectivity of a node is the number of its neighbors. The neighborhood connectivity of a node n is defined as the average connectivity of all neighbors of n, Figure 4. The neighborhood connectivity distribution gives the average of the neighborhood connectivity of all nodes n with k neighbors for k = 0,1,….

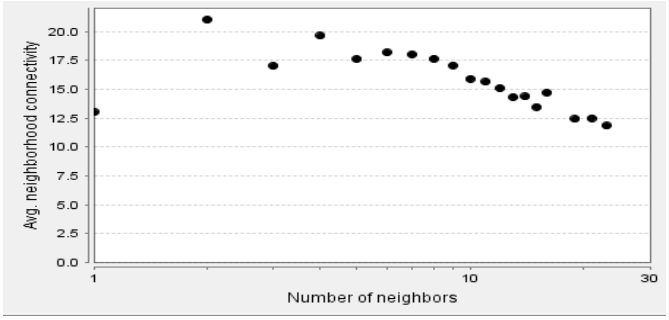

Figure 4: Neighborhood connectivity

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

2. Closeness centrality: It is the degree to which this node is close to all nodes. Figure 5, shows the closeness centrality plotted against number of neighbors. It is calculated based on shortest paths, it is given by,

$Cc(v) = \frac{\Sigma}{t \neq v \in s(v,t)}$ where s(*v*, *t*) is the shortest path between v and t.

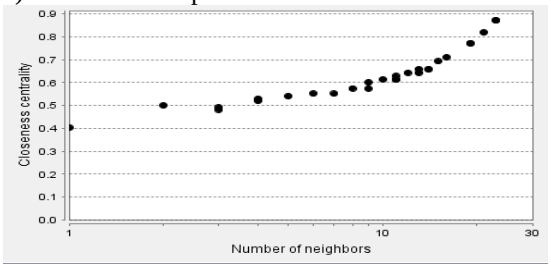

Figure 5: Closeness centrality

3. Shared Neighborhood Distribution: P(n, m) is the number of partners shared between the nodes n and m, that is, nodes that are neighbors of both n and m. Figure 6 shows the shared neighbors distribution for the given number of node pairs (n,m) with P(n,m) = k for k = 1,….

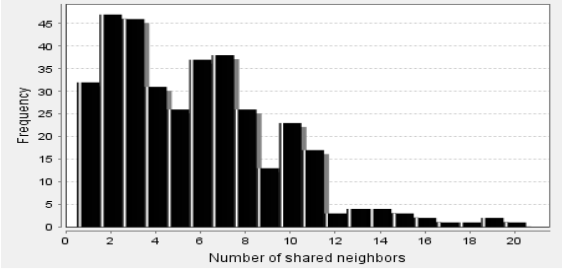

Figure 6: Shared neighbor distribution

4. Shortest Path Distribution: Figure 7 shows the shortest path distribution. The length of the shortest path between two nodes n and m is L(n,m). The shortest path length distribution gives the number of node pairs (n,m) with L(n,m) = k for k = 1,2,….

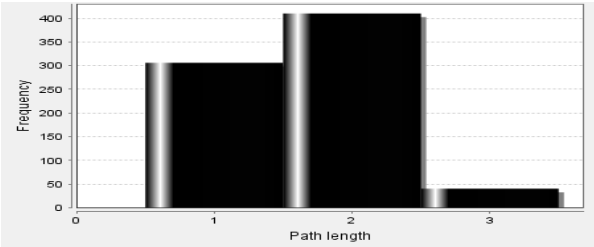

Figure 7: Shortest path-length distribution

VII. CONCLUSION

In this work, a novel approach comprising the features viz, filtering function, mutual information and gene-gene interaction function have been used on the cancer data to compute regulatory relationship between gene pairs and statistical analysis of reconstructed network. The microarray data considered here consists of 54675 genes having 12 sets each of breast and prostate cancer data and 12 each of normal cell data. Our study yields 6 major outcomes; first we identify differentially expressed genes in dataset, second, the interactions between differentially expressed gene

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

have been identified; third, genes regulating most of the other genes were identified; fourth, provides the statistical analysis of reconstructed network revealed a large number of interactions in the used data; fifth, provides the highly connected gene for 10 different network and sixth, helps to identify the prognostic molecular markers in the reactive stroma of breast and prostate cancer using G2SBC. From this analysis it is found that the genes GOLM1, CSMD2, MICAL2, TMEM167A, TBC1D2, POSTN, AEBP1, ZNF668, ZFAND3, TXNL1, VOPP1, TRIP13 are common for causing both breast and prostate cancer. The result provides an excelled understanding of the interaction mechanism of the breast and prostate cancer data and provides new insight into the biomedical world.

NETWORK NO.	NO. OF GENES	NO. OF INTERACTION	TOP FIVE GENES WITH HIGHEST DEGREE	ANALYSIS TIME(sec)
Network 1	10	44	CLDN23(7),GOLM1(7), CSMD2(6),RNFT2(6),MICAL2(6)	0.095
Network 2	20	138	AW190406(15),TMEM167A(14), CLDN23(13),MICAL2(13),GOLM1(12	0.271
Network 3	30	306	AW190406(23),TMEM167A(23), MICAL2(21),TBC1D2(19),GOLM1(19	0.672
Network 4	40	708	COMP(37),ITGBL1(36), TMEM167A(31), AW190406(30), MICAL2(29)	0.581
Network 5	50	1266	COMP(47),ITGBL1(46), CTHRC1(45),ASPN(44),POSTN(44)	1.404
Network 6	60	1884	AEBP1(57),AI040305(56), ARMC9(55),AW190406(54),ASPN(54)	1.798
Network 7	70	2670	ZNF668(67),ZFAND3(66), TXNL1(65),VOPP1(65),TNS3(64)	4.349
Network 8	80	3786	ZFAND3(77),ZNF668(77), VOPP1(76),TNS3(74),TRIP13(74)	6.885
Network 9	90	5178	COL10A1(87),COMP(87), ITGBL1(86),ASPN(84),COL11A1(84)	14.468
Network 10	100	6774	COL10A1(97),COMP(97), ITGBL1(96),CTHRC1(94),ASPN(94)	30.139

Table 2: Ten different networks, number of genes involve each, five highly connected genes with their degrees.

REFERENCES

[1] Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
[2] Basso,K. *et al*.(2005) Reverse engineering of regulatory networks in human B cells.*Nat. Genet.*, **37**, 382–390.
[3] Margolin,A.A. *et al.* (2006a) Reverse engineering cellular networks. *Nat. Protoc.*, **1**,663–672.
[4] Marbach, D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.
[5] Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**,78.
[6] Holter,N.S. *et al.* (2001) Dynamic modeling of gene expression data. *Proc. Natl Acad.Sci. USA*, **98**, 1693– 1698.
[7] Tegner,J. *et al.* (2003) Reverse engineering gene networks: integrating genetics
[8] Kauffman,S. *et al.* (2003) Random Boolean network models and the yeast transcriptional network. *Proc. Natl Acad. Sci. USA*, **100**, 14796–14799.
[9] Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.
[10] Cantone,I. *et al.* (2009) A yeast synthetic network for in vivo assessment of reverse engineering and modeling approaches. *Cell*, **137**, 172–181.
[11] di Bernardo,D. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol*.,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

23, 377–383.
[12] Honkela,A. *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA*, **107**, 7793–7798.
[13] Gardner,T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.
[14] Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc.B*, **58**, 267–288.
[15] Wang,Y. *et al.* (2006) Inferring gene regulatory networks from multiple microarray datasets. B*ioinformatics*, **22**, 2413–2420.
[16] Altay,G. and Emmert-Streib,F. (2010) Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, **26**,1738–1744.
[17] Brunel,H. *et al.* (2010) MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis.*Bioinformatics*, **26**, 1811–1818
[18] Smet,R.D. and Marchal,K. (2010) Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, **8**, 717–729.
[19] Meyer,P.E. *et al.* (2008) minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.
[20] Margolin,A.A. *et al.* (2006b) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
[21] Adamcsek,B. *et al.* (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 2006, **22**, 1021–1023.
[22] Identification of Prognostic Molecular Features in the Reactive Stroma of Human Breast and Prostate Cancer Anne Planche, Marina Bacac, Institute of Pathology, CHUV, and Faculty of Biology and Medicine, University of Lausanne,