World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:6, No:5, 2012

Clustering Approach to Unveiling Relationships between Gene Regulatory Networks

Hiba Hasan, Khalid Raza

***Abstract*—**Reverse engineering of genetic regulatory network involves the modeling of the given gene expression data into a form of the network. Computationally it is possible to have the relationships between genes, so called gene regulatory networks (GRNs), that can help to find the genomics and proteomics based diagnostic approach for any disease. In this paper, clustering based method has been used to reconstruct genetic regulatory network from time series gene expression data. Supercoiled data set from Escherichia coli has been taken to demonstrate the proposed method.

***Keywords*—**Gene expression, gene regulatory networks (GRNs), clustering, data preprocessing, network visualization.

I. Introduction

THE Central dogma applies that DNA makes by replication, DNA makes RNA by transcription and DNA again makes protein by translation. RNA can be made from DNA by reverse transcription which by PCR can be used for various purposes. Gene expression data is basically extracted from a 96 pixel chip. The dye, for example, coomassie brilliant blue is applied on the pixels. The different proteins fluoresce differently on the pixels and the various kinds of proteins are identified by viewing different colors. Gene expression in prokaryotes are identified in a type of fragment in genome called operon which consists of three enzymes transgalactosidase, permease and the thiogalactosidase and three regions operator, promoter and a Z region (not identified). The operator biases the protein attaching to the promoter. When the protein attached with the inducer the protein is controlled by the enzymes, mainly by transgalactosidase to form various products. When the protein is attached with repressor the protein is not recognized and the regulation does not occur. The same approach is applied on the data set of *Escherichia coli* of the supercoiling gene expression data which has 4608 genes. Supercoiling gene expression data can easily be recognized and executed by the dyes. As such by various techniques the gene regulatory networks can be identified [1].

The huge data set can reveal the interaction between the genes which can thus be used as utilization of the data achieved for identification of the drug target owing to high throughput screening technology such as ChIP, DNA microarray and protein-protein interactions. The huge dataset can be compressed by these approaches and the various outcomes can be achieved by identifying the issues [2] [3].

Hiba Hasan is with the Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi, India (e-mail: hbhsn96@gmail.com).

Khalid Raza is with the Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi, India (phone: +91-9891-4782-55; e-mail: kraza@jmi.ac.in).

II. Previous Works

Various techniques have been proposed for identification of the gene regulatory networks (GRNs) from the time-series gene expression profiles. REVEAL [4], reverse engineering algorithm proposes the mutual information technique which implies Shannon entropy which can find dependency between the attributes which thereby could produce the gene regulatory network. The proposed work in [4] has been produced for Bayesian and regulatory matrix methods. Boolean networks involve the production of GRNs on the basis of true and falseness of the data by giving it the values 0 and 1. The Bayesian methods involve the probabilistic approach and regulation matrix involves the dataset involving the time series and steady state data. Other machine learning approaches have also been proposed. Wang and Gotoh [6] constructed cancer specific GRN using soft computing rules. We the extensive review, refer [2] [5] [11].

Most of the previous study has many major drawbacks such as dimensionality problem, computational complexity problem and an experimental measurement problem. There are too many genes with lesser time points [7]. A priori information is needed with guidelines provided of the designed data.

III. Materials and Methods

In this paper, a clustering technique has been applied to find the regulatory relationships among genes.

A. Methods

To identify a GRN the proposed work is follows:

- Data set is identified and downloaded
- Preprocessing of data set is done
- Clustering of data set is overviewed
- Post processing of the clustered data set
- Visualization is done as GRNs are synthesized.

B. Data Set

Data set can be downloaded by various sites such as SMD (Stanford Microarray Database) or GEO (Gene Expression Ontology). The current data set has been downloaded from SMD that consists of 4608 rows (genes) and 24 columns (environment in which the data set is applied) of the supercoiled data set from Escherichia coli. Clustering-based approach has been applied to produce the GRNs.

C. Data Preprocessing

Data preprocessing by various techniques such as normalization is undergone. The comparison of attributes of data by various techniques such as log transformation is known as normalization. Log transformation can be identified by dyes Cyt3 and Cyt5 and the ratios are further overviewed. The scaling factor may be used to identify the data sets attributes.

D. Clustering

The data set is then clumped into various pattern and architectures by various techniques such as hierarchical clustering in which no centroid related vector-based search is applied. It can be classified into two parts - the hierarchical divisive clustering and hierarchical agglomerative clustering which are based division and clumping of the data set respectively. Hierarchical agglomerative clustering can again be classified into various types:

- Single linkage: The minimum distance between the correlated attributes is considered. The grouping is based on minimum dissimilarity between members of each group.
- Complete linkage: Greatest distance between the non-correlated attributes is considered. It is also known as maximum or furthest neighborhood method.
- Average linkage: Average distance between the correlated attributes is used. It can be weighted or unweighted depending on the size of the clusters created.
- Centroid method: Mean of the various correlated attributes is taken as centroid and calculated which result in merging of clusters.
- Ward's method: The squared distance between the mean centroid is overviewed.

k-means clustering: The various means taken from the correlated data is produced and hence the centroid is deciphered. The minimum difference between the centroids nodes, called seeds, are classified and considered. By taking the subsets by 'rounds of clustering', the clustered data is produced. It is a supervised form of clustering.

Self-organising map (SOM) and trees: It is the supervised form of clustering in which a gene is picked at random and a reference vector is placed and hence a map is created.

Principal component analysis (PCA): It is the unsupervised form of clustering which involves the clustering of the data set along various dimensions of the data and hence a single clustered data set is evolved which is organized into a axis of the dimensions of the data set.

E. Further Processing of the Data

Further processing of the data set involves the reproduction of clustered data set into a form of dendrograms, scatter plot etc. to consider the roots and various dimensions of the data set. Dendrograms can be used to identify the genetics of the dimensions of the data (Fig. 1). In this data set, for example, the root is deeply attached to the latter part of the dimensions of the data set.

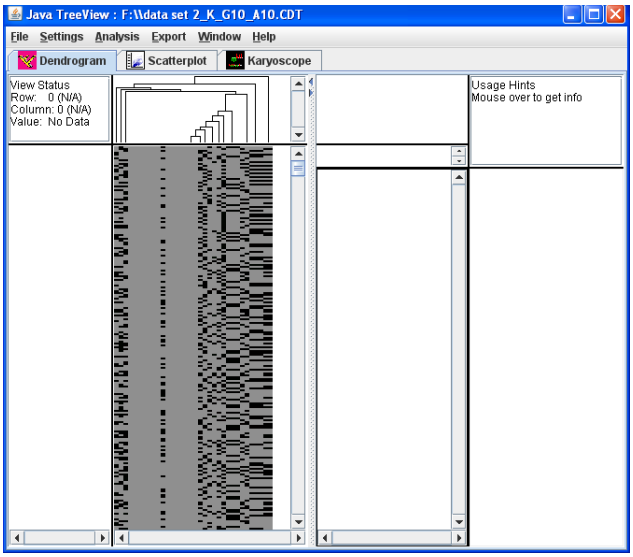

Fig. 1 Dendrogams attached from root to later part of the data set

Scatter plots involves the dimensions of the data set. The data set produces a plot that involves the highly clumped points of data set in the latter part of the data set (Fig. 2).

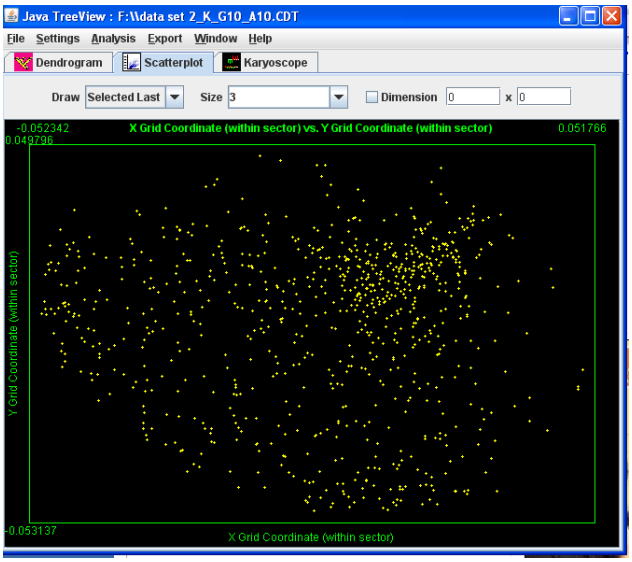

Fig. 2 Genes in the later dimensions of the data set

IV. Visualization of Data Set for GRNs

Gene regulatory networks are used to produce nodes and edges of the data set using the software tool Cytoscape (freeware installed from website www.cytoscape.org). The GRNs when plotted involves the distinguishing of the un-clustered (Fig. 3) and clustered data (Figs. 4 & 5). The clustered data and unclustered data involve the difference between the edges of the data set. The unclustered data set involves lesser edges as compared to the clustered data and the clustered data set is highly interlinked and has more edges and connections as compared to the unclustered data which gives rise to the upregulation and downregulation of the data set as a result of clustering. The non-regulated genes are self-

regulating and are linked to the other node which can be compensated to produce an effective interface (Fig. 6). The various parameters used in the clustering, for example, Euclidean distance, correlation coefficients, standard deviation recognize the same network [8].

The post processing of data set also result in the conclusion that the data set can be used for further purposes such as drug discovery. The result involves the drug identification based results such as the latter part of the dimensional data is firmly attached with the root of the dendrogram [9] [10].

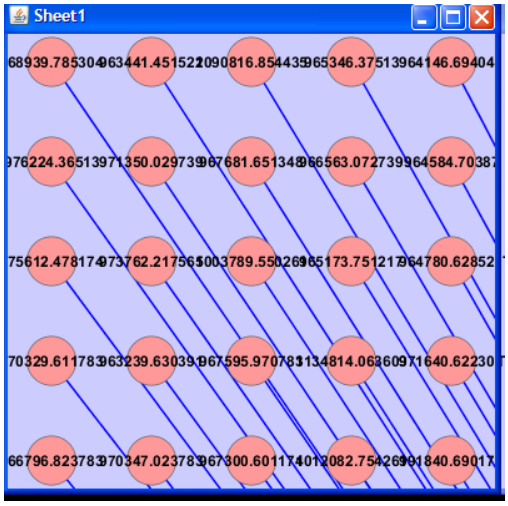

Fig. 3 Un-clustered data with no parameters

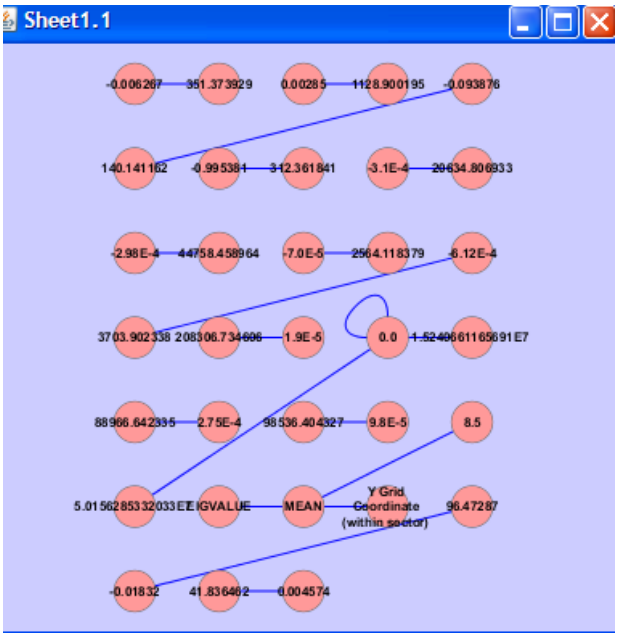

Fig. 4 Network result (SOM)

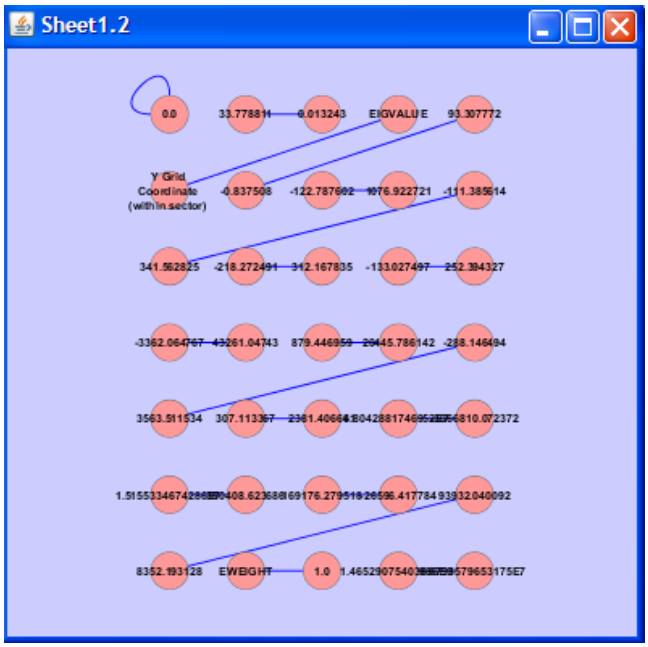

Fig. 5 Network result (SOM)

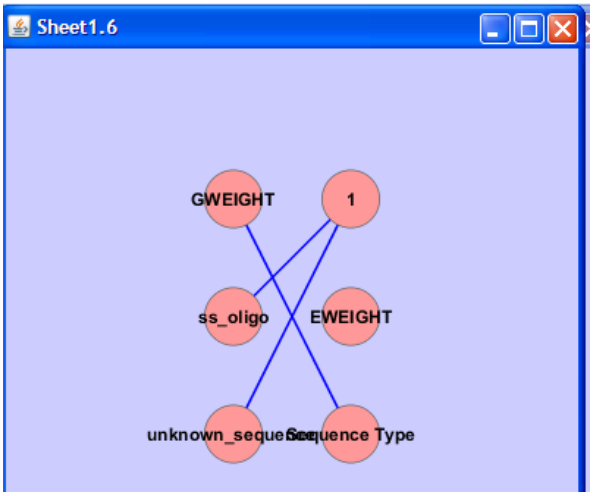

Fig. 6 Showing interacting genes in the later part of the data set

V. Results and Conclusions

GRNs can be used to identify the interactions between the genes, correlation between genes, drug identification and genomics and proteomics based research. For achieving the dependency between the genes by first transformation of the data set and thus reproducing the result in the form of the network we can find the clustered analysis of the genes. Thus we can find the correlation between the genes. The dendrogram obtained by phylogenetic analysis of the data we can thus obtain that the enzymes are keenly supporting the later part of the data dimensionality which can be a part of the identification of the drug target. The log transformation by log sigmoid function could thereby influenced to obtain the interaction of the genes by using the significance of edges of the network produced.

The enzymes functioning involved in the regulation and

expression of the data set are inheriting themselves from the earlier part of the data set as seen in the GRNs. So we can find these properties of the genes involved in the earlier part of the data set. So if data is not harmed, it can be used in drug discovery. Secondly the parameters involved do not change the functionality stream of the data which give rise to the interface that enzymes are originally involved in the cell regulation. Supercoiled data can be used in the drug discovery and other parts of genomics and proteomics based research [12].

In this paper, reconstruction of genetic regulatory networks from time series gene expression data using clustering technique has been done. The supercoiled data of *E. coli* bacterium has been considered for the reconstruction using proposed method. The proposed method includes preprocessing of the raw data set, its clustering using SOM, post processing of clustered data using Java TreeView and finally construction of genetic regulatory network using Cytoscape software tool.

REFERENCES

[1] Helen C. Causton, John Quackenbush and Alvis Brazma, "A Beginner's guide Microarray Gene Expression data Analysis," *Blackwell Publishing*.

[2] K.H. Cho, S.M. Choo, S.H. Jung, J.R. Kim, H.S. Choi and J. Kim, "Reverse engineering of gene regulatory networks," *IET systems Biology, 2007, pp.149-163.*

[3] Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu and Luonan Chen, "inferring gene regulatory networks from gene expression data by PC-algorithm based on conditional mutual information," *Oxford university press.*

[4] Shoudan Liang, Stefanie Fuhrman and Roland Somogyi, "REVEAL, A general reverse engineering algorithm for inferemce of genetic network architectures," P*acific symposium on Biocomputing 3,1998, pp. 18-29*.

[5] Khalid Raza and Rafat Parveen, "Evolutionary Algorithms in Genetic Regulatory Networks Model", *Journal of Advanced Bioinformatics Applications and Research*, 3(1):271-280, 2012.

[6] Xiaosheng Wang and Osamu Gotoh, "Inference of Cancer-speifc gene Regulatory Networks Using Soft Computing Rules," *Gene Regulation and Systems Biology, 2010,19-34.*

[7] Yuji Zhang, Jianhua Xuan, Benildo G de los Reyes, Robert Clarke and Habtom w Ressom, "Reverse engineering module networks by PSO-RNN hybrid modeling,", *BMC Genomics, 2009.*

[8] Brueckner F, Armache KJ, Cheung A, et al., "Structure–function studies of the RNA polymerase II elongation complex". *Acta Crystallogr. D Biol. Crystallogr.* 65 (Pt 2): 112–20, 2009

[9] PP Amaral, ME Dinger, TR Mercer and JS Mattick, "The eukaryotic genome as an RNA machine". *Science* 319 (5871): 1787–9, 2008.

[10] B Schwanhäusser,D Busse, G Dittmar, J Schuchhardt, J Wolf, W Chen and M Selbach, "Global quantification of mammalian gene expression control". *Nature* 473 (7347): 337–42, 2011.

[11] Khalid Raza and Rafat Parveen, "Soft Computing Approach for Modeling Genetic Regulatory Networks", *2nd International Conference on Artificial Intelligence, Soft Computing and Applications (AIAA-2012)*, Proc. published in *Advances in Intelligent Systems and Computing*, AISC 178, pp 1-11, Springer-Verlag Berlin Heidelberg, 2012.

[12] Chesler EJ, Lu L, Wang J, Williams RW, Manly KF, "WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior". *Nat Neurosci* 7 (5): 485–86, 2004.