Reconstruction of Gene Regulatory Network using Bayesian Network

Abhishek Bholal, Sandeep Mittal2
1Chitkara University School of Engineering and Technology
Chitkara University, Himachal Pradesh, India
abhisek@chitkara.edu.in
2Department of Computer Science and Engineering,
R D Engineering College, Ghaziabad
sandeep.mittal2363@gmail.com

Abstract. Gene Regulatory Networks (GRN) are complicated govern systems, that refers to the intercommunication of genes and it also helps to understand and visualize all the biological interconnections. Reconstruction of GRN helps in finding out various hidden facts that can be helpful in solving various complexities related to living life, curing many diseases and inventing new drugs for the humans. There are number of computational methods available in literature for reconstructing GRNs. In this paper, a method is proposed for reconstructing GRN i.e. based upon Bayesian network. The mentioned method in this paper is considered as one of the most capable method for reconstructing GRN. The proposed method functions in two stages. In its first stage, there is two level clustering is employed to diminish the size of given dataset. In its second stage, Bayesian Network will be utilized to reconstruct the network for the sub-clusters that are gained from the first stage. The proposed method is also validated and evaluated using different parameters.

Keywords: Gene regulatory Network, K-mean Clustering, Bayesian network, Gene Expression Data

1. Introduction

Gene is the basic building block of living life and gene controls it’s all biological processes. All the biological activities are directly or indirectly connected to genes. Due to its complicated and mysterious nature, study of gene regulatory networks (GRN) has utmost importance therefore numerous methods have been proposed in the past to reconstruct GRN[1-6].Ramesh Ram et al. have proposed a Fuzzy logic model that searches Microarray data for activator/repressor relationship. This method eradicates likely fallacious predictions from the classical Fuzzy model, so that permitting a larger search space for inferring regulatory relationships among number of Genes [7]. Young Hwan Chang et al. have developed a method i.e. based on the concept of repairing Sparse Low-Rank Matrices. This method is important from the view that it considers perturbations can affect the result inferred by GRNs. It estimates the error, occurred due to these perturbations and identifies the common graph structures among biased inferred structures [8]. Sudip Mandal et al. have specified Neural Model for reconstructing GRN, which is effective at that particular time to specify relationship among number of Genes [9]. This model is implemented in the form of Weighted Matrix. Further, another method has been specified i.e. proposed by Nilanjan Santra et al. They have known the importance of GRN, because it helps in revealing the disease motifs and also in solving the complexities related to living life. They have used Firefly Algorithm for reconstructing and learning GRN effectively. Although the method supposed by this work has accuracy rate i.e above 50 % [10].Raed I. Hamed et al. have discussed a fuzzy Petri net (FPN) method i.e. used for reconstructing Gene Regulatory Networks. The approach specified by their work is more propitious than the Fuzzy Logic. The concept used by this work is better as compared to simple Fuzzy Logic approach because it removes the plausible untrue predictions from the classical Fuzzy model so that enabling a larger search space for inferring regulatory relationship [11].

This paper has been organized in four main sections i.e. Bayesian Networks, datasets & methodology, results and conclusion. Section 2 discusses about Bayesian Networks and how they are inferring GRN effectively. Section 3discusses data sets and methodology used in this work. Next Section 4 analyses result and discuss it. Finally, section 5highlights conclusion based on the proposed work and future directions.

2. Bayesian Network Method

Bayesian Network methods are one of the mostly used inference architectures for reconstructing GRNs. In this paper, Bayesian Network Methods have been used for reconstructing GRNs. These methods are also called probabilistic models, because they make use theorem of probability i.e. each node is connected to another on the basis of some probabilistic value. Bayesian Methods are directed acyclic graphs, which consist of nodes describing characteristics and acyclic edges present between them highlighting
independancy. Each of the nodes is joined to the other on the basis of Probability Distribution Table, highlighting the chances of connecting each node to the other. Bayesian methods are considered to be highly capable because they make use of both graph concept and probability concept. These methods are quite understandable and have been used successfully in many of the applications and research works. These methods are mathematically capable because of having concept of probability and reconstruct the GRNs effectively by giving accurate results.

3. Proposed Methodology & Datasets

The reconstruction of GRN can be said as one of the key tasks in the area of Bioinformatics because for reconstructing GRN we need calculative and capable computational method that can infer GRN effectively. Bayesian Computational method makes use of Genomic Datasets as input those are obtained from the Microarray analysis. But handling of complete Datasets obtained from Microarray Gene Expression Data is impossible that’s why Clustering technique has been used for monitoring a large number of Genes.

There are some of the authenticated source like gene expression omnibus (GEO)repository of NCBI that are considered for obtaining the needed gene expression datasets[12-13]. In the proposed framework the Gene Expression Dataset of E. coli is used and the methodology that has been used to used and implemented to reconstruct the GRN is shown in Fig 1.

Fig. 1 Proposed methodology

Step 1: Pre-processing: In this step, pre-processing is done on the dataset obtained from the GEO data repository. It eliminates the Genes those are having missing names from the Gene Expression Data. It also removes the gene profiles that have low variance.

Step 2: Clustering

In this step Clustering is implemented on the gene expression data [14]. Clustering is a technique of grouping the similar things into one group and different objects into different Groups. This technique can group the Genes having similar expressions under various conditions. There are number of Clustering techniques present in literature, but in the proposed method K-mean clustering has been used [15]. It is a clustering technique that classifies the objects (Genes) based on some characteristics into K number of groups (clusters). In this clustering, computational approach begins, by positioning K points into the space represented by the objects that are being clustered.

Step 3: GRN Reconstruction for sub-Clusters

Now, clusters have been created in the previous step. In this step, Bayesian Network is used for reconstructing GRN for each of the sub-clusters [16]. Each sub-cluster will be treated as an individual, independent and complete Dataset and proposed method will be implemented and GRN will be reconstructed for each of the sub-Clusters.

Step 4: Merging

After conducting the previous step, there has been number of sub-networks or network is created for each of the sub-clusters [17]. Now, in this step merging of sub-networks is done. This step enables the making of main network. For creating main network, all sub-networks (related to each of the cluster) are concatenated and main network is produced.

Step 5: Evaluation and validation of Results

In this step the results are evaluated based on the number of parameters e.g. True Positive Rate, True Negative Rate, F-score and Total Time taken.
For validation Gene-Mania is used to validate accuracy of results.

4. Result and Discussion

The datasets used in this work has been downloaded from publicly available Gene Expression Omnibus (GEO) data repository of NCBI [18]. In this section, the results obtained from the proposed methodology, has been discussed. At the end of first step i.e. pre-processing step, the obtained gene expression data will be free from most of the anomalies [19-20]. After pre-processing, the clustering technique has been applied so that various clusters of genes can be formulated. After clustering technique, GRN is constructed for each of the sub-clusters individually as shown in Fig 2. These all sub-networks may vary in shape and size all though this work has applied same method for reconstructing GRN for each of the sub-clusters.

![Fig. 2 Sub-network for each of the sub-clusters](image)

All the GRN of sub-cluster reconstructed in previous step are merged into a single Network as shown in Fig 3.

![Fig. 3 Merged GRN](image)
In the previous step all the sub-networks belonging to each of the sub-clusters are merged to form single network. Now results are evaluated and validated using Gene-MANIA platform [21].

Various parameters used for the evaluation of results includes: true positive rate, True Negative Rate, F-score and Total time taken. The results for each parameter that are taken into consideration for evaluating performance, is shown in the Table 1.

| Parameter                  | Calculated Value |
|----------------------------|------------------|
| True Positive Rate         | 0.3              |
| True Negative Rate         | 0.285            |
| F-score                    | 0.29             |
| Total Time taken (in Seconds) | 102             |

Based on the above parameter’s performance measure of the proposed method for different number of genes has been also done and shown in Table 2.

| Experiment No. | Number of Genes | True Positive Rate | True Negative Rate | F-score |
|----------------|-----------------|--------------------|--------------------|---------|
| 1              | 7               | 0.80               | 0.57               | 0.66    |
| 2              | 8               | 0.69               | 0.60               | 0.64    |
| 3              | 9               | 0.60               | 0.60               | 0.60    |
| 4              | 10              | 0.52               | 0.52               | 0.52    |
| 5              | 11              | 0.50               | 0.52               | 0.50    |

5. Conclusion and Future Scope

This work has discussed all the facts related to GRN and stated how previous work lack in reconstructing GRN effectively. The work done in this paper, we tried to eradicate some of the gaps which are present while reconstructing GRN efficiently. For future perspectives more work needs to be done and investigated to improve the result. It is one of the challenging fields to which needs more intense works as it helps in uplifting the standard of human beings by inventing personalized drugs and diagnosis tools. The study of GRN also helps in inspecting the Gene’s behavior towards changing environment and solving the complexities related to living life. Therefore, the work to reconstruct GRN must be continued so that results can be further improved.

References

[1] Fernando M. Delgado, Francisco Gomez-Vela: Artificial Intelligence in Medicine, Volume 95 (April 2019)
[2] Sudip M, Goutam Saha,Rajar K Pral :Reconstruction of Dominant Gene Regulatory Network from Microarray Data using Rough Set and Bayesian Approach (2013)
[3] Paras B, Shailendra S, Abhishek, A Review, On Bayesian Network Techniques for Inferring Gene Regulatory Networks (2015)
[4] Jure Bordon, MihaMoskon, Nikolaj Zand MihaMraz: Fuzzy logic as a computational tool for quantitative modelling of Biological Systems with uncertain Kinetic data (2015)
[5] A Brazma, P Hingamap, J Quackenbush, G Sherlok: Minimum Information about a microarray experiment (MIAME) – towards standards for Microarray Data (2001)
[6] R Seidel- Discrete and Computational Geometry: The nature and meaning of perturbations in Geometric Computing (1998)
[7] Ramesh Ram, Madhu Chetty, Trevor I. Dix: Fuzzy Model for Gene Regulatory Network (2006)
[8] Young Hwan C, Roel Dobbe, Palak Bhushan, Joe W. Gray, and Claire J. Tomlin: Reconstruction of Gene Regulatory Networks based on Repairing Sparse Low-Rank Matrices (2016)
[9] Sudip Mandal, Gautam S, Rajat K Pal: Neural Network Based Gene Regulatory Network Reconstruction (2015)
[10] Nilanjan S, Surama Biswas, Sriyankar Acharyya: Neural Modeling of Gene Regulatory Network using Firefly Algorithm (2015).
[11] Raed I. Hamed, Syed I. Ahson, Rafat Parveen: Fuzzy Petri Nets Approach (2010)
[12] A Kolinski, J Skolnick, JS Fettow: Structural Genomics and its importance for gene function analysis (2000)
[13] JJ Geeman, P Buhlmann- Bioinformatics: Analyzing Gene Expression Data in terms of Gene sets: methodological issues (2007)
[14] AR Deshpande, L Lobo-International Journal of Engineering: Text Summarization using Clustering Technique (2013)
[15] L Morisette, S Chartier- Tutorials in Quantitative Methods for Psychology: The K-means clustering technique: General considerations and implementation in Mathematics (2013)
[16] ABhalerao, A Katkar, J Gahlot: Systems and Methods for managing sub-clusters within dependent clustered computing systems subsequent to partition events (2015)
[17] C Fontaine, PR Guimaraes Jr, S Kefi, N Loeuille: The Ecological and Evolutionary implications of merging different types of networks (2011)
[18] HO, MLazaar, Y Tabii: Effect of Feature Selection on Gene Expression Dataset Classification Accuracy (2018)
[19] ST Sherry, MH Ward, M Kholodov, J Baker: dbSNP: the NCBI database of Genetic variation (2001)
[20] E Clough, T Barret- Statistical genomics: The Gene Expression Omnibus Database (2016)
[21] J V, K Zuberi, H Rodriguez, R Arnold: Novel Function discovery with Gene-MANIA: a new integrated resource for gene function prediction in E. Coli (2015)