iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45879

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Reconstruction of Gene Regulatory Network for Colon Cancer Dataset

Suhas A Bhyratae[1], Divya N[2], Elton Glenvill Pinto[3], Vinuta S Badami[4], Amogh P Kadamannaya[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*UG Students, Department of Computer Science and Engineering , Sahyadri College of Engineering and Management, Mangalore, Karnataka*

Abstract: ***Molecular networks involve interacting proteins, RNA, and DNA molecules, which underlie the major functions of living cells. DNA microarray probes how the gene expression changes to perform complex coordinated tasks in adaptation to a changing environment at a genome-wide scale. Microarray is a technology that has been widely used to probe the presence of genes in a sample of DNA or RNA. This technology helps to check the expression levels of thousands of genes together. The DNA microarray was established as a tool for the efficient collection of mRNA expression for a large number of genes. The mapping function route maps pairs of genes that present similar positive, and negative interactions and also defines how the range of each gene is going to be segmented. From all the combinations a function transforms each pair of labels into another one that classifies the type of interaction. This project addresses the challenge of reconstructing molecular networks and gene regulation from gene expression data. Reconstruction of gene regulatory networks which can also be called reverse engineering is a process of identifying gene interaction networks from the experimental microarray gene expression profiles through computation techniques. The main features involved in the computation of interaction in the filtered genes are the discretization mapping function, gene-gene mapping function, and filtering function.***

Keywords: Bioinformatics, Gene Regulatory Networks, Colon Cancer, Gene Interactions

I. INTRODUCTION

Cancer is one of the most destructive diseases and with its growing number, its detection and treatment become a necessity. Researchers have developed numerous methods which are based on gene expression. In recent years, the extensive amount of genetic information generated by new-generation approaches has led to the need for new data handling methods. The integrative examination of diverse-nature gene information could provide a much-sought overview to study complex biological systems and processes. Biological networks are the representation of multiple interactions within cells, a global view intended to help recognize how relationships between molecules dictate cellular behaviour. Recent advances in molecular and computational biology have made possible the study of intricate transcriptional regulatory networks that describe gene expression as a function of regulatory inputs specified by interactions between proteins and DNA. Cellular biology plays an important role in the understanding of life sciences. One such branch of cellular biology deals with genetics. Genetics is the understanding of heredity and variation in living organisms. A gene is the basic unit of inheritances that causes the repetition of certain characteristics in a cell and a better understanding of cell function. A gene regulatory network is a collection of genes that influence/regulate other genes or themselves. Gene regulatory networks can be broadly defined as a group of genes that are activated by particular signals and stimuli, and once activated, orchestrate their operation to regulate certain biological functions, such as development, metabolism, and the cell cycle. These gene networks are therefore known as dynamic objects that continuously sense the environment and orchestrate their operation accordingly. The core of this operation lies in the central dogma of biology which describes how operative information stored in the DNA is used to generate operating elements, mostly proteins. Proteins are produced from an intermediate product which is known as RNA. First, coding regions of DNA are transcribed to synthesize these RNA molecules. Thereafter, proteins are generated through the translation of the RNA molecules. These proteins, in turn, affect the production of other proteins or catalyse and regulate reactions responsible for various cellular activities. The organization and feedback can be perceived as a working definition of a gene regulatory network. Gene expression is the process by which information from a gene is used in the fusion of a functional gene product. These products are often proteins. Several steps in the gene expression process are transcription and translation. Transcription is the special replicating of one side of the DNA molecule i.e. the sense strand that results in the production of a single strand of RNA. The original DNA is not changed and the process can be repeated. Translation is the reading of the RNA code, by the ribosome, to make proteins or polypeptides. Translation is often called protein synthesis.

II. LITERATURE SURVEY

Microarrays are one of the latest breakthroughs in experimental molecular biology [3], which allow monitoring of gene expression for tens of thousands of genes in parallel and are already producing huge amounts of valuable data. There is also a greedy algorithm to identify groups of related genes [4], Clustering algorithms are used to analyse genes in order to group the genes with similar behaviour. The algorithm will allow the researcher to modify all the criteria such as the discretization mapping function, gene-gene mapping function, and the filtering function. It provides the user much flexibility to obtain clusters based on the level of precision needed. Analysis and handling of such data are becoming one of the major bottlenecks in the utilization of the technology. The raw microarray data are images, which have to be transformed into gene expression matrices – tables where rows represent genes, columns represent various samples such as tissues or experimental conditions, and numbers in each cell characterize the expression level of the particular gene in the particular sample.

In recent years gene regulatory networks (GRNs) have attracted a lot of people and many approaches have been introduced for their statistical inference from gene expression data [6]. However, despite their popularity, GRNs are mostly misunderstood. Specifically, their meaning, the consistency among different network inference methods, ensemble methods, the assessment of GRNs, the estimated number of existing GRNs, and their usage in different application domains. XGBoost for gene regulatory networks (XGRN), is a supervised algorithm, which combines gene expression data with previously known interactions for GRN inference [1], whose key idea is to train a regression model for each known interaction of the network and then utilize this model to predict new interactions. XGRN determines a regression model based on gene expression of the two interactions and then provides predictions using the gene expression of other candidate interactor

We also have other methods for inference of gene regulatory networks (GRNs) from transcriptomic data that are used in cancer research [5]. The methods are classified into three categories according to the analysed model. The first category includes techniques that use pair-wise measures between genes, which include correlation coefficient and mutual information. The second category includes approaches that determine the genetic regulatory relationship using multivariate measures, which then consider the expression profiles of genes concurrently. The third category includes approaches using supervised and integrative approaches. Some of the most important statistical approaches used for modelling gene regulatory networks and protein-protein interaction networks are the statistical graphical modelling methods, state-space representation models, and information-theoretic methods [7].

Inference of gene regulatory network from expression data is a competitive task [10]. There are large number of methods for inferring gene regulatory networks from expression data, however, both their absolute and comparative ability remains poorly understood [11], inferring directed gene regulatory networks based on soft computing rules that can identify critical cause-effect regulatory relations of gene expression [9]. Most GRN inference methods are not specific to cancer transcriptome data, and such methods are required for a better understanding of cancer pathophysiology. In addition, more systematic methods for validation of the estimated GRNs need to be developed in the context of cancer biology.

ARACNE, an algorithm, which uses microarray expression profiles [13], is specifically designed to scale up to the intricacy of regulatory networks in mammalian cells, yet general enough to address a wider range of network deconvolution problems. This method uses an information-theoretic method to exterminate the majority of non-direct interactions inferred by co-expression methods. ARACNE shows promise in recognizing direct transcriptional interactions in mammalian cellular networks, a problem that has challenged the current reverse-engineering algorithms.

In [2] a cancer-specific gene regulatory network constructed using a simple and novel statistics-based approach is given. First, significant genes differentially expressing themselves in the disease condition have been identified using a two-stage filtering approach t-test and fold-change measure. Next, regulatory relationships between the identified genes have been computed using the Pearson correlation coefficient. The obtained results have been validated with the available databases and literature.

III. METHODOLOGY AND IMPLEMENTATION

The first step in our approach is to read the gene expression data. This gene expression data consists of empty cells and null values, which can be removed using pre-processing and normalization processes, i.e., the second step. The normalization process is carried out using MATLAB. Then we transform continuous variables, models, or functions into a discrete form by Discretization mapping. Once each gene expression level has been labelled, we will focus on the Interaction between every pair of genes. Then the interactions are filtered using a filtering function. The result of the Gene-Gene interaction matrix is imported into the network visualization and analysis tool, Cytoscape.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

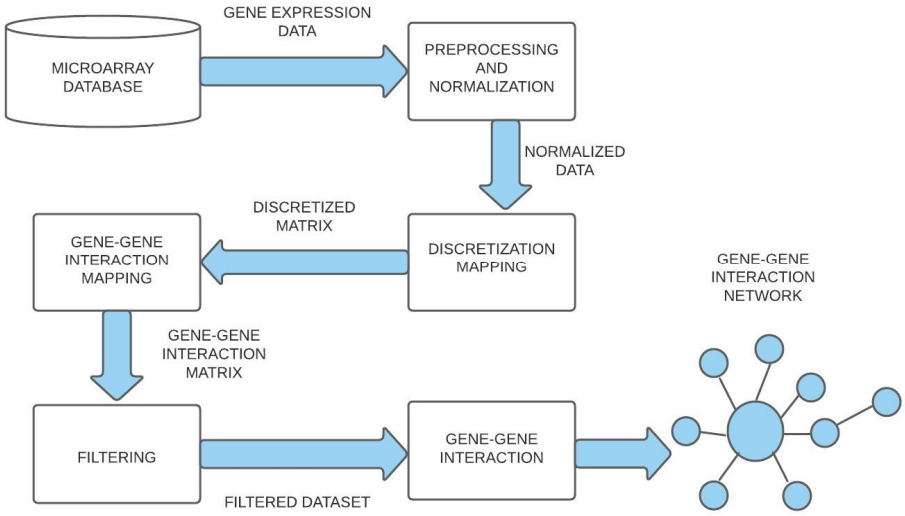


Fig 1: Architecture Diagram

A. Preprocessing and Normalization

In reference to the architecture design in Figure 5.1, the first step in our approach is to read the gene expression data. This gene expression data consists of empty cells and null values which can be removed by making use of pre-processing and normalization process i.e., the second step. The normalization process is carried out using MATLAB. The data set is quite huge and a lot of the data corresponds to genes that do not show any significant changes during the experiment. To make it easier to find these interesting genes, the first thing to do is to reduce the size of the data set by discarding genes with expression profiles that do not show anything that is of interest. There are 6400 expression profiles. There are a number of techniques to reduce this to some subset that contains the most significant genes. If you traverse through the list of genes, you will see several spots marked as EMPTY. These are empty spots on the array, and while they might have data associated with them, these points can be noise. In the expression profile, we notice several places where the expression level is marked as NaN, which denotes that no data was been collected for this gene at the particular time step. One approach to dealing with these empty values would be to impute them using the median or mean of data for that particular gene over time. This example uses a less rigorous method of simply flushing away the data for any genes where one or more expression level was not measured. The function isNan( ) is used to detect the genes with missing data and commands are used to flush the genes with missing data.

B. Discretization Mapping

The first step is to transform the gene expression level matrix into a discretized matrix by using the discretization mapping α, which is defined over a three symbol alphabet Ω = I, M, E. To carry out this discretization, we need to establish an alphabet Ω, which is used to deliver labels for the mapping, and a mapping function α, which is used to change labels from Ω to the numerical values. The definition of Ω and α is given by the user: characters for Ω and a discretization mapping table for α, in which the user can also make use of symbols ∞, μ and σ, standing for infinite, mean, and standard deviation. Segmentation is completed by discretizing the range of values. In this way, different labels are acquired according to the gene expression level. However, the discretization is local i.e., the same expression range for two different genes might transform into different labels. If the gene expression level is in (-∞,μ-σ) then the label "I" is given(inhibited); if it is in [μ-σ,μ+σ], then the label is "M" (medium); and finally, if it is in (μ+σ,+∞), then "E" (expressed)

C. Gene-Gene Interaction

Once each gene expression level has been labeled according to the mapping function, we will focus on the interaction between every pair of genes. Firstly, a new alphabet Π is required to allocate a label to any possible combination of gene pairs. In general, the size of the set Π is, at maximum, the square of the size of the set Ω, although usually needs to be lower.

In the second step, the gene-gene interaction mapping has combinations, but the magnitude of the alphabet Π corresponds to {Z, S, P, N, Q}. Where Z stands for null, S for similar, P for positive, N for negative, and Q for both expressed. The interaction mapping function β is also defined by the user, as a mapping table. Once the discretized matrix is constructed, we then check for the interaction between genes. As the microarray has M genes and N experiments, for each gene M-1 interactions with the remaining genes are needed. There will be M*(M-1) interactions. The new matrix M" encodes the information of all possible interactions, although not everyone might be interesting.

Table I
GENE-GENE INTERACTION RULES

Gene 1	Gene 2	Value
I	I	Z
I	M	S
I	E	P
M	I	S
M	M	S
M	E	S
E	I	N
E	M	S
E	E	Q

D. Filtering Function

The fact that two genes are inhibited under most or all of the experimental states has no biological significance. Therefore, this situation can be easily discarded. When two genes are both expressed under nearly all or all the experimental conditions that might have a biological explanation. In fact, many types of research only focus on this aspect of the interaction expressed–expressed. In this work, we are also interested in other cases, such as example, when many of the time an inhibited gene is related to an expressed gene, and vice-versa. And this situation is significantly interesting when the complementary is true as well, i.e., if gene1 is expressed then gene2 is inhibited and if gene1 is inhibited then gene2 is expressed. The last situation is more difficult to detect and is one of the main goals of this work. Depending on the value of the filtering function, we find which genes are related to other genes. This gives some indications about the strength of interactions and gives us a specific criterion for each gene regarding the remainder. The value which is provided by the filtering function might be different for each gene.

E. Constructing the Gene Regulatory Network

The result of the Gene-Gene interaction matrix is imported into the network visualization and analysis tool, Cytoscape. Cytoscape's roots are in Systems Biology, where it is used as a tool for integrating bimolecular interaction networks with high-throughput expression data and other molecular state data. Although Cytoscape is applicable to any system of molecular components and interactions, it is most powerful when it is used in conjunction with massive-sized databases of protein-DNA, protein-protein, and genetic interactions that are increasingly accessible for humans and model organisms. Cytoscape allows the visual aggregation of the network with expression profiles, and other molecular state information, and links the network to databases of functional annotations. The central organization structure of Cytoscape is a network or a graph, with genes, molecules, and proteins represented as nodes and interactions between them represented as links, i.e. edges, between the nodes.

IV. RESULTS

The final gene-gene interactions are then represented in a visual format as Gene regulatory network using Cytoscape. The nodes in the network are the genes and the edges connecting them are the interactions between them. Using the Cytoscape tool we can perform an analysis of the network to find the degree of interactions of each gene, which helps us determine the genes that have more possibilities of being responsible for colon cancer.

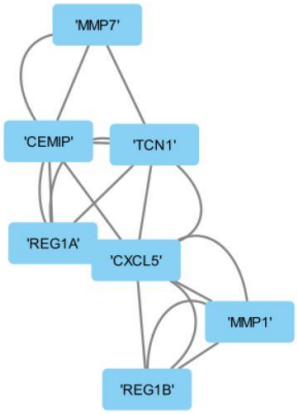


Fig 2: Gene Regulatory network for 10 genes

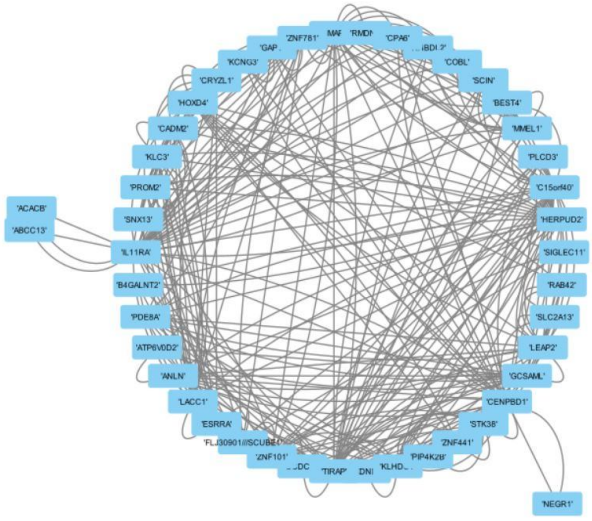


Fig 3: Gene Regulatory network for 50 genes

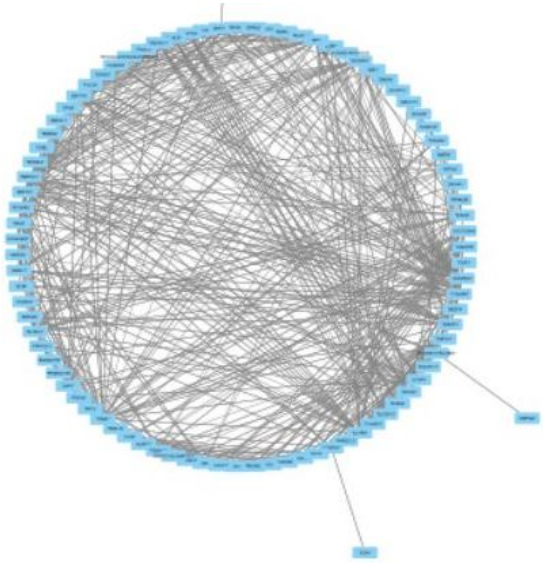


Fig 4: Gene Regulatory network for 100 genes

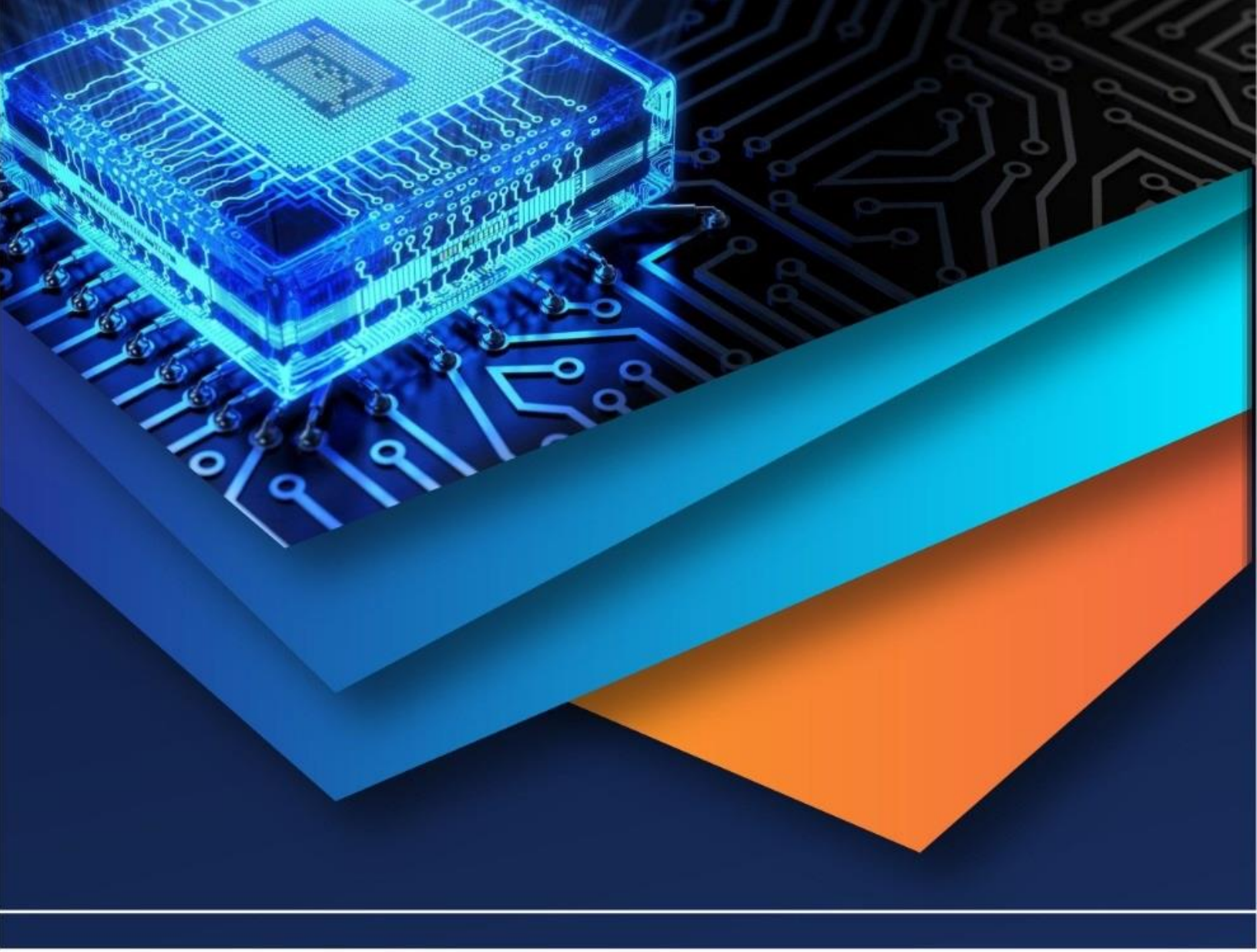
V. CONCLUSION AND FUTURE WORK

The disruption of the gene regulatory networks is what causes the intricate molecular interactions that underlie cancer. The first stage in diagnosing cancer is to identify the genes that are malignant and the pathways that they influence through the gene regulatory network. A directed regulatory network can effectively show interactions between gene pairs. In order to recreate gene regulatory networks under particular illness settings, this research provides a straightforward method to extract differentially expressed Genes and detect associations between gene pairs. This work is focused in designing the reconstruction of GRN using the mRNA gene expression profiles from GRN such that the development of new model is done. This work can be used by geneticists, oncologists and biochemists to conduct further analysis and deeper understanding of gene regulatory networks and interaction between genes. The current work included the transcription part and the future works in our project include the post translation, a process called protein synthesis where reading of RNA code is carried out by the ribosomes to make proteins.

REFERENCES

[1] Dimitrakopoulos, G.N. XGRN: Reconstruction of Biological Networks Based on Boosted Trees Regression. Computation 2021, 9, 48.
[2] K. Raza and R. Jaiswal, "Reconstruction and Analysis of Cancer specific Gene Regulatory Networks from Gene Expression Profiles," International Journal on Bioinformatics & Biosciences, vol. 3, issue 2, pp. 25-34, 2013. arXiv:1305.5750v1, 2013.
[3] Alvis Brazma,Jaak Vilo, "Gene expression data analysis", FEBS Letters 480 Issue1.Functional Genomics,24 August 2000.

[4] Norberto Díaz–Díaz, Domingo S. Rodríguez–Baena, Isabel Nepomuceno, and Jesús S. Aguilar–Ruiz, "Neighborhood-Based Clustering of Gene-Gene Interactions" BioInformatics Group Seville; Seville and Pablo de Olavide University. Spain.
[5] Cho, S.B. Estimation of Gene Regulatory Networks from Cancer Transcriptomics Data. Processes 2021, 9, 1758.
[6] M. D. Frank Emmert-Streib and B. Haibe-Kains, "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks," in Frontiers in Cell and Development Biology, 2014; 2: 38, NCBI, 2014.
[7] Noor, A.; Serpedin, E.; Nounou, M.; Nounou, H.; Mohamed, N.; Chouchane, An overview of the statistical methods used for inferring gene regulatory networks and protein-protein interaction networks. Adv. Bioinform. 2013, 2013, 953814.
[8] Xing, L.; Guo, M.; Liu, X.; Wang, C.; Zhang, L. Gene Regulatory Networks Reconstruction Using the Flooding-Pruning Hill Climbing Algorithm. Genes 2018, 9, 342.
[9] "X. Wang & O. Gotoh, (2010). "Inference of Cancer-specific gene regulatory networks using soft computing rules", Gene Regulation and Systems Biology, vol. 4, pp. 19–34.
[10] Maetschke, S.R.; Madhamshettiwar, P.B.; Davis, M.J.; Ragan, M.A. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. Brief. Bioinform. 2014, 15, 195–211.
[11] Marbach, D.; Prill, R.J.; Schaffter, T.; Mattiussi, C.; Floreano, D.; Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. Proc. Natl. Acad. Sci. USA 2010, 107, 6286–6291.
[12] W. Jiang and X. Li, "Constructing disease-specific gene networks using pair-wise relevance metric: Application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements," in BMC Systems Biology volume 2, Article number: 72 (2008), 2008
[13] Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R.D.; Califano, A. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform. 2006, 7, S7.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY