2010 International Conference on Communication, Computers and Devices, Kharagpur, INDIA December 10-12
PAPER IDENTIFICATION NUMBER: 218

Determination of Genetic Network from Micro-Array Data using Neural Network Approach

Soumya Kanti Datta
C.C.S. Department
Institut Eurecom
Sophia Antipolis, France
dattas@eurecom.fr

Srirupa Dasgupta
Lecturer, I.T.
GCELT, Saltlake
Kolkata, India
srirupadasgupta@rediffmail.com

Sounak Mitra
E.C.E. Department
TICT, Rajarhat
Kolkata, India
sounak_mitra06@yahoo.co.in

Dr. Goutam Saha
Assistant Prof, I.T.
GCELT, Saltlake
Kolkata, India
dr_goutamsaha@yahoo.com

***Abstract* - This paper focuses on a simple method for finding out the set of genes which have very significant contribution in case of 'Burkholderia Pseudomalli' for its growth and then find out the genetic network formed by them. Computation for this purpose has been carried out using Microarray gene expression time series dataset of 'Burkholderia Pseudomalli' bacteria at its various phases of growth. The dataset has been obtained from GEO data base of NCBI website. This Microarray data set represents the external manifestation of internal genetic activity resulting into genetic network. From the 5289 by 47 genetic time series data, efforts were made to detect the responsible gene set which has actively participated in the growth activity of the bacteria and the genetic network thereof. Here, 'fidelity matrix' approach has been adopted to reduce dataset by detecting the most responsible gene for the purpose. From this set of responsible gene sets a suitable genetic network searched using Artificial Neural Network (ANN). Out of the many possible genetic networks derived from running ANN several times, only the network, common to all the obtained networks, is chosen. Once the responsible gene set along with its network is determined, it can lead to further investigations on metabolic pathway engineering, drug discovery etc.**

Keywords – Micro-array data, Fidelity Matrix, Perceptron Learning Model

I. INTRODUCTION

Every living species in this earth is directed by the genetic code it inherits. Genes directly encode proteins that make up the cell to function properly. The genes are first transcribed into RNAs and are later translated to proteins. A required group of proteins park themselves on the promoter region of a gene to switch it on or off or regulates the protein production rate. Thus each gene is influenced by the other gene / genes and sometimes itself through the regulatory proteins. Thus inside the cell, expression level of the working genes always changes with time. Working genes mean among the billions of genes only 3 to 4% of them take part in the protein formation activity or the cell maintenance activity.

They are called the ESTs i.e. Essential Sequential Tags. Clearly, these ESTs form a network in the timeframe. Thus it is understood that the state of translation and transcription inside the cell is continuously updated from one state to another just like a feed-forward network in which the previous state determines the next state. The whole process can be conceptualized as a network where all genes and their products actively participate in a regulatory event. Such network is called a genetic network. Genetic network will vary for the same species in different situations. Discovery of the hidden genetic network is the essence of discovery of life activity in a species.

Expression level of ESTs can be measured by performing an experiment called 'Micro-array data analysis' is available in the website www.ncbi.nlm.gov.in for public domain. Research and development in bioinformatics has made it possible to generate data that reflect the status of entire cell at an instant of time and provides a detailed understanding about the genetic network of a cell. A set of such data taken at specific intervals of time, frames a time series of protein amounts. Such kind of time series data acts as raw material for designing mathematical models of genetic network. Genetic networks provide knowledge about functional pathway in a given cell, representing processes such as metabolism, gene regulation, transport, and signal transduction. Any disorder taken place in the genetic network due to mutation or some other cause may lead to diseases or behavioral disorder in the living species.

Several approaches have been made to model genetic networks such as application of clustering algorithms **[1]**, Boolean networks, Bayesian networks **[2], [3]** and the use of continuous models **[4], [5], [6]**. In this work, artificial neural network has been used to

Indian Institute of Technology Kharagpur, INDIA

identify the genetic network. Computation for this process has been carried out using the Microarray time series data of "**Burkholderia Pseudomalli**". The dataset obtained from the GEO database of www.ncbi.nlm.gov.in and consists of activity values of 5289 genes at 47 instants of time. The whole work can be broadly classified into two phases: data reduction and search for genetic network using neural network. As computation with the huge Microarray data is very difficult, efforts have been made to work out the responsible genes actively taking part in the overall bacterial growth process. Second phase of the work employs artificial neural network on the Microarray data of the resulting genes in the data reduction phase. From such genetic network, proteins and genes responsible for a particular disease can be found out. This paves the way for 'drug designing' **[9]**.

II. DATA REDUCTION

Conventional K-means clustering algorithm **[10], [11], [15], [16], [17]** is a very useful algorithm for data reduction. This algorithm clusters similar type of data into a cluster. But in this work the concept of fidelity matrix has been adopted. The reason behind choosing fidelity matrix method is that the process is very close to the actual biological process taking place inside a cell. The purpose of data reduction in this work is twofold. Firstly, size of the Microarray dataset is 5289 by 47 and manipulating such huge amount of data is impossible. Secondly, this simplification process identifies the most responsible genes. The steps of the data reduction can be summarized as below:

- **Representation of Microarray data**: The Microarray data represents the expression value of 5289 genes at 47 time instants. A matrix **mat[5289, 47]** is formed where each row and column represent a gene and a time instant respectively. Thus mat[i, j] [for i = 1(1)5289, j=1(1)47] corresponds to the expression value of i^{th} gene at j^{th} instant of time.
- **Elementary column operation**: All the elements of the column k are subtracted from the corresponding elements of column k+1 and the result is stored in column k [for k=1(1)46]. This operation finds the change in expression of all the genes at successive time instants. If the resulting value is very near to zero then it is understood that the gene contribute very little over that time interval in the growth process and vice versa. The last column of the original matrix mat[5289, 47] is discarded thereby reducing the size to 5289 by 46 as the last column retains the originals values.
- **Deviation from mean:** The average expression value of each gene is calculated. For each gene, the average is subtracted from all the corresponding gene expressions.
- **Thresholding of the fidelity matrix:** The absolute value of each of the elements of 5289 by 46 matrix is calculated. This is the required 'fidelity matrix'. Then each expression value is compared with a threshold of 0.78 such that any expression less than threshold is reduced to zero or else that is kept as it is.
- **Obtaining the contributing genes:** The genes with only zeros as their expressions are discarded and rest are collected in another array. This manipulation results in 25 contributing genes. Thus the operation segregates 25 most contributing genes from the 5289 genes.

Thus data reduction is achieved and 25 genes which are most responsible for the actual bacterial growth are worked out. The gene numbers are as follows: 3, 60, 129, 502, 560, 934, 962, 991, 1161, 1193, 1221, 1277, 1552, 1868, 2552, 2563, 2598, 2722, 2914, 3123, 3380, 4216, 4220, 4417, and 4784. In the next phase of the work, neural network is applied on the expression values of these 25 genes at two particular time spans.

III. SEARCH FOR GENETIC NETWORK USING ARTIFICIAL NEURAL NETWORK

Genetic network models **[18]** are worked out in order to extract the 'gene regulation matrix' that describes which gene(s) regulate(s) which gene(s) and what are the effects of environmental inputs to such network **[13], [14]**. The work is based on the assumption that the regulatory effect on a particular gene expression data can be expressed by neural network (Fig. 1). Each node of the network represents a particular gene and regulatory interactions among the genes are given by the wiring among the nodes.

Each layer of the network represents the level of expression of genes at an instant of time, say t. in principle, in a fully connected network, all genes can control all other. In reality a gene is regulated by only a few genes. The result of the work justifies the statement. The state of the entire network is updated in every instant of time. The state of a gene expression at current instant is determined by the gene expression(s) of the previous instant. Thus the output of a node at instant $t+\Delta t$ is calculated based on the expression levels of the genes at time t, $[x_j]$ and the connecting weights $[w_{ij}]$ among the genes. If we consider g_i to be the regulatory effect on gene i, then g_i can be given by the following equation,

$$g_i = \sum_j x_j \, . \, w_{ij} \qquad (1)$$

This has the advantage of replacing the gene interactions with a weight matrix and consequently the matrix is available for mathematical manipulation. Since the wiring between the nodes represent the gene regulatory interactions and the wiring corresponds to the weights (w_{ij}) between two genes, our interest lies with the weights of the trained network. To achieve the purpose, Perceptron Learning Model **[21], [22], [23]** has been chosen for the work. Perceptron Learning Model employs a learning law for the weight adaptation in the McCulloch-Pitts model **[20]**. The work presented here is a typical reverse engineering work **[12]**.

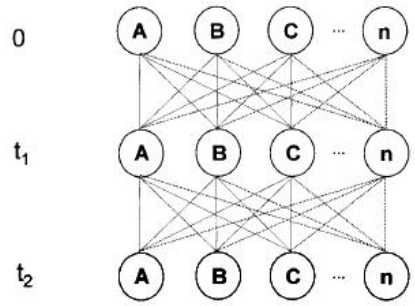

Figure 1. Artificial Neural Network

The learning principle is illustrated here.

Let, $X_k = [x_1(k), x_2(k), \ldots\ldots, x_n(k)]$ be the input vector at time instant k. The input vector comprises of the microarray data (at 10am) of the 25 genes resulted from the data reduction phase.

$W_k = [w_1(k), w_2(k), \ldots\ldots, w_n(k)]$ be the weight vector at time k. The weight vector is always initialized randomly using real numbers because that is what happens in the actual biological process.

ΔW_k be a change in the weight vector at instant k.

Y_k be the output scalar of the McCulloch-Pitts neuron at time instant k.

$D_k = [d_1(k), d_2(k), \ldots\ldots, d_n(k)]$ be the desired output at instant k. In this work, this vector consists of the microarray data (at 10:30am) of the 25 genes resulted from the data reduction phase.

e_k = error signal at the instant k.

α = the learning rate, $\alpha > 0$.

Convergence of the Perceptron Learning Algorithm is another important issue in this regard which is explained below:

We define Net_k following the McCulloch-Pitts Model as

$$Net_k = \sum x_i(k) \, . \, w_i(k) - \theta \qquad (2)$$

where, θ is a bias term. It can be shown that,

$$Net_k = X_k^T W_k \qquad (3)$$

Thus, $Y_k = f(Net_k)$

$= 1$, if $Net_k \geq 0$

$= 0$, otherwise (4)

The Perceptron learning rule which is used to adapt weights of the neural network is given by,

$$W_{k+1} = W_k + \Delta W_k \qquad (5)$$

where, $\Delta W_k = \alpha e_k X_k$ (6)

The algorithm that has been developed to find out the gene regulatory network using the Perceptron Learning Model is summarized below. In this work only one hidden layer has been used.

1. Compute the activation for the input pattern X_k.
2. Compute the error where error is the difference between the desired output and Y_k.
3. Modify the weights among the nodes using "(5)" and "(6)".
4. Repeat steps 1, 2 and 3 for each of the input pattern.
5. Repeat step 4 as long as error does not reduce to zero or become negligible for all the given input patterns.
6. Collect the input to hidden layer weights and hidden layer to output layer weights in two different matrices.

7. Compute the two standard deviations of all the elements of the two matrices.
8. The elements of the matrices are compared with a threshold of 1.2 times the corresponding standard deviations such that, an element less than the threshold is reduced to zero, otherwise it is kept as it is.
9. The non-zero weight-paths between the input to hidden layer and hidden layer to output are kept and others are truncated.
10. Thus a genetic network is obtained.

While training the neural network, the initial weights are randomly determined. Thus, at the end of each running the above mentioned steps different networks are formed. The obtained networks are compared with each other to obtain the common stable network. For this work, the stable network is obtained after 35 such iterations.

IV. RESULTS

The gene expressions at 10:30 am are influenced by genes expressions at 10am as summarized below.

TABLE I. Regulatory interaction among genes

Genes at 10:30am	**Influencing genes of 10am**
3	502, 1161, 2914
60	2552, 4784, 1552, 1868
129	4784, 1161
502	991, 2563, 2722, 1552, 2563, 1868
560	1552, 2563, 1868
934	4417
962	Switched off
991	2552, 4784, 4417, 3
1161	2598, 560
1193	502, 991, 1277
1221	1868, 4417, 560
1277	1868
1552	560, 502, 2598
1868	2563, 560, 1277, 1552, 991, 1868, 2722
2552	129, 3123, 4220, 2563, 2598, 991, 4220, 2563
2563	991, 4220
2598	2563
2722	991
2914	Switched off
3123	1193, 2563, 1868, 1277
3380	4220, 1277
4216	2598, 1868
4420	Switched off
4417	Switched off
4784	4220, 991, 2552, 1868

The actual gene regulatory network identified in this work is given below.

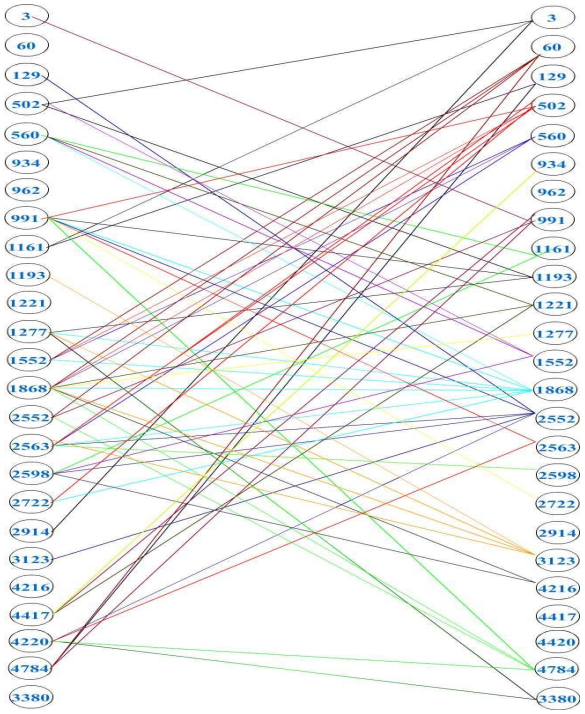

Figure 2. The identified Gene Regulatory Network

This can be inferred from the above discussion that expressions of genes numbered 502, 1161, and 2914 (at 10 am) influence the expression of gene 3 (at 10: 30 am). This means that proteins produced by the above mentioned genes at 10am sits on the promoter region of the genes 3 at 10:30 am and has affected its functionality. So on can be inferred about the other gene expressions as shown above.

V. DISCUSSION

As seen from the drawn network above that gene 991, 1868, 2563 at 10am have influenced 6, 8, 6 number of genes at 10:30am. Thus these 3 genes can be termed as most contributing genes for the concerned time spans. It has been mentioned that the microarray data has expression value of 5289 genes at 47 instants of time. The above procedure of searching for genetic network can be repeated for the other time spans to reveal the entire genetic network for the bacterial growth. The most responsible or contributing gene for the entire bacterial growth can be found out analyzing the entire network. The entire network needs

to be verified by studying metabolic pathway engineering **[19]**. It has also been studied through mathematical models that include sets of Ordinary Differential Equations (ODE). Gepasi **[7]** and DBsolve **[8]** are examples of simulation softwares for ODE cellular modeling. Qualitative analysis in metabolic pathway is very necessary as it can open new vistas in applications in biomedical engineering, biotechnology and drug designing.

From the developed genetic network further work can be carried out to modify its present metabolic pathway to fetch some important products from the bacteria that could not be done in its normal activity. This means in growth process many enzymes play important role at different instants of time. This can be clearly apprehended from the genetic network. In further modification, we can replace one enzymatic action by another one as seen from the genetic network. This may result into production of costly antigens which otherwise cannot be produced in this way. This technique has paved a way the reduction of cost of many costly medicines and leads to new areas of research in metabolic pathway engineering.

The same network can also be worked out using Bayesian Network technique. Then the resulting networks of the two processes i.e. one obtained by the method discussed in this paper and the other obtained using Bayesian Network can be compared to find out which one is more approximate of the actual genetic network which can be known from metabolic pathway engineering.

Computation with numerous sizes of the microarray data is a curse for the present hardware of the computers. Though the advent of softwares like R is reliable for such work, but it is very much time consuming. But the method of data reduction suffers from the disadvantage that it discards 5264 genes in the process of finding the more contributing genes and the effect of those genes on the entire network is not studied. This may result in considerable deviation from the original and actual genetic network of the mentioned bacteria. Investigation should be done to establish a hardware that will facilitate working with this huge amount of data generated in bioinformatics' experiments. The technological advancements in VLSI industry and EDA Tools will be a definite help for realizing such hardware.

REFERENCES

[1] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome – wide expression data. Proc. Nat. Acad. Sci, 95(25): 14863 – 14868, 1998.

[2] N. Friedman, M. Linial, I. Nachman and D.Pe'er. Using Bayesian networks to analyze expression data. J. Comput. Biol. , 7: 601 – 620, 2000.

[3] D. Husmeier. Reverse Engineering of genetic networks with Bayesian networks. Biochem. Soc. Trans., 31: 1516 – 1518, 2003.

[4] J. C. Liao et al. Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Nat. Acad. Sci., 100(26): 15522 – 15527, 2003.

[5] J. Tegner, M. K. Yeung, J. Hasty and J. Collins. Reverse engineering gene networks: Integrating genetic perturbations with dynamic modeling. Proc. Nat. Acad. Sci., 100(10): 5944 – 5949, 2003

[6] M. K. Yeung, J. Tegner and J. Collins. Reverse engineering genetic networks using singular value decomposition and robust regression. Proc. Nat. Acad. Sci., 99(9): 6163 – 6168, 2002.

[7] Gepasi, M. P. : A software package for modeling the dynamics, steady status and control of biochemical and other systems. Comput. Appl. Biosci 9(1993) 563 – 571.

[8] Goryanin, I., Hodyman, T.C., Salkor, E. : Mathematical simulation & analysis of cellular metabolism and regulation. Bioinformatics 15(1999) 749 – 758

[9] Joshua Stender, "Microarrays to Functional Genomics: Generation of Transcriptional Networks for Microarray experiments", December 3, 2002, Department of Biochemistry.

[10] Prof. Abraham B. Korol, "Review – Microarray cluster analysis and applications", Institute of Evolution, University of Haifa, lecture thesis.

[11] Patrix D'Haeseller , Shoudan Liang and Ronald Somogyi, "Genetic Network Interface: From Co-Expression Clustering to Reverse Engineering", lecture thesis.

[12] Abhijit Choudhari,"Reverse Engineering of Genetic Networks", lecture thesis.

[13] Ron Shamir, "Analysis of Gene Expression data, Lecture 9: May 5, 2005, Spring Semester, 2004".

[14] Benny Chor and Ron Shamir,"Analysis of Gene Expression data Lecture 9: June 13, 2002, Spring Semester, 2002".

[15] William Shannon, Robert Culverhouse, Jill Duncan, "Analyzing microarray data using cluster analysis".

[16] Famili, A., Liu, Z., "Evaluation and Optimization of Clustering in Gene Expression Data Analysis", Journal of Bioinformatics, 2003, Oxford University Press(in press).

[17] Jennifer S. Hallinan, "Cluster analysis of the p53 genetic regulatory network: Topology and Biology". IEEE conference 2004

[18] "Finding Genetic network using Graphical Gaussian Model" Abhishek Bag, Bandana Barman, Dr. Goutam Saha, accepted in **ICIIS 2008** held in IIT, Kharagpur in December, 2008; page 56 to 63

[19] Niranjan Baisakh and Swapan Datta, "Metabolic Pathway Engineering for Nutrition Enrichment", chapter 19. Plant breeding, Genetics, Biochemistry division, International Rice Research Institute, Philippines.
[20] McCulloch, W.S. and Pitts, W., "A logical calculus of the ideas immanent in the nervous acitivity, *Bull. Math. Biophys*., vol. 5, pp. 115 – 133, 1943.
[21] Minisky, M. and Papert, S., *Perceptrons,* MIT Press, Cambridge, 1988.
[22] Rosenblatt, F., *The Perceptron: a perceiving and recognizing automation*, Technical Report 85-460-1, Cornell Aeronautical Laboratory, 1957.
[23] Rosenblatt, F., *The Perceptron: a probabilistic model for information storage in the brain*, Psych. Rev., vol. 65, pp. 365-408, 1958.