A CLONAL BASED ALGORITHM FOR THE RECONSTRUCTION OF GENETIC NETWORK USING S-SYSTEM

Jereesh A S [1], Govindan V K [2]

[1] Research scholar, [2]Professor, Department of Computer Science & Engineering, National Institute of Technology, Calicut, Kerala, India, jereesh.a.s@gmail.com, vkg@nitc.ac.in

Abstract

Motivation: Gene regulatory network is the network based approach to represent the interactions between genes. DNA microarray is the most widely used technology for extracting the relationships between thousands of genes simultaneously. Gene microarray experiment provides the gene expression data for a particular condition and varying time periods. The expression of a particular gene depends upon the biological conditions and other genes. In this paper, we propose a new method for the analysis of microarray data. The proposed method makes use of S-system, which is a well-accepted model for the gene regulatory network reconstruction. Since the problem has multiple solutions, we have to identify an optimized solution. Evolutionary algorithms have been used to solve such problems. Though there are a number of attempts already been carried out by various researchers, the solutions are still not that satisfactory with respect to the time taken and the degree of accuracy achieved. Therefore, there is a need of huge amount further work in this topic for achieving solutions with improved performances.

Results: In this work, we have proposed Clonal selection algorithm for identifying optimal gene regulatory network. The approach is tested on the real life data: SOS Ecoli DNA repairing gene expression data. It is observed that the proposed algorithm converges much faster and provides better results than the existing algorithms.

Index Terms: Microarray analysis, Evolutionary Algorithm, Artificial Immune System, S-system, Gene Regulatory Network, SOS Ecoli DNA repairing, Clonal Selection Algorithm.

--***---

1. INTRODUCTION

DNA microarray is a modern technology, which is used to analyze the interactions between thousands of genes in parallel [7]. Exploiting the hybridization property of CDNA, the transcript abundance information is measured in microarray experiment. Microarrays have numerous applications. A particular set of genes are activated for a particular condition. Identification of activated genes will be useful for recovering or activating the conditions artificially. Even though the technology is well developed, direct biological methods available for finding gene expression are complex. Analysis of protein expression data is very expensive due to the complex structures of proteins.

Microarray data analysis involves methodologies and techniques to analyze the data obtained after the microarray experiments. The major part of the microarray data analysis is the numerical analysis of normalized data matrix. Gene expression analysis is a large-scale experiment, which comes under functional genomics. Functional genomics deals with the analysis of large data sets to identify the functions and interactions between genes [24]. A set of algorithms and methods are defined for the analysis of microarray data. There

is a tradeoff between the time and accuracy for using an algorithm for analyzing the microarray data.

Gene Regulatory Network (GRN) is a network of set of genes, which are involved, in a particular process. In GRN, each node represents gene and links between genes define the relationships between those genes. Gene regulatory network is the network based approach to represent the interactions between genes. The expression of a particular gene depends upon the biological conditions and other genes. Gene microarray experiment identifies the gene expression data for a particular condition and varying time periods. Identifying such network will lead to various applications in biological and medical areas. Objective of this paper is to propose a new method, which leads to substantial improvements in processing time and accuracy. High dimensionality of the microarray data matrix makes the identification of GRN complex. In this paper optimization of S-system model using artificial immune system is proposed.

The rest of this paper is organized as follows. A brief survey of some of the existing work is given in Section 2. Section 3 presents the mathematical model used for the modeling of gene regulatory network and algorithm for the optimization

process. Section 4 describes the experimental setup and compares the results of the new proposal with the existing approach. Section 5 is a discussion based on the results obtained by the proposed method on the real life data set called SOS Ecoli DNA repairing gene expression. Finally, the paper is concluded in Section 6.

2. LITERATURE SURVEY

There have been several mathematical models applied for the gene regulatory network reconstruction. One of the basic mathematical models identified was based on Random Boolean Network [1]. According to this model, the state of a particular gene will be either in on or off state. The state space for Boolean network is 2^N where N is the number of genes in microarray. This model gives the information about gene states, but does not provide expression levels of genes.

Zhang et al. [25] suggested Bayesian network model based on joint probability distribution. This model uses DAG (Directed Acyclic Graph) structure for modeling. Since the gene regulatory network is having the property of cyclic dependency between gene nodes, this type of model is not efficient for inferring gene network.

Another important work [17] proposed is the modeling of Gene regulatory network using ANN (Artificial Neural Network) with the standard back propagation method. The number of inputs and outputs required for this model is N, where N is the number of genes in microarray data set. The structural complexity of ANN model will increase as the number of genes increases; hence, this model is not efficient for large data sets.

Reverse engineering using the evolutionary algorithms can be applied for solving the optimization problems. Genetic algorithm is one of the major evolutionary algorithms that can be used to construct the gene network. Spieth et al. [21] proposed a memetic inference method for gene regulatory network based on S-system. This is a popular mathematical model proposed by Savageau [20]. The memetic algorithm uses a combination of genetic algorithm and evolution strategies [21].

A multi objective phenomic algorithm proposed by Rio D'Souza et al. [8] is an advanced method, which concentrates on multiple objectives like Number of Links (NoL) and Small World Similarity Factor (SWSF). Rio DSouza et al. in [9] proposes an Integrated Pheneto-Genetic Algorithm (IPGA), which makes use of the approach of S-system model [20] with memetic algorithm proposed by Spiethet al. [21]. The memetic algorithm [21] makes use of genetic algorithm to identify the populations of structures of possible networks. For N genes, out of N combinations of solutions, GA is used to identify the best solution by optimizing the error or fitness value. Memetic algorithm is a superior method than the existing evolutionary

algorithms such as standard evolutionary strategy and skeletalizing (extension of standard GA) for the particular problem [21]

Nonetheless, the above algorithms are standard algorithms, the tradeoff between time, space and accuracy factors of the algorithms are still issues need to be addressed. In this paper, we make a new proposal to optimize the model parameters for the reconstruction of gene network for achieving improved performance.

3. PROPOSED METHOD

3.1 MODEL

S-systems are a type of power law formalism, which was suggested by Savageau [20] and defined as follows.

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{N} x_j(t)^{G_{i,j}} - \beta_i \prod_{j=1}^{N} x_j(t)^{H_{i,j}} \qquad (1)$$

Where G_{ij} and H_{ij} are kinetic exponents, α_i and β_i are positive rate constants and those values are optimized using Evolution strategies. According to the S-system equation [1], $2N*(1+N)$ values are to be optimized for each individual in a population, where N is the total number of genes in a microarray data set. We propose to employ an optimization technique known as Clonal selection algorithm, which is faster than the genetic algorithm. Clonal selection algorithm is a technique used in artificial immune systems. A brief description of artificial immune system and Clonal selection algorithm is given in the following:

3.2 ARTIFICIAL IMMUNE SYSTEM (AIS)

Artificial Immune System is based on the theory of biological immune system. In biological immune system, the foreign materials, which are trying to intrude the body, will be identified and prevented. These foreign materials are called pathogens. Each pathogen has molecules called antigen which will be identified by the antibody. There are two types of immune systems in body called innate immune system and adaptive immune system [2]. Innate immune system is a static method, which is generic to all bodies. These are the basic level of protection from pathogen [6]. Adaptive immune systems are self-adaptive natured immunities, which work with the antigens. This type of immunity remembers previous attacks and strengthens the immunity process. In artificial immune system, the principles of biological immune system are used to solve the various computational problems. Clonal selection is one of the theories, which explain the process of immunity.

3.3 CLONAL SELECTION ALGORITHM

The response of immune system to infection explained by Burnet is a well-known theory in immunology [4]. In this

work, Clonal selection is used to explain the processing of adaptive immune system to antigens. In 2002, Castro and Zuben proposed a Clonal based algorithm called CLONALG [6]. Clonal selection algorithms follow the biological adaptive immune system, which consists of antibodies and antigens [2]. This type of algorithms considers solution set as antibody. The set of antibodies is called as population. At each generation selection, cloning, affinity maturation and reselection are happening to the population and trying to generate new population with better affinity. In this algorithm, affinity is calculated with the help of fitness value. As there are no recombination/ crossover steps in Clonal selection algorithm, it is faster than the genetic algorithm and hence the basic Clonal selection algorithm is used to optimize the S-systems model. The Clonal algorithm for the optimization of the S-systems model is given below:

Algorithm 1: CLONAL based Algorithm
Require: Max N of Generation; error tolerance
Ensure: Optimal antibody

   a.   Start.
   b.   Generation: = 0
   c.   Pop(Generation) := Init(Clonal pop)
   d.   Evaluate_Fitness (Pop (Generation))
   e.   while termination criteria not met do
      i.   Selected_Pop(Generation):=Selection(Pop(Generation))
      ii.   Cloned_Pop(Generation):=Clone(Selected_Pop(Generation)
      iii.   Pop(Generation):=Maturation(Cloned_pop(Generation))
      iv.   Evaluate_Fitness (Pop(Generation)
      v.   Pop(Generation+1):=Re_Selection(Pop(Generation))
      vi.   Generation := Generation + 1
   f.   end while
   g.   Stop.

Fitness function: The proposed method uses the following fitness function proposed by Tominaga et al. [23]:

$$f = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{X_{i,t}^{cal} - X_{i,t}^{exp}}{X_{i,t}^{exp}} \right)^2 \qquad (2)$$

Where $X_{i,t}^{cal}$, $X_{i,t}^{exp}$ are the expression value of gene i at time t from the estimated (calculated) and experimental data respectively.

4. EXPERIMENTAL SET UP AND RESULTS

For the experimentation, the standard artificial gene regulatory network, given in Table 1, used by various researchers [12, 13, 14, 16, 18, 19] is made use of. This network consists of 5 genes. The Runge-kutta algorithm is used to infer standard microarray data using the S-system model [13]. In order to confirm the ability of proposed method to infer the gene

regulatory network we generated 10 sets of expression data artificially. Initial values of these sets are randomly generated in the range [0, 1] as shown in Table 2.The 10 sets of time series data are obtained using equation(1) and S-system parameters given in Table 1,with T=11 and G=5; so totally 10*11*5=550expression values are observed. A sample Time dynamics of the 5 dimensional regulatory system inferred is shown in Fig.1where duration of 0.0 to 0.5 is divided into 11 equi-distance samples, and 10 points are computed between each sampling point.

In order to confirm the effectiveness of the proposed model, both the proposed algorithm and the standard memetic algorithm have been implemented and applied to a standard artificial genetic network [12, 13, 14, 16, 18, 19]. Since these algorithms are stochastic in nature, we have to test on multiple data sets for the experiment. After computing the model parameters, the microarray data set is regenerated and compared with the original. We have used350000 fitness evaluations in the comparative study. Mean Squared Error (MSE) [23] is used as the error evaluation measurement metric.
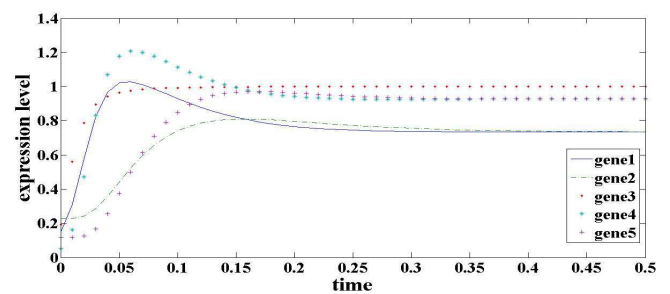

Fig. 1: A sample Time dynamics of the 5-dim regulatory system using parameters in Table 1.

Fig. 2 shows comparison of average error (MSE) versus fitness evaluation courses obtained for memetic and proposed method for 3.5 lakhs fitness evaluation. Since memetic algorithm uses genetic algorithm for the optimization purpose, over all error will be reduced after some iterations. In memetic algorithm, S-system parameters are optimized for the reconstruction of gene regulatory network. In this algorithm for each generation in genetic algorithm, evolutionary strategy with covariance matrix adaptation (CMA) has to be performed. Evolutionary strategy is a local optimal evolutionary algorithm, which is much similar to genetic algorithm. Due to hybrid nature of the algorithm, huge amount of computation is required for the processing. For the memetic algorithm, convergence happens after 20 lakhs fitness evaluations [21]. The proposed method converged after 3.5 lakhs fitness evaluations whereas, at this point, standard memetic algorithm is far away from convergence. Hence, it is also observed that the proposed algorithm converges much faster than the existing memetic algorithm

.

Table 1: S-system model parameters for the target network model [12, 13, 14, 16, 18, 19]

i	α_i	$G_{i,1}$	$G_{i,2}$	$G_{i,3}$	$G_{i,4}$	$G_{i,5}$	β_i	$H_{i,1}$	$H_{i,2}$	$H_{i,3}$	$H_{i,4}$	$H_{i,5}$
1	5.0	0.0	0.0	1.0	0.0	-1.0	10.0	2.0	0.0	0.0	0.0	0.0
2	10.0	2.0	0.0	0.0	0.0	0.0	10.0	0.0	2.0	0.0	0.0	0.0
3	10.0	0.0	-1.0	0.0	0.0	0.0	10.0	0.0	-1.0	2.0	0.0	0.0
4	8.0	0.0	0.0	2.0	0.0	-1.0	10.0	0.0	0.0	0.0	2.0	0.0
5	10.0	0.0	0.0	0.0	2.0	0.0	10.0	0.0	0.0	0.0	0.0	2.0

Table 2: Initial expression values for 10 Data sets

| Data sets Genes	Set 1	Set2	Set3	Set4	Set5	Set6	Set7	Set8	Set9	Set10
G1	0.8231	0.2851	0.9961	0.9991	0.7937	0.1479	0.6264	0.9556	0.6724	0.4216
G2	0.3933	0.2586	0.0400	0.0770	0.5441	0.2278	0.9497	0.6866	0.2542	0.6126
G3	0.6273	0.4616	0.5457	0.9494	0.8954	0.1921	0.3645	0.9983	0.3055	0.7605
G4	0.5855	0.2377	0.0971	0.0282	0.9090	0.0518	0.4206	0.7768	0.6902	0.5935
G5	0.5401	0.8144	0.8121	0.6938	0.6359	0.1169	0.9943	0.3467	0.5378	0.5618
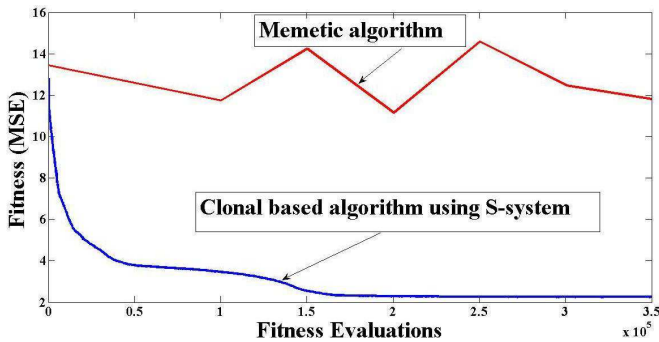
Fig. 2: Comparison of average error (MSE) obtained for memetic algorithm and the proposed approach; the proposed algorithm converges at about 3.5 lakhs fitness evaluations.

5. DISCUSSION

5.1 ANALYSIS OF REAL LIFE DATA USING THE PROPOSED METHOD

In order to assure the performance of a method, it should be evaluated on a real life data. We employed a famous real life dataset called SOS DNA repair system in E.coli [22] to study the performance of the proposed method. Fig.3 graphically describes the interactions during the repairing of DNA of E.coli., when DNA damage is occurred. According to this system, when a damage happens immediately RecA protein will identify the damage and will invoke the processing of cleavage of LexA protein without any help of enzymes. Thus, the concentration of LexA will be decreased. Due to reduction of LexA other proteins in the SOS system will activate the repairing process of the DNA. LexA protein is acting as a repressor in the system. After the repairing of DNA, concentration of RecA will be dropped; in effect, automatic cleavage of RecA will stop. Finally, concentration of LexA will increase and repress the other genes. This will lead to a stable state and will continue in this state till the next damage happens.

SOS Data is obtained from the website www.weizmann .ac.il/mcb/UriAlon/Papers/SOSData/ as a result of

experiments done by Uri Alon lab of Weizmann institute of science. They have 4experimental results obtained, each of 8 proteins and 50 time points. As the first time-point represents 0 seconds all initial expression values are zeros. Since the first time-point contains no information it was removed and the remaining 49 time points were used for the modeling. From the previous literatures [3, 5, 10, 12, 15, 16] it was identified that out of 8 genes, 6 major genes (uvrD, umuD, lexA,recA, uvrA and polB) and last 2 experimental results are required for the accurate prediction of SOS Gene regulatory system. Each values of the gene expression values are normalized in the interval [0,1].
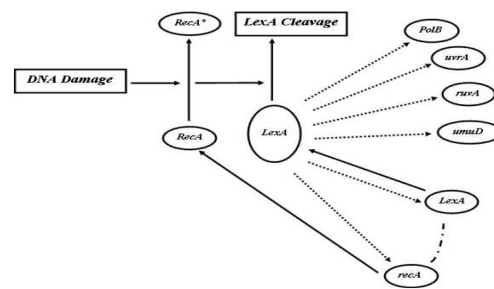

Fig3. SOS DNA repair system of E.coli.

Implementation of the proposed approach on SOS Data set inferred the gene network of Fig.4. Since in the given microarray, data is real one it is concealed with noise, and hence the accuracy of the proposed algorithm depends on the degree of noise. As the biological systems are so complex, even with the biological experiments it is difficult to extract all the hidden facts in the system. Therefore, the SOS DNA repair system of Ecoli identified in Fig. 4 may not contain all the relationships. There is still a possibility of finding new relationships. The gene network obtained by the proposed method identified inhibitions from LexA to LexA, uvrD, uvrA, recA and polB. Proposed method also identified regulations from recA to recA and recA to lexA correctly. There are also some more relations, as given in Table 3 reported by other researchers, identified by the proposed method.

Table3. Relations identified by the proposed approach that are also already identified by previous researchers

Gene	Predicted relation and the references where these are already identified
uvrD	uvrD -| uvrD(12, 5, 15, 11), uvrD -| umuDc (15), uvrD -| lexA (15), uvrD→polB (10, 11)
LexA	LexA-| LexA(3, 5, 10, 12, 16), LexA-| uvrD(5, 10, 12, 16), LexA-| recA(3, 12, 16), LexA-| uvrA(5, 10, 11, 12), LexA→ uvrA(11, 15), LexA-| PolB(5, 11, 12, 16), LexA→ PolB(11, 15),
umuDc	umuDc -| umuDc (3, 5, 15, 11), umuDc -| recA(16, 3, 11), umuDc -| polB(11), umuDc→uvrA(11), umuDc -| lexA (3, 15, 11)
recA	recA→uvrA(11), recA -| uvrA (15, 10), recA -| umuDc(12, 15, 10, 11)
uvrA	uvrA -| uvrA (16, 12, 5, 15, 11), uvrA -| recA (11), uvrA -| umuDc (16, 10), uvrA -| lexA (15, 10), uvrA→uvrD(16, 12, 10, 11), uvrA -| polB(16, 11)
polB	polB -| polB(12, 5, 11), polB→uvrD (11), polB -| recA(16, 11), polB→uvrA (11), polB -| uvrA(11)

Therefore, out of the total 33 relations, 30 relations are already proposed by previous researchers. The remaining may be the relations, which were not found yet, or false positives. Hence, it is demonstrated that the proposed algorithm can be used for the real life applications.
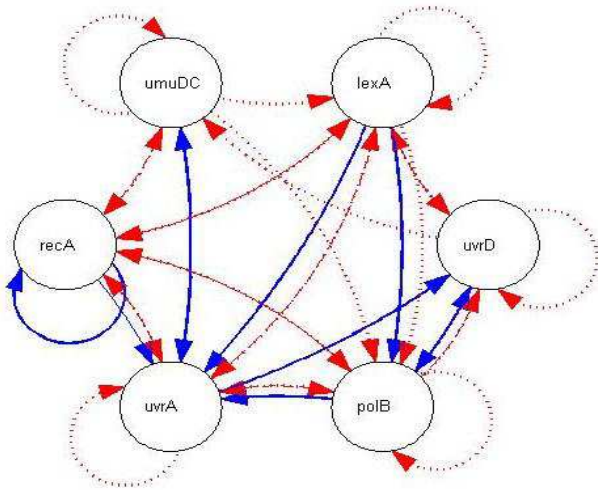

Fig4. SOS DNA repair system of E.coli. Identified by the proposed method (dashed lines indicate the inhibition and solid lines indicate the activation); 33 relations are identified.

CONCLUSIONS

Gene regulatory network reconstruction is a major issue in bioinformatics. Existing methods for GRN reconstruction either take longer computations for convergence or poor in accuracy of identifying the relations. This paper proposes a Clonal based approach using S-system model. The model parameters are computed using optimization employing the basic Clonal selection algorithm. Performance of the model is compared with the existing standard memetic algorithm and found to be superior with respect to execution time and accuracy. Convergence is achieved with much lesser number of fitness evaluations than the standard memetic algorithm. The results obtained on SOS DNA repair system of E.coli. demonstrate that the proposed approach identified most of the relations identified by the previous researchers. This amply proves that the approach is powerful and applicable to real life data.

REFERENCES

[1]. Akutsu, T., Miyano, S., Kuhara, S., et al. (1999). Identification of genetic networks from a small number of gene expression patterns under the boolean network model.In Pacific Symposium on Biocomputing, volume 4, pages 17–28. World ScientificMaui, Hawaii.

[2]. Al-Enezi, J., Abbod, M., and Alsharhan, S. (2010). Artificial immunesystemsmodels,algorithms and applications. International Journal.

[3]. Bansal, M., Della Gatta, G., and Di Bernardo, D. (2006). Inference of generegulatory networks and compound mode of action from time course gene expressionprofiles. Bioinformatics, 22(7), 815–822.

[4]. Burnet, F. (2008). A modification of jerne's theory of antibody production using theconcept of clonal selection. CA: A Cancer Journal for Clinicians, 26(2), 119–121.

[5]. Cho, D., Cho, K., and Zhang, B. (2006). Identification of biochemical networks by s-tree based genetic programming. Bioinformatics, 22(13), 1631–1640.

[6]. De Castro, L. and Von Zuben, F. (2002). Learning and optimization using the Clonal selection principle. Evolutionary Computation, IEEE Transactions on, 6(3), 239–251.

[7]. Dubitzky, W., Granzow, M., Downes, C., and Berrar, D. (2003). Introduction to microarray data analysis. A Practical Approach to Microarray Data Analysis, pages1–46.

[8]. DSouza, R., Sekaran, K., and Kandasamy, A. (2012a). A multiobjective phenomic algorithm for inference of gene networks. Bio-Inspired Models of Network, Information, and Computing Systems, pages 440–451.

[9]. DSouza, R., Sekaran, K., and Kandasamy, A. (2012b). A phenomic algorithm for inference of gene networks using s-systems and memetic search. Bio-Inspired Modelsof Network, Information, and Computing Systems, pages 229–237.

[10]. Hsiao, Y. and Lee, W. (2012). Inferring robust gene networks from expression databy a sensitivity-based incremental evolution method. BMC bioinformatics, 13, 1–21.

[11]. Huang, H., Chen, K., Ho, S., and Ho, S. (2008). Inferring s-systemmodels of genetic networks from a time-series real data set of gene expression profiles. In Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress onComputational Intelligence). IEEE Congress on, pages 2788–2793. IEEE.

[12]. Kabir, M., Noman, N., and Iba, H. (2010). Reverse engineering gene regulatory network from microarray data using linear time-variant model. BMC bioinformatics, 11(Suppl 1), S56.

[13]. Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and s-system. Bioinformatics, 19(5), 643–650.

[14]. Kimura, S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S., and Konagaya, A. (2005). Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm.Bioinformatics, 21(7), 1154–1163.

[15]. Kimura, S., Nakayama, S., and Hatakeyama, M. (2009). Genetic network inference as a series of discrimination tasks.Bioinformatics, 25(7), 918–925.

[16]. Kimura, S., Sonoda, K., Yamane, S., Maeda, H., Matsumura, K., and Hatakeyama, M. (2008). Function approximation approach to the inference of reduced ngnet models of genetic networks.BMC bioinformatics, 9(1), 23.

[17]. Narayanan, A., Keedwell, E., Gamalielsson, J., and Tatineni, S. (2004). Singlelayer artificial neural networks for gene expression analysis.Neurocomputing, 61, 217–240.

[18]. Noman, N. and Iba, H. (2007). Inferring gene regulatory networks using differential evolution with local search heuristics.IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 4(4), 634–647.

[19]. Perrin, B., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and dAlche Buc, F. (2003). Gene networks inference using dynamic bayesian networks.Bioinformatics, 19(suppl 2), ii138–ii148.

[20]. Savageau, M. (20). years of s-systems. Canonical Nonlinear Modeling. S-systems Approach to Understand Complexity, pages 1–44.

[21]. Spieth, C., Streichert, F., Speer, N., and Zell, A. (2004). A memetic inference method for gene regulatory networks based on s-systems. InEvolutionary Computation, 2004. CEC2004. Congress on, volume 1, pages 152–157. IEEE.

[22]. Sutton, M., Smith, B., Godoy, V., and Walker, G. (2000). The sos response: recent insights into umudc-dependent mutagenesis and dna damage tolerance.Annual review of genetics, 34(1), 479–497.

[23]. Tominaga, D., Koga, N., and Okamoto, N. (2000). Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. In Proceedings of the Genetic and Evolutionary Computation Conference, pages 251–258.

[24]. Wikipedia (2012). Epistasis and functional genomics— Wikipedia, the free encyclopedia. [Online; accessed 22-June-2012].

[25]. Zhang, B. and Hwang, K. (2003). Bayesian network classifiers for gene expression analysis. A Practical Approach to Microarray Data Analysis, pages 150–165.

BIOGRAPHIES

Jereesh A S received Bachelor's degree in Computer science and engineering from the Rajiv Gandhi Institute of technology Kottayam in the year 2007 and received Master's degree in Computer science and engineering (Information Security) from the National Institute of technology Calicut in the year 2010. He is currently a research scholar pursuing for Ph.D degree in the Department of Computer science and engineering at National institute of Technology Calicut. His research interests include the Bioinformatics, data mining and evolutionary algorithms.

V K Govindan received Bachelor's and Master's degrees in electrical engineering from the National Institute of technology Calicut in the year 1975 and 1978, respectively. He was awarded PhD in Character Recognition from the Indian Institute of Science, Bangalore, in 1989. His research areas include Image processing, pattern recognition, data compression, document imaging and operating systems. He has more than 125 research publications in international journals and conferences, and authored ten books. He has produced seven PhDs and reviewed papers for many Journals and conferences. He has more than 34 years of teaching experience at UG and PG levels and he was the Professor and Head of the Department of Computer Science and Engineering, NIT Calicut during years 2000 to 2005. He is currently working as Professor in the Department of Computer Science and Engineering, and Dean Academic at National Institute of Technology Calicut, India