Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 4, Number 2 (2011) pp. 221-231

http://www.ripublication.com/jcib.htm

Modelling Gene Regulatory Network from Microarray Data Using Modified Genetic Algorithm

Vineetha S.[1], Chandra Shekara Bhat C.[2] and Sumam Mary Idicula[3]

[1]Dept. of Computer Science,
Rajiv Gandhi Institute of Technology Kottayam, Kerala, 686501, India.
E-mail: svineetha@hotmail.com
[2]National Institute for Interdisciplinary Science and Technology,
Trivandrum, Kerala, 695019, India.
E-mail: chakrakody@yahoo.co.in
[3]Dept. of Computer Science,
Cochin University of Science & Technology Cochin, Kerala, 682022, India.
E-mail: sumam@cusat.ac.in

Abstract

In this paper a modified Genetic Algorithm (GA) was used to model gene regulatory network from microarray data. GA can effectively model gene regulation and interaction to accurately reflect the underlying biology. However, this approach requires several generations and much computational power to identify the interacting genes. This paper uses some statistical techniques to reduce the search space, thereby reducing the number of generations. This will allow the algorithm to run in a shorter amount of time with minimal effect on the result. Our approach was tested on a set of genes whose regulatory functions are identified using fuzzy GRN algorithm. The method derives a regulatory network structure which is consistent with the network structure obtained using feed forward neural fuzzy network and fuzzy GRN algorithm. The inferred knowledge can be used to provide guidance for the experiments by suggesting likely relations inferred from the observed data.

Keywords: Gene Regulatory Network, Genetic Algorithm, Microarray Data, Fuzzy Logic, Neural Fuzzy Logic, Correlation Analysis.

Introduction

The recent advances of genome-scale sequencing and array technologies have made it possible to monitor simultaneously the expression pattern of thousands of genes. A

major focus on microarray data analysis is the reconstruction of gene regulatory network (GRN), which aims to find the underlying network of gene-gene interactions from microarray dataset of gene expression [1]. The inference of GRN based on microarray is referred to as 'reverse engineering'. Changes in expression levels of genes across different samples provide information that allows reverse engineering techniques to extract gene regulatory features like activation and inhibition, which enable to construct regulatory relations among those genes. However, these approaches suffer several difficulties including (i) dimensionality problem of microarray datasets (ii) exponential complexity of the algorithm (iii) presence of noise in expression values.

A wide variety of soft computing techniques have been proposed to infer GRN from microarray data, such as Boolean networks[2][3], Bayesian networks[4][5], Artificial Neural Network[6][7], graphical models[8], fuzzy logic approach[9][10][11][12][13]. The reconstruction of gene network from biological perspective to computational perspective has been developed since 1999[14]. GRN modelling using Kouffman Boolean network presented by Akutsu ***et. al***. [2] assumes that a gene is either on or off. The main drawback of Boolean network lies in ignorance of the effect of gene at intermediate levels, causing information loss because of binary quantization process. In addition, Boolean networks presume that the transition between genes activation states are synchronous, which is biologically impossible. Models using Bayesian and regression networks done by Kato ***et. al***.[4] are effective in handling with noise, incompleteness and stochastic nature of gene expression data. However they are unable to represent the dynamical aspects of gene regulation. Dynamic Bayesian networks to handle the temporal information of gene regulatory network are under investigation. Woolf and Wang [13] have applied fuzzy rules to every possible combination of genes to find the activator/repressor relationship in a normalized subset of *saccharomyces cervisae* data. This approach is slow and computationally complex.

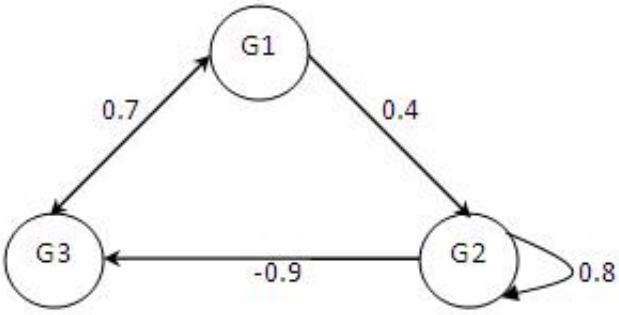
Vineetha et. al. [15] used feed forward neural fuzzy network (DNFN) to simulate gene regulatory network from the expression profile of circulating plasma RNA of colorectal cancer patients from microarray experiments. This method combines both the advantages of neural network and fuzzy logic. Combination of neural networks with fuzzy knowledge base helps to reduce the searching space and time for achieving optimal solution. Moreover no pre assignment of network structure and fuzzy rules are required since all of them are constructed via online learning.

In this paper genetic algorithm was applied to infer gene regulatory network for the plasma RNA dataset of colon cancer patients. Previous research had confirmed [16][17] that the GA can infer network structure with significant accuracy. Since GA is a probabilistic search, several generations and greater computation power were required to model smaller network with good sensitivity and precision. In this paper some feature selection techniques were used to identify a small subset of informative genes from the huge microarray experimental data. Instead of randomly initializing the genotypes, statistical techniques were used to identify the significant regulations. By using this method, the number of generations required to model regulatory network was reduced drastically.

Colorectal cancer or colon cancer (CRC) is one of the most common cancers in western countries. Discovery of genes commonly regulated in colon cancer may provide insights to the common molecular mechanism of this type of cancer. The blood of cancer patients is known to contain fragments of RNA released from the tumor [18]. Gene expression profiling of circulating plasma RNA is a valuable tool to detect cancer and a potential clinical instrument to study tumor progression and therapy responsiveness. Plasma RNA is highly attractive for cancer analysis since the sample collection is easy and reproducible, and allows reiterative extractions during treatment response [19].

Modelling Methodology

GRN modelling presented by Weaver *et.al*. [20] represent regulatory relationships between genes as linear coefficients of weights, with the "net" regulation influence on a gene's expression being the mathematical summation of the independent regulatory inputs. This representation is practical for analysis of microarray experiment, which provides snapshots of concentration at various samples. The regulatory networks generated from microarray data with this approach, display stable gene expression levels, consistent with known biological system [20]. The GRN is represented by a weighted graph G = (V,E,W), where V is the set of nodes(genes), E is the set of edges(regulatory relationships) and W is the weight matrix. Figure 1 shows an example of a gene network and the corresponding weight matrix. The value of w_{ij} is limited to a range between -1 and 1.The positive w_{ij} means gene i activating gene j, as opposed to a negative value representing an inhibition. Zero indicates no influence. The expression level of a transcriptional regulatory network containing N genes is represented by a vector $\vec{x}$. Each column of weight matrix W represents all regulatory inputs to a gene.



	G1	G2	G3
G1	0	0.4	0
G2	0	0.8	- 0.9
G3	0.7	0	0

Figure 1: A Sample Gene Network and the corresponding Weight Matrix.

The net regulatory input to a gene i, S_i is determined by taking the weighted sum of the expression level of other genes.

$$S_i = \sum_{j=0}^{n} w_{ij} x_j \quad (1)$$

The response of gene i to the regulatory input is defined using the sigmoid function as

$$x_i = \frac{m_i}{1+e^{-si}} \quad (2)$$

where m_i is the maximal expression level.

Genetic Algorithm Implementation

In this paper the modelling methodology described above was used to represent the gene regulatory network. The weighted matrix was generated and optimized using genetic algorithm. Simple regulatory networks with few numbers of genes can be inferred using this method effectively [21][16]. However, when applied to large microarray dataset this approach is slow and computationally complex. So, in order to reduce the number of generations, some preprocessing steps were done to initialize the genotypes.

Correlation analysis can be used to capture the regulation and co-regulation among the genes and have been proven useful for identifying biologically relevant groups of genes and samples [17].High correlation between gene A and B can be caused by (i) A regulates B or vice versa (ii) A and B are co-regulated by other genes (iii) there is no casual relationship just coincidence. Here regulations may be indirect, i.e. interactions through immediates. Pearson's correlation coefficient is widely used to measure the similarity between the expression patterns of two genes. Pearson's correlation coefficient views each object as a random variable with p observations and measures the similarity between two objects by calculating the linear relationship between the distributions of the two corresponding random variables. However, empirical study has shown that it is not robust with respect to outliers, thus potentially yielding false positives which assign a high similarity score to a pair of dissimilar patterns. If two patterns have a common peak or valley at a single feature, the correlation will be dominated by this feature, although the patterns at the remaining features may be completely dissimilar. This observation evoked an improved measure called Jackknife correlation .Given two data objects O_i ,O_j ,Jackknife correlation coefficient defined as Jackknife(O_i, O_j)=min{$\rho_{ij}(1)$,$\rho_{ij}(2)$, …, $\rho_{ij}(p)$} where $\rho_{ij}(l)$ is the Pearson's correlation coefficient of data objects O_i and O_j with the l^{th} feature deleted. Use of the Jackknife correlation avoids the "dominance effect" of single outliers.

Genetic algorithms are search procedures, based on the mechanics of natural genetics, able to provide robust search in complex problem spaces. For the GA implementations, gene networks have to be coded into chromosomes. Each row of the weight matrix is aligned in an array to be the one dimensional real number array chromosome. The fitness function of the GA is defined by the Euclidean error δ between the generated expression pattern and the target expression pattern.

$$\delta = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (3)$$

A Genetic Algorithm operates through a simple cycle of stages as shown in figure 2. Each cycle in Genetic Algorithm produces a new generation of possible solutions for a given problem. In the first phase, an initial population, describing representatives of the potential solution, is created to initiate the search process. For that the correlation coefficient of the target gene with every other gene is calculated. If the correlation is significant, a random number is generated to initialize the corresponding data point of each of the genotypes in the population. Depending on the fitness of the chromosomes, they are selected for a subsequent genetic manipulation process. In this paper, Roulette Wheel selection algorithm was used to breed a new generation. In Roulette Wheel selection, the individuals are given a probability of being selected that is directly proportionate to their fitness. In order to enhance the adaptability of the GA as well as the reverse engineering method, the genetic manipulation process consisting of two steps is carried out. In the first step, the crossover operation that recombines the bits (genes) of each two selected strings (chromosomes) is executed. The second step in the genetic manipulation process is termed mutation, where the bits at one or more randomly selected positions of the chromosomes are altered. The mutation process helps to overcome trapping at local maxima. The offsprings produced by the genetic manipulation process form the next population to be evaluated. The process is repeated until the specified numbers of generations are completed. Finally, the chromosome with the best fitness score has the weights of the regulators of the target gene. The algorithm is repeated for each of the target gene.

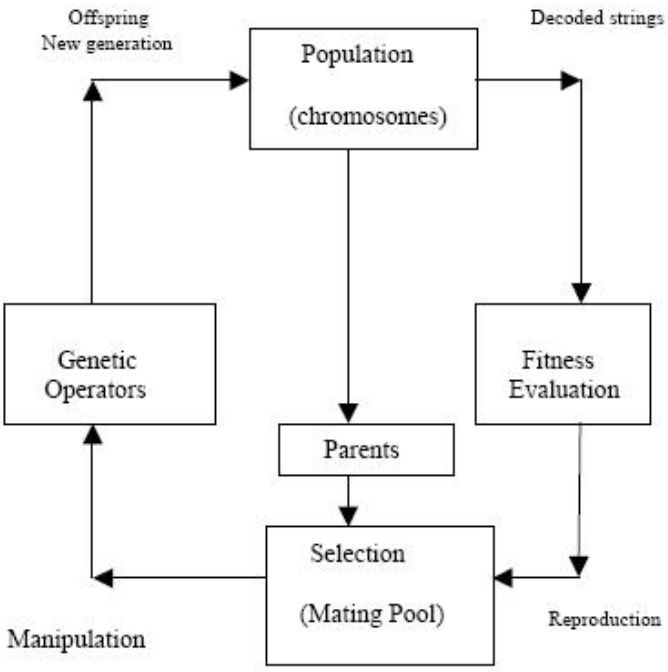


Figure 2: Cycle of stages in Genetic Algorithm.

Results and Discussion

In the present paper the gene expression profiles of circulating plasma RNA from colorectal cancer patients were analyzed. The entire dataset includes the expression levels of 15552 genes obtained by performing genomic profiling of plasma RNA from colorectal cancer patients (12) and from healthy donors (8) through cDNA microarray hybridization. Each sample was competitively hybridized against a common reference formed by a pool of blood samples from 26 healthy donors [22][23].

Vineetha *et.al*.[15] selected 100 most significant genes from the above data set using t-test whose p - values are less than 0.005. They clustered these genes using hybrid clustering technique[24]. Using fuzzy logic algorithm[13] and with the cluster centroid as input, they have identified the activator/repressor regulatory relationship between 27 genes. In this paper, all these 27 genes were again considered to improve the efficiency of inferring the regulatory network using modified genetic algorithm. The program was implemented in Matlab. Table 1 shows the GA parameters used in this experiment. Due to the large search space, the basic GA requires long time to converge. By incorporating the correlation techniques, modified GA runs fast and gives good results compared to the basic GA. This algorithm helps to integrate the biological expert knowledge with heuristic search.

Table 1: Algorithm Parameters.

	No. of Variables	Population	Generations	Crossover	Mutation	Time (secs)
Basic GA	26	200	1500	0.8	0.2	200
Modified GA	26	100	150	0.8	0.2	50

The weight matrix generated using modified GA was used to model Gene regulatory network. Each node in the GRN represents a gene and the presence of an edge between two nodes indicates an interaction between connected genes. The edges labeled positive weight denote activation and the edges labeled negative weight denote repression.

The set of all regulatory relations predicted by modified GA are shown in table 2. The similar results were acquired using other models such as feed forward neural fuzzy network and fuzzy logic algorithm. The modified GA was successful in determining almost all regulatory relations determined by other models. The advantage of the new approach was that it can approximate the relationships among genes without heavy computation. The regulatory relations determined by feed forward neural fuzzy network are shown in table 3.

From the previous gene expression analysis of Colorectal Cancer, three of the selected markers –PSMA3, EPAS1, UBE2D3- were found to be significantly higher in cancer patients compared to healthy donors. Our model can be used to identify those genes which are affected by the up regulation of the above cancer markers. The identification of circulating markers for CRC would optimize early stage diagnosis

and the monitoring for disease recurrence. Figure 3 shows the regulatory pathways of EPAS1, PCBP2, PSMA3 in plasma RNA dataset as observed using modified GA.

Table 2: Regulatory Relations Obtained Using Modified Genetic Algorithm.

Activator	***Target***
EPAS1	KIAA0632
	KIAA0758
	PCBP2
	RPL10A
	PSMA3
	HBE1
	LDHA
LDHA	KIAA0632
	KIAA0758
	MVP
	IL7R
	PCBP2
	ANXA1
	RPL10A
	PSMA3
	HBE1
OXA1L	PCBP2
	ANXA1
	PSMA3
ANXA1	KIAA0632
	IL7R
	DKFZP564B167
	RPL10A
	OXA1L
	PSMA3
DGKZ	KIAA0632
TRIP10	TNKS

Activator	***Target***
PCBP2	MVP
	IL7R
	OXA1L
RELN	DOC1
	TNKS
KIAA0758	PCBP2
	DKFZP564B167
	PSMA3
	LDHA
ACP1	KIAA0758
	MVP
	PCBP2
	DKFZP564B167
	PSMA3
FLJ20701	DKFZP434N061
PSMA3	MVP
	ACP1
	IL7R
	PCBP2
	ANXA1
	OXA1L
	LDHA
DKFZP564B167	KIAA0758
	ACP1
	DGKZ
IL7R	MVP
	PCBP2

Activator	***Target***
PPP1CC	DKFZP564B167
	DKFZP434N061
	DGKZ
KIAA0632	ANXA1
	DGKZ
	HBE1
MVP	IL7R
	PCBP2
	ACP1
	DKFZP564B167
RPL10A	KIAA0632
	ANXA1
	LDHA
	PSMA3
	HBE1
Repressor	Target
RELN	PCBP2
	HBE1
TRIP10	MVP
	HBE1
	PCBP2
TIAF1	DKFZP564B167
SP38	PCBP2
FLJ10849	MVP
	PCBP2
	PSMA3

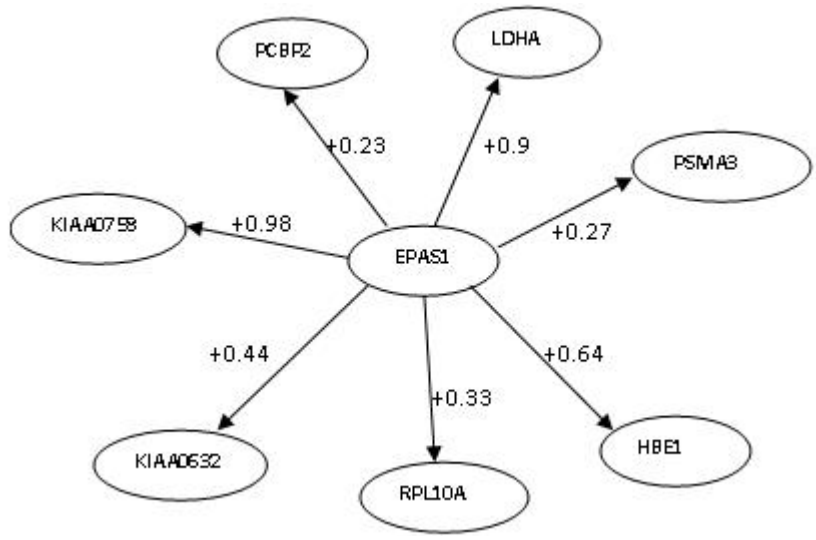
Table 3: Regulatory Relations Obtained Using DNFN.

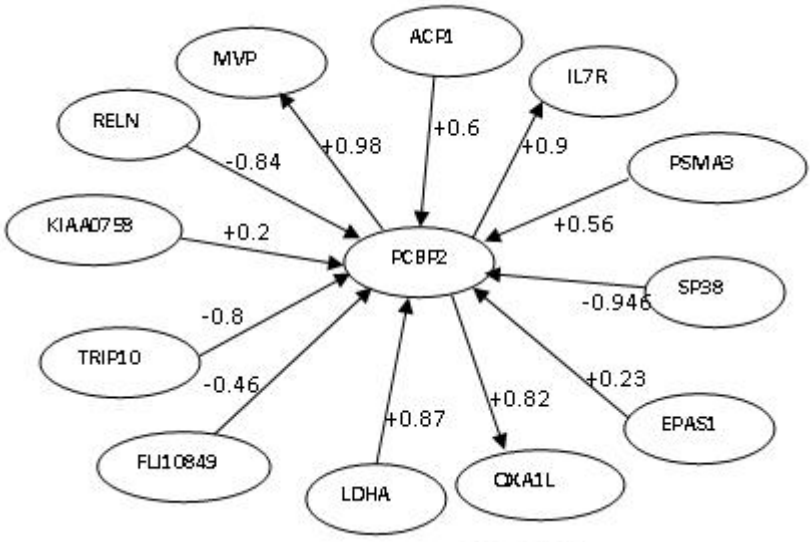
Activator	***Target***
EPAS1	KIAA0632
	KIAA0758
	MVP
	PCBP2
	ACP1
	ANAX1

Activator	***Target***
KIAA0758	MVP
	PCBP2
	ACP1
	DKFZP564B167
	DGKZ
	OXA1L

Activator	***Target***
DKFZP564B167	PCBP2
	DGKZ
TRIP10	TNKS
	RPL10A
	RELN
PPP1CC	DKFZP434N061

	RPL10A	ACP1	MVP		DGKZ
	OXAIL		PCBP2	KIAA0632	ANAX1
	LDHA		HBE1		DGKZ
	PSMA3	ANXA1	DKFZP564B167	DGKZ	DKFZP564B167
	HBE1		DKFZP434N061	FLJ20701	HBE1
LDHA	MVP		DGKZ	MVP	HBE1
	PCBP2		OXA1L	Repressor	Target
	RPL10A		PSMA3	FLJ10849	ACP1
	HBE1	PSMA3	OXA1L	RELN	DKFZP564B167
			LDHA		HBE1
			HBE1	TRIP10	HBE1





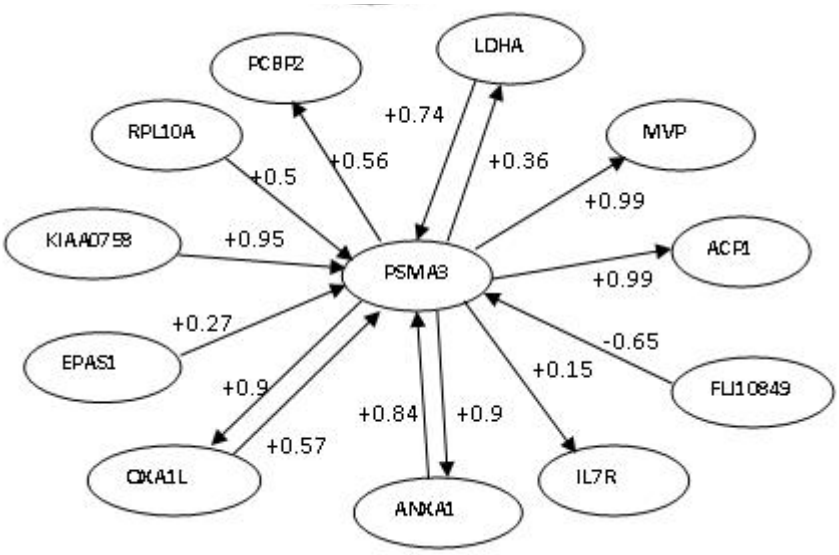


Figure 3: Regulatory pathways of EPAS1, PCBP2, PSMA3 observed by modified GA.

Conclusion

In the present paper, Genetic algorithm was applied for optimizing the weight matrix for gene regulatory network. By incorporating statistical technique viz. correlation analysis, search space has been immensely reduced. As a result, computational time for modified GA was minimal compared to that of basic GA. By using this approach a high resemblance was observed and actual expression pattern has been acquired. The relations obtained were compared with the regulatory network simulated by other models such as fuzzy GRN algorithm and feed forward neural fuzzy network and the results obtained are almost comparable.

The regulatory relationships between 27 differentially expressed genes in the plasma RNA of CRC patient were modeled. These findings may perhaps provide new insights into cancer diagnostics, prognostics and therapy.

References

[1] Xu, and Luonan Chen, 2006, "Inferring gene Regulatory networks from multiple microarray datasets", Bioinformatics. 22, pp. 2413-2420

[2] Akutsu T., Miyano S., and Kuhara S., 1999, "Identification of genetic networks from a small number of gene expression patterns under the boolean network model". Pacific Symposium on Biocomputing , pp.17-28

[3] S. Liang, S. Fuhrman, and R. Somogyi, 1998, "Reveal, a general reverse engineering algorithm for inference of gene network architectures", Pacific Symposium on Biocomputing, pp.18-29

[4] Kato M., Tsunoda T., and Takagi T. 2000, "Inferring genetic networks from DNA microarray data by multiple regression analysis", Genome Informatics. 11, pp.118-128
[5] N. Friedman, M.Linial, I.Nachman, and D.Pe'er, 2000, "Using Bayesian network to analyze gene expression data", Journal for Computational Biology, 7, pp. 601-620
[6] Marnellos, G., and Mjolsness, E. D., 2001, "Modelling Neural Development". MIT Press, Cambridge, MA, pp. 27-28
[7] Vohradsky, J., 2001, "Neural network model of gene expression", The FASEB Journal.15, pp. 846-854
[8] Jiayin, W., Yufei, H., Maribel, S., Yufeng Wang, and Jianqiu (Michelle) Zhang, 2006, "Reverse Engineering Yeast Gene Regulatory Networks Using Graphical Models", ICASSP IEEE, pp.1088-1091
[9] Bor-Sen C., Shih K. Y., Chung-Yu L. , and Yung-Jen Chuang, 2008, "A Systems biology approach to construct the gene regulatory network of systematic inflammation via microarray and database mining", BMC Medical Genomics.1
[10] Mauricio F., and Fernando G, 1999, "Design of Fuzzy Systems Using Neurofuzzy networks" , IEEE Transactions on Neural Networks.10
[11] Ramesh, R., Madhu, C., and Trevor, I. D., 2006, "Fuzzy Model for Gene Regulatory Network", IEEE Congress on Evolutionary Computation. pp.1450-1455
[12] Ressom, H., Wang, D., Varghese, R. S., and Reynolds R., 2006, "Fuzzy Logic-Based Gene Regulatory Network", Proceedings of the IEEE International Conference on Fuzzy Systems, pp.1210-1215
[13] Woolf P. I, and Wang Y., 2000, "A fuzzy logic approach to analyzing gene expression data", Physiological Genomics.3, pp. 39-15
[14] De Jong H., 2002, "Modelling and simulation of genetic regulatory systems: a literature review", Journal of Computational Biology.9 , pp. 67-102
[15] Vineetha S., Chandra Shekara Bhat C., and Sumam Mary Idicula, 2010, "Gene Regulatory Network from Microarray Data using Dynamic Neural Fuzzy Approach", Proceedings of the International Symposium on Biocomputing
[16] Shin Ando, and Hitoshi Iba, 2003, "Inference of Gene Regulatory Model by Genetic Algorithms", IEEE ,pp.712-719.
[17] Shin Ando, and Hitoshi Iba, 2003, "Estimation of gene Regulatory Network by Genetic Algorithm and Pairwise Correlation Anlysis", IEEE, pp.207-214
[18] American Cancer Society, 2008, "Cancer Facts and Figures", Atlanta
[19] Manuel C., Vanesa G., Jose , M. G., Isabel A., Luis L., Ramon D., Luis A., Lo'pez Ferna'ndez, Angel Z., Fe'lix, B., and Manuel Serano, 2007, "Genomic profiling of circulating plasma RNA for the analysis of cancer", Clinical Chemistry 53
[20] D.C Weaver,C.T Workman, and G.D stormo, 1999, "Modelling regulatory networks with weight matrices", Pacific Symposium on Biocomputing ,4, pp.112-123

[21] M.E.Mamakou, G.Ch.Sirakoulis, I.Andreadis , and I.Karafyllidis,2005, "Adaptive Reverse Engineering of Gene Regulatory Networks Using Genetic Algorithms", EUROCON

[22] Collado M., Garcia V., Garcia J.M., and Alonso I., 2007, "Genomic profiling of circulating plasma RNA for the analysis of cancer", Clinical Chemistry ,53, pp.1860-1863

[23] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov

[24] Vineetha S., Chandra Shekara Bhat C., and Sumam Mary Idicula, 2009, "Analysis of Circulating Plasma RNA from Colorectal Cancer Patients using Unsupervised Hybrid Clustering Algorithm", International Conference on Advances in Optoelectronics, Information and Communication Technologies(ICOICT). pp.591-593

[25] Chia-Feng J., and Chin-Teng L, 1999, "A Recurrent Self-Organizing Neural Fuzzy Inference Network", IEEE Trans. Neural Networks,10, pp.828-845

[26] H. R. Berenji, and P. Khedkar, 1992, "Learning and tuning fuzzy logic controllers through reinforcements", IEEE Trans. Neural Networks, 3, 724-740

[27] J. S. Jang, 1993, "ANFIS: Adaptive-network-based fuzzy inference system", IEEE Trans. Syst., Man, Cybern., 23, pp. 665-685

[28] Lev A Soinov, Maria A Krestyaninova , and Alvis Brazma ,2003, "Towards reconstruction of gene networks from expression data by supervised learning", Genome biology,3

[29] Zainal, A., Hasibuan, R. , Fadilah, R., Muhammad, F. P., and Rahmat, B., 2009, "Adaptive Nested Neural Network (ANNN) based on Human Gene Regulatory Network for gene knowledge Discovery Engine",IJCSNS.9`