Reconstruction of Gene Regulatory Network using Modified Ant-based Algorithm

Rakhi Wajgi, Manali Kshirsagar, Dipak Wajgi, Gauri Chaudhary, Gauri Dhopavkar

Abstract*: Healthcare is a major area of research since few years. Ample amount of biological data getting accumulated daily due to advancement in technologies. Microarray is such technology which captures expressions of thousands of genes at a time. Interactions occur among genes are represented in terms of special network known as Gene Regulatory Network (GRN). It is constructed from Differentially Expressing Genes(DEFs). GRN is a graphical representation containing genes as nodes and regulatory interactions among them as edges. It helps in tracking pathways where usual gene interaction changes leading to malfunctioning of cells and results in illness. Also, now a day's people are diagnosed with new diseases like dengue, swine flu, Nipah, Corona virus infection for which exact molecular pathways are yet to be invented through GRN. Therefore, in this paper, a nature inspired algorithm is used for reconstruction of GRN using differentially expressing genes.

***Keywords* : Microarray, Genes, Cellular Biology, Gene Regulatory Network, Differentially Expressing Genes**

I. INTRODUCTION

Genes contain blue print of living organisms. All cell activities are controlled by synthesis of proteins whose disproportionate share causes malfunctioning in cellular activity. Some gene products known as proteins are required by cells under all growth conditions. Those are called housekeeping genes. These include genes that encode proteins such as DNA polymerase, RNA polymerase, and DNA gyrase. Some gene products are required under specific growth conditions . These include enzymes that synthesize amino acids, break down specific sugars, or respond to a specific environmental condition such as DNA damage [1]. To analyze the insight of biological activities, analysis of gene expressions is necessary . Advanced technology like microarray plays an important role in gene expression analysis as it captures expressions of thousands of genes under different conditions simultaneously. Those genes which behave differently under stress conditions are called as Differentially Expressing Genes (DEGs).

Identifying gene interactions is a major challenge in post genomic era. It helps in knowing how cells maintain their form. Though vast amount of biological data getting accumulated day by day, a technique is needed which will successfully model uncertainty lies in gene expressions in terms of GRN.

A. Definition and Concepts

Definition 1: The Gene Regulatory Network is a graph G(E, N), where N represents set of genes and E represents set of regulatory interactions through which genes communicate with each other.

Definition 2: A positive regulation between gene g1 and gene g2 is indicated by a directed edge arising from source gene g1 to the target gene g2 and is denoted as g1 → g2. Gene g1 positively regulates gene g2; iff binding of gene g1 at specific promoter causes gene g2 to express. In this case gene g1 is called activator gene and gene g2 is called target gene.

Definition 3: A negative regulation between gene g1 and gene g2 is indicated by an undirected edge arising from source gene g1, and closed at target gene g2, g1 ⊣ g2 and is denoted as gene g1 negatively regulates gene g2; iff inactivation of gene g1 at operon site causes gene g2 to express. In this case gene g1 is called inhibitor gene and gene g2 is called target gene. For showcasing GRN, graphical representation is preferred as it is simple and perfect layout to show interaction between genes. Interaction between genes can be shown using any preferred way not necessarily as mentioned in definitions2 and 3. Figure 1 shows sample GRN of budding yeast. Green arrows and red blunt-end ones are activating and inhibiting interactions, respectively . For self-pointed arrows, orange blunt-end indicates self-degradation .

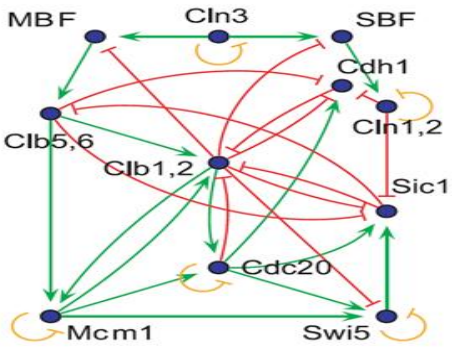

Figure1. Sample GRN of budding yeast [2]

In past few years there are many methods proposed in [3][4][5][6][7][8] for inference of GRN but still this research area has a wide scope because of inability to reach to maximum detection of true positive interactions between genes for complex disorders.

Revised Manuscript Received on March 11, 2020.

***** Correspondence Author

Rakhi Wajgi*, Computer Technology Department, YCCE, Nagpur, Maharashtra, India. Email:wajgi.rakhi@gmail.com

Manali Kshirsagar Principal RGCER, Nagpur, Maharashtra , India, Email:manali_kshirsagar@yahoo.com

Dipak Wajgi, Computer Engineering Department, SVPCET, Nagpur, Maharashtra, India, Email:wajgi@rediffmail.com

Gauri Chaudhary: Computer Technology Department, YCCE, Nagpur, Maharashtra, India. Email: chaudhary_gauri@yahoo.com

Gauri Dhopavkar: Computer Technology Department, YCCE, Nagpur, Maharashtra, India. Email:gauri.ycce@gmail.com

Based on the chronological order, existing models are classified into two major categories i.e. conventional and non-conventional. A conventional model includes Boolean Network, Bayesian Network, Linear Differential Model and non-conventional model includes Neural Network model and Model based on Evolutionary algorithms. Paper[9] gives detailed review of existing mathematical models used for reconstruction of GRN along with database and experimental setup used. Disadvantages of some of the important models are given in Figure 2.

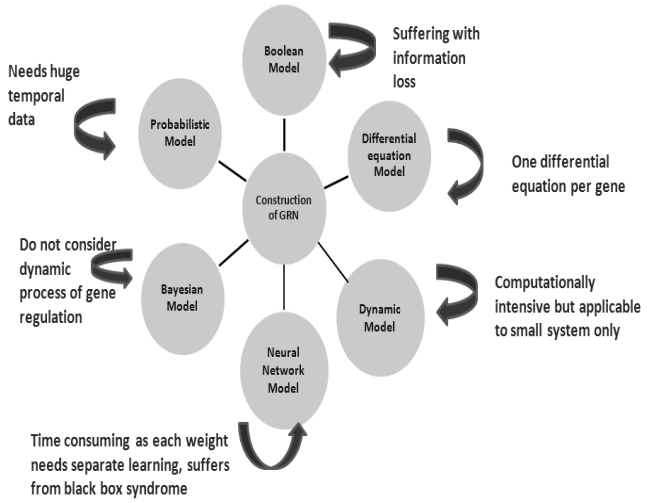

Figure 2. Disadvantage of existing models used for construction of GRN

II. METHODOLGY

Research in bioinformatics demands use of advanced tools for processing huge amount of ambiguous and uncertain biological data . Discovering patterns hidden in the gene expression data across number of samples which are correlated with specific condition has a tremendous opportunity and challenges for functional genomics and proteomics [10][11][12]. Unfortunately, employing any kind of pattern recognition algorithm to such data is hindered by the curse of dimensionality (limited number of samples and very high feature dimensionality) [13]. Algorithms belonging to Swarm Intelligence Category have capability to handle enormous data and generate solution from it in simpler way. Errors generated in microarray are more tolerable in SI algorithms than in deterministic algorithms. Errors are treated as contributing factor for population diversity, a desirable property for convergence of SI algorithms [14]. Therefore Ant Colony Optimization based algorithm is proposed which will generate GRN from any number of genes in less time by considering relationship between genes.

In 2005, Karaboga [15] gave an interesting idea of artificial ants based algorithm known as Ant colony optimization (ACO) algorithm. Ants are blind, but yet know how to find the shortest distance between the food source and there native place. Ants use pheromones laid by the other ants as footmarks to follow and hence ant reaches the shortest path by using knowledge gained by the other ants and this behavior is imitated in the form of an algorithm that can be used for optimization problems, including gene interaction network optimization [16]. In [16], ACO is used for inference of GRN but author is able to find number of interactions equal to number of genes. It is major drawback because one target gene has many controlling parent which regulates its expressions [17]. Inspired by the foraging activities of ants, ant colony optimization [18] is a class of metaheuristics that provide a generic framework of communication between simple agents (artificial ants), whose task is to construct candidate solutions to the optimization problem under consideration. One type of heuristic that has not been used previously is Ant-Based algorithm. The difference between the Ant Colony Optimization and the Ant-Based algorithm is that in both cases artificial ants maneuver based on the local information and deposited pheromones as they travel but in Ant-Based algorithm cumulative pheromone levels are used to build candidate solution. In Ant Colony Optimization, each ant builds the individual solution and leaves pheromone on the edges which act as guide for remaining ants but in Ant-Based algorithm each ant builds a part of solution and together efforts of all ant gives rise to final solution. We have combine features of both the algorithms and proposed a hybrid approach known as Sequential

Algorithm 1 Algorithm for reconstruction of GRN using SDCAA

Initialize pheromone matrix $\tau_{ij}(t2) = [x]_{NXN}$ based on correlation coefficient between genes for first two sample points.
Initialize Tabu list for each gene g_i as $T_i = \{g_i\}$
Initialize Interaction type as $I_{ij} = \{0\}_{NXN}$
Initialize $D_i = \{0\}$ which contains degree of each gene
$\alpha := 1$ the parameter controlling influence of pheromone on the edge
$\beta := 2$ the parameter controlling desirability of edge between gene i and j
$\rho := 0.5$ is pheromone evaporation rate
$E := [e_{ij}]_{NXN}$ Contains gene expressions of DEGs

Procedure

1: Initialize one ant at each gene
2: Initialize pheromone between pair of genes using equation 1
3: **while** stopping criteria not met
4: **for** t = 2 to M **do**
5: **for** each ant k **do**
6: Move ant k from gene i to j with probability

$$P_{ij}^{k}(1,t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{K \epsilon allowed_k}[\tau_{ik}(t)]^{\alpha}.[\eta_{ik}]^{\beta}} & if\ j \epsilon allowed_k \\ 0 & otherwise \end{cases}$$

7: Update pheromone for edges which are selected by ant using expression

$$\tau_{ij}(1,t) \coloneqq (1-\rho)\tau_{ij} + 1/\eta_{ij}$$

8: Update pheromone for the edges which are not selected using

$$\tau_{ij}(1,t) \coloneqq (1-\rho)\tau_{ij}$$

9: Update Tabu list T_i ,Interaction type I_{ij} and degree vector D_i of each gene
10: **end for**
11: **end while**
12: **if** stopping criteria met **then** go to step 13 **else** empty Tabu list and go to step 3
13: Based on threshold value of pheromone construct adjacency matrix $A[i][j]_{NXN}$ between genes
14: **if** $\tau_{ij} < Th$ **then** $A[i][j] = 0$
15: **else** $A[i][j] = 1$
16: **end if**

Where M is maximum and m is minimum value of expression for i^{th} gene. Value of pheromone decides the regulatory interaction between genes. Scaling factor 3 is used in order to have large enough differences in pheromone values so that it is easy to select suitable edge.

The algorithm adds new edge between existing gene g1 in GRN and the new gene g2 which has highest probability of getting selected. Due to this, total number of edges at the end is more than number of genes which is advantageous in biological point of view. Amount of pheromone evaporates if edge connecting already added ge ne is not selected in further iteration. By selecting edges having pheromone value above threshold restricted the degree of each node in GRN. The algorithm stops when GRN with maximum deposition of pheromone is generated. The algorithm is compared with existing approaches on the basis of true positive edges matched with benchmark networks. Specificity and sensitivity of reconstructed GRN is also calculated to check the performance of proposed algorithm. GRN is constructed from adjacency matrix generated from Algorithm1 which is given as input to Cytoscape [19]. It is mainly used for graphical display of any kind of biological networks. Solid edges are used to represent positive regulation and dotted edges to represent negative regulation along with label 1 and -1 respectively. In order to compare the result of SDCAA with other existing approaches, GRN is constructed using following datasets.

A.Urilon dataset: It contains 9 significant genes responding to DNA breakdown. This is the most preferred dataset used for validity of new method. This dataset contains missing values which are first imputed and the GRN is constructed from it. Four different experiments were conducted with different UV light intensities. Using these experiments, expressions of eight major genes, such as uvrD, uvrA, lexA, recA, umuDC, ruvA, polB and uvrY, have been documented as shown in the Figure 3. The displayed relationships express known regulatory interactions between genes. Normal arrow heads denote activation, while diamond-shaped arrow heads denote repression or inhibition.

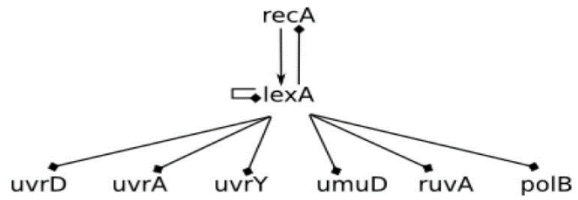

Figure3. Structure of the SOS DNA repair transcriptional network of E. coli[5].

SOS DNA repair transcriptional network of E.coli using SDCAA algorithm is shown in Figure 4.

III. RESULTS

We have compared GRN modeled using SDCAA with different existing approaches on the basis of known interaction which is shown in Table 1. In this dataset, LexA is a major repressor gene which represses expressions of all other genes.

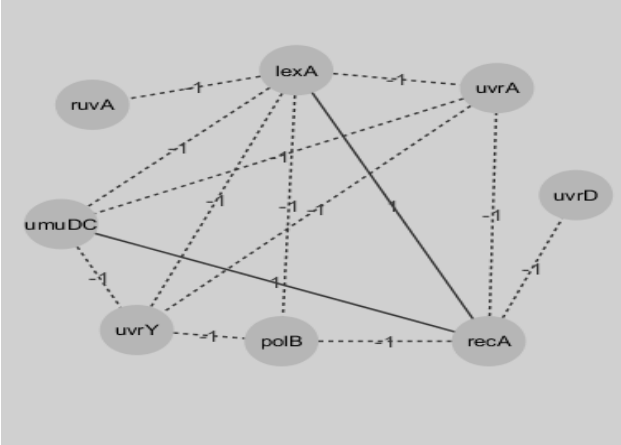

Figure 4 SOS DNA repair GRN of E. coli using SDCAA

In the table Y indicates that known interaction mentioned in first column is correctly predicted using respective method. In methods [20][21][22][23] two genes, uvrY and ruvA were not considered as missing values are there in their expressions for initial time samples. Positive predictive value is calculates using formula as:

$$PPV = \frac{TP\ edges}{Total\ number\ of\ edges}$$

False Positive edges are the edges which are incorrectly identified as significant.Method mentioned in [24] reports a less conservative prediction which included all nine true relations but more FP = 7 leading to a lower precision value (PPV = 0.56). Neural network technique is used in [5] which is suffering with black box syndrome. Apart from true edges, list of spurious edges of SOS DNA repair GRN is listed in Table 2. Sensitivity of GRN in Figure 4 using proposed algorithm is 66% and specificity is 33%.

Table 1 Comparative analysis of SDCAA with other methods for SOS GRN

Known interactions	**[25]**	**[26]**	**[20]**	**[24]**	**[21]**	**[22]**	**[23]**	**[5]**	**[16]**	**SDCAA**
lexA->lexA	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*Y*	*Y*	*N*	*N*
lexA->recA	*Y*	*Y*	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*
recA->lexA	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*Y*	*N*	*N*	*Y*
lexA->uvrA	*Y*	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*Y*	*Y*	*Y*
lexA->uvrD	*N*	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*N*
lexA->uvrY	*N*	*N*	-	*N*	-	-	-	*Y*	*N*	*Y*
lexA->umuD	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*N*
lexA->ruvA	*N*	*N*	-	*N*	-	-	-	*Y*	*N*	*Y*
lexA->polB	*N*	*N*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*Y*
Spurious edges (FP)	*5*	*10*	*6*	*7*	*15*	*16*	*11*	*5*	*6*	*6*
Precision (PPV)	*0.28*	*0.33*	*0.50*	*0.56*	*0.29*	*0.30*	*0.39*	*0.62*	*0.25*	*0.43*

Table 2 List of Spurious edges of SOS DNA repair DNA

Sr.No.	**Gene 1**	**Gene 2**	**Interaction**
1.	*umuDC*	*uvrA*	*negative*
2.	*umuDC*	*recA*	*negative*
3.	*uvrA*	*recA*	*negative*
4.	*uvrA*	*uvrY*	*negative*
5.	*uvrY*	*polB*	*negative*
6.	*polB*	*recA*	*negative*

A.Yeast cell cycle ($\alpha - factor$): It is also called Spellman dataset [21] containing gene expressions of yeast while undergoing cell cycle regulation. Figure 5 shows the standard GRN of yeast cell cycle constructed using GeneNetweaver. GRN is constructed using 10 DEGs.

Figure 6 shows GRN constructed using SDCAA algorithm. Positive interaction is shown using continuous line and negative interaction is shown using dotted line. Total number of edges in the network is 15 out of which 6 are spurious edges which are shown in Table 3.

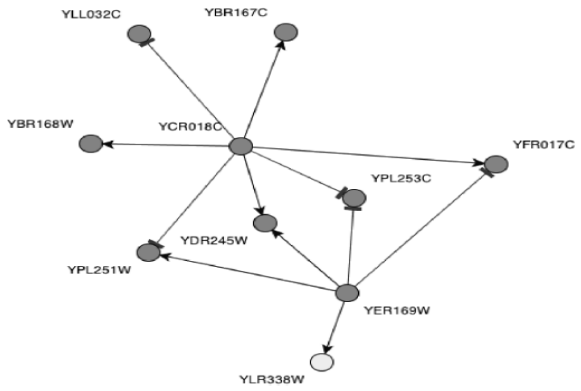

Figure 5. Standard GRN of Yeast from GeneNetweaver

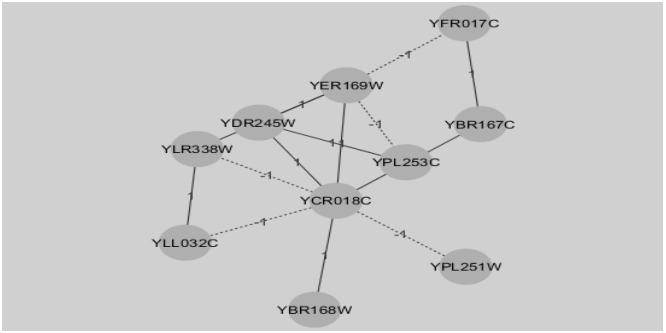

Figure 6 GRN of cell cycle regulated genes in yeast using SDCAA

Table 3 Gene interaction for cell cycle regulated genes in yeast

Gene1	Gene2	Predicted Interaction type	Standard interaction P/N
YCR018C	*YBR167C*	*1*	*P*
YCR018C	*YLL032C*	*-1*	*P*
YCR018C	*YBR168W*	*1*	*P*
YCR018C	*YPL251W*	*-1*	*N*
YCR018C	*YDR245W*	*1*	*P*
YCR018C	*YER169W*	*1*	*N*
YER169W	*YLR338W*	*1*	*P*
YER169W	*YFR017C*	*-1*	*P*
YER169W	*YPL253C*	*-1*	*N*
YER169W	*YDR245W*	*1*	*P*
YFR017C	*YBR167C*	*1*	*N*
YCR018C	*YLR338W*	*-1*	*N*
YPL253C	*YDR245W*	*1*	*N*
YLL032C	*YLR338W*	*1*	*N*
YFR017C	*YER169W*	*-1*	*N*

IV. DISCUSSION AND CONCLUSION

Sensitivity of SOS and yeast network is 66 % and 60% respectively and specificity is 33% and 50% respectively. SDCAA can build GRN of any size in considerable amount of time. It is flexible and less time consuming which makes it better choice for reconstruction of GRN. Time complexity of SDCAA is O(m*n) where m is number of sample points and n is number of genes.

REFERENCES

1. J. Biju, S. Anuparna and K. Govindswami, “Microarrays ‘Chipping’ in Genomics”, Indian Journal of Biotechnology, 1(3):245-254,2002
2. Y. Wu, X. Zhang, J. Yu and Q. Ouyang, “Identification of Topological Charactristics Responsible for the Biological Robusteness of Regulatory Networks”, PLOS Computational Biology,5(7), 2009
3. Noor, E. Serpedin, M. Nounou, and H. Nounou “Inferring Gene Regulatory Networks via Nonlinear State-Space Models and Exploiting Sparsity”, IEEE/ACM Transactions On Computational Biology And Bioinformatics, 9(4):1203-1211, 2012
4. R. Ram and M. Chetty, “A Markov-Blanket-Based Model for Gene Regulatory Network Inference”, IEEE/ACM Transactions On Computational Biology And Bioinformatics , 8(2): 353-367, 2011
5. K. Kentzoglanakis and M. Poole “A Swarm Intelligence Framework for Reconstructing Gene Networks: Searching for Biologically Plausible Architecture” IEEE/ACM transaction on computational Biology and Bioinformatics , 9(2):358-370, 2012
6. M. Tan, M. Alshalalfa, R. Alhajj, and Faruk Polat, “Influence of Prior Knowledge in Constraint-Based Learning of Gene Regulatory Networks”, IEEE/ACM Transactions On Computational Biology And Bioinformatics, 8(1):130-142,2011
7. R. Xu, G. Venayagamoorthy, and D. Wunsch, “Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization”, Neural Networks, 20 (2007):917–927,2007
8. H. Iba, and A. Mimura, “Inference of gene regulatory network by means of interactive evolutionary computing”, Journal of Information Sciences,145 (2002):225-236,2002
9. Chanda Panse, Manali Kshirsagar, “Survey On Modelling Methods Applicable to Gene Regulatory Network”, International Journal on Bioinformatics & Biosciences, Vol.3, No.3, September 2013
10. D. Jiang, C. Tang, and A. Zhangi, “Cluster Analysis for gene expression data: A survey”, IEEE Transactions on Knowledge and Data Engineering, 16(11):1370-1386,2004
11. P. Larranaga, B. Calvo, R. Sanatana, C. Bielza, J. Galdiano, I. Inza, J. Lozana, R. Armananzas, G. Santafe, A. Perez and V. Robles, “Machine learning in bioinformatics”, Briefings in Bioinformatics,7(1):86-112,2006
A. Stantnikov, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis”, Bioinformatics,21(5):631-643,2005
12. D. Tasoulis, V. Plagianakos and M. Vrahatis, “ Computational Intelligence Algorithms and DNA Microarrays”, Studies in Computational Intelligence(SCI), 94, pages 1-31,2008
13. M. Clerc and J. Kennedy, “The Particle Swarm-Explosion, Stability and Convergence in a Multidimensional Complex Space,” IEEE Trans. Evolutionary Computation, 6(1):58-73,2002
14. S. Kim, J. Kim, and K.Cho, “Inferring gene regulatory networks from temporal expression profiles under time-delay and noise”, Journal of Computational Biology and Chemistry, 31(2007):239-245, 2007
15. K. Raza and M. Kohli, "Ant Colony Optimization for Inferring Key Gene Interactions", Proceeding. of Ninth INDIACom-2015, 2nd International Conference on Computing for Sustainable Global Development, pages 1242-1246, 2015

16. J. Donkers and K. Tuyls, "Belief Networks for Bioinformatics", Studies in Computational Intelligence(SCI), 94, pages 75-112,2008
17. M. Dorigo, V. Maniezzo, and A. Colorni, "The Ant System: Optimization by a Colony of Cooperating Agents," IEEE Trans. Systems, Man and Cybernetics, Part B, 26(1):29-41,1996.
18. Cytoscape manual available at website http://www.cytoscape.org/manual/Cytoscape2_6Manual.pdf
19. D. Cho, K. Cho, and B. Zhang, "Identification of Biochemical Networks by s-Tree Based Genetic Programming," Bioinformatics, 22(13):1631-1640, 2006.
20. S. Kimura, K. Sonoda, S. Yamane, H. Maeda, K. Matsumura, and M. Hatakeyama, "Function Approximation Approach to the Inference of Reduced NGnet Models of Genetic Networks," BMC Bioinformatics, 9(23), 2008
21. S.Kimura, S. Nakayama, and M. Hatakeyama, "Genetic Network Inference as a Series of Discrimination Tasks," Bioinformatics, 25(7):918-925, 2009
22. M. Kabir, N. Noman, and H. Iba, "Reverse Engineering Gene Regulatory Network from Microarray Data Using Linear Time-Variant Model," BMC Bioinformatics, 11(1), 2010
23. R. Xu, D. Wunsch, and R. Frank, "Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization", IEEE/ACM Transactions On Computational Biology And Bioinformatics, 4(4):681-692,2007
24. B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and D. Buc, "Gene Network Inference Using Dynamic Bayesian Networks," Bioinformatics, 19(2):138-148, 2003
25. N. Noman and H. Iba, "Reverse Engineering Genetic Networks Using Evolutionary Computation," Genome Informatics, 16(2):205-214,2005
26. P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, P. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Molecular Biology of the Cell, 9(12):3273-3297, 1998
27. Grubman, A., Chew, G., Ouyang, J.F. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neuroscience (2019) doi:10.1038/s41593-019-0539-4

AUTHORS PROFILE

Rakhi Wajgi, completed her Master from BITS Pilani Rajasthan in 2008. She completed her doctorate in computer Science and Engineering discipline in 2019. She has more than 20 research publications in her credit. She has more than 10 years of teaching experience in engineering colleges. Currently she is working as an Assistant Professor at YCCE, Nagpur.

Dr. Manali Kshirsagar, received her B.E. in Computer Technology, M.E. Computer Science & Engineering and Ph.D. in the faculty of Computer Science and Information Technology. Her specialization at the PG and Ph.D. level has been in the areas of Data Warehousing , Data Mining and Bioinformatics. She has about 20 technical and research publications to her credit which have appeared in International Journals, International and National conferences. She is guiding 08 Ph.D. students at present.

Dipak W. Wajgi, received his M.Tech. in Computer Science and Engineering from Ramdeobaba College of Engineering and Management , Nagpur, India. He has around seventeen years of teaching experience in engineering colleges and currently pursuing his PhD in CSE. He has published more than 15 papers in various reputed journals and conferences.

Gauri Chaodhary, received her M.Tech. in Computer Science and Engineering Yeshwantrao Chavan College of Engineering Nagpur, India. She has 8 publications in her credits. She worked as manager in Global Logic Nagpur. Currently pursuing her PhD in Data Mining and also working as ERP in-charge MGI Nagpur.

Gauri Dhopavkar, received her MTech from G.H. Raisoni College of Engineering, Nagpur. She received her PhD in the domain of Natural Language Processing. She has 20 publicationsin her credit. Currently she is working as Head of the department Computer Technology YCCE. Her area of research includes Natural Language Processing and software Engineering.