INFERRING GRAPH-BASED VISUALIZATION OF GENE REGULATORY NETWORKS IN 3D USING GENETIC ALGORITHM

Sumon Ahmed, Mohammad Shahedul Islam, Nasimul Noman and Kazi Sakib

Abstract—**In recent years, the estimation of genetic network has been accelerated with the advent of high throughput DNA microarray technology, which has enabled the researchers to measure gene expression levels of thousands of genes simultaneously. With this rapid increase of the gene network size, a very effective means to identify the collaborating genes in the network is visualization, which will facilitate the network behavioral study. However, in the literature, no definitive method has been found for drawing a large effective graph from this enormous data. This paper used a layered approach to visualize the Gene Regulatory Network (GRN), which is an optimization problem requiring maximizing the fitness function. A new fitness function has been proposed based on edge-edge crossing, node-edge crossing and node distance. A Genetic Algorithm (GA) based meta-heuristic search technique is incorporated in the proposed methodology for efficiently attaining the visualization of target networks. The suitability of the proposed method is investigated through various experiments using two real data sets of 82 genes and 552 genes. The experimental results illustrate the suitability of the proposed method in improving the understandability of biological systems, by reducing node-edge and edge-edge crossings in the generated outputs (graphs).**

Index Terms—Genetic algorithm, gene regulatory network, visualization

I. INTRODUCTION

The revolutionary advancement in molecular biology, particularly in high-throughput genomics and proteomics continue to produce massive biological data which have opened the door of various researches. In a microarray experiment, it is possible to survey gene expression levels of thousands to tens of thousands of genes under various environmental conditions [1]. As increasing numbers of gene expression data are becoming available, the researchers are now more interested in studying the relationships among genes at the large scale. Visualization is one of the effective techniques may seduce the genomics community by its potentials to resolve the hidden mysteries of genetic code. Obtaining a layout of gene regulatory network gives a whole new perspective on genetic interactions and leads to new experimental designs in the field of *Systems biology*.

Recently, the method of inferring gene regulatory network has been developed very rapidly which leads the target network size to become larger. For example, networks of about 500 nodes are very common in typical studies [2, 3]. Visualization enables the whole structure grasped at a glance. For the sake of visualization, a network containing small number of genes (20 or 30) may be arranged in trial-and-error manner. However, it is not possible to deal with a network of more than 500 nodes manually. Thus, the developments of computer aided automatic graph layout generating techniques are very essential for the researchers to form new hypothesis about the biological systems.

The analysis of gene expression networks and metabolic pathways has resulted in various types of GRN models, such as Boolean Network [4], Linear Model [5], Bayesian Network [6], Neural Network [7], Differential Equations [8], Linear Time-Variant Model [9], S-system Model [10] and models including stochastic components on the molecular level [3, 11]. In the most of these models, directed arcs have been used to represent the causality relationships among genes.

For visualizing GRNs of thousands of genes, the well known and commonly used models are the spring model [12] and the fish-eye lens model [13]. In the spring model, Itoh et al. have arranged the nodes of a network in a two-dimensional plane based on the spring dynamics where nodes do not have fixed positions and the distance between the nodes changes dynamically, based on the number of nodes present in the network. The fish-eye lens model gives emphasis upon local areas to provide visualization of bio-molecular networks. The main problem associated to these models is, it becomes very difficult to arrange so many nodes due to space capacity. Moreover, either whole or part of a network can be visualized, so it becomes very hard to grasp a local area while looking over the whole structure. To overcome these difficulties, layered approach [14, 15, 16] has been proposed to visualize GRNs that uses three-dimensional layers.

The top layer only contains controlling genes where the middle layer contains both controlling and controlled genes and the bottom layer comprised with controlled genes only. Layered method tries to minimize edge-edge crossing while generating visualization for GRNs; the method does not pay any attention to node-edge crossing which is more important issue for a good visualization. The problem of node-edge crossing should be avoided, as they cause more serious problems in the biological pathways. The node-edge crossings may introduce the confusion where edges are outgoing and

Manuscript received July 19, 2014.

Sumon Ahmed is with the Institute of Information Technology, University of Dhaka, Bangladesh (e-mail: sumon@univdhaka.edu).

Mohammad Shahedul Islam is with the Department of Computer Science, University of Texas at San Antonio, USA (e-mail: MohammadShahedulIslam@gmail.com).

Nasimul Noman is with the School of Electrical Engineering and Computer Science, University of NEWCASTLE, Australia (e-mail: NASIMUL.NOMAN@newcastle.edu.au).

Kazi Sakib is with the Institute of Information Technology, University of Dhaka, Bangladesh (e-mail: sakib@univdhaka.edu).

incoming. This will result the misunderstanding of the whole genetic network structure [17]. Moreover stochastic search methods i.e. hill climbing and simulated annealing have been used which are more likely to be trapped in local minima or maxima.

To deal with these problems, in this work, a new approach has been proposed to visualize GRNs that has the following salient features for the sake of generating good visualization:

- Design of a new fitness function that gives emphasis on edge-edge crossing, node-edge crossing and node distance.
- Development of an efficient, effective, and generalized algorithm to train model parameters.
- Verification of the proposed method, using multiple data sets to generate clear visualization for GRNs which indicates the efficacy of the proposed method.

The proposed methodology offers a good compromise between the biological proximity and mathematical flexibility for generating layouts of gene regulatory networks. A natural computational approach, GA has been used in this work to identify model parameters, because it is proven to be very effective in solving different complex problems arising in different domains [18, 19].

The inference capability of the proposed method has been highlighted in different layout generating experiments with data sets of varying network sizes. The experimental result has been compared to some previous work [15] and shows the practicability of the proposed method. The proposed approach generates good layouts of gene regulatory networks containing less node-edge and edge-edge crossing. The visualization result also gives an insight about the biological clusters of genes and the interrelationships among these clusters which will enable the researchers to follow a new prototype of the biological systems.

The rest of the paper is organized as follows. The next section describes the layered approach of network layout policy. Section III explains the proposed methodology. The fourth section presents the experimental results to verify the effectiveness of the proposed method. Section V concludes the paper with some general discussions.

II. NETWORK MODEL AND LAYOUT POLICY

In a biological system, very few genes or proteins interact with another particular gene [1]. From biological observations it has been revealed that one gene is affected by other four to eight genes [15]. Thus, from the viewpoint of drawing networks, the proposed model needs to grasp the characteristics of a gene network, where each node can have at most 8 relations. A cause-effect graph has been used to visualize gene regulatory network where causality is represented by directed arcs. The direction of a arc shows the cause-effect relation between the genes.

As a visualization example consider the isomorphic graphs [20] shown in Fig. 1. having 6 nodes and 8 arcs. When the two graphs of Fig. 1. are compared to each other, the structure on the right is found easier to understand, because it contains less edge-edge crossing.

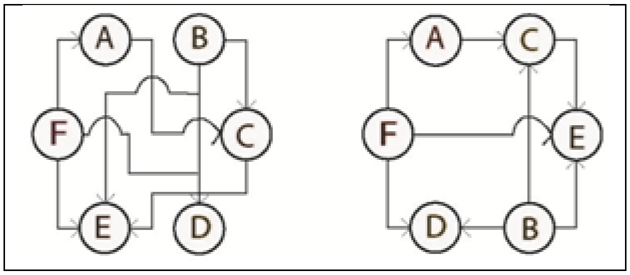


Fig. 1. Easy sample of useful layout

To generate a useful layout, genes have been classified into three categories, based on their causal relationships [15, 16], which is shown in Fig. 2. 'Parent Only' genes are placed in the top layer and are not regulated by other genes but they regulate some genes. 'Parent & Child' genes control some genes and are being controlled by some other genes. This type of genes has been placed in the middle layer. 'Child Only' genes are placed in the bottom layer and are being regulated by other genes; however they do not regulate any gene.

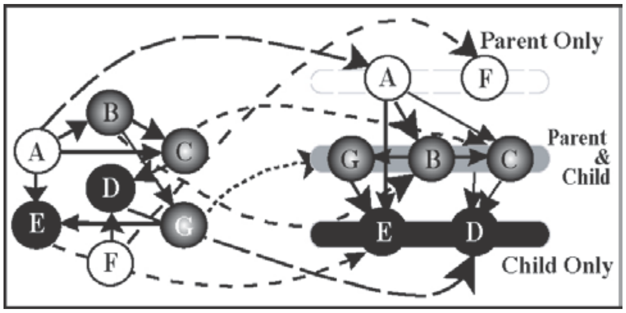


Fig. 2. Layering of genes (redrawn from [16])

Each layer comprises of equally interval co-centric regular hexagon(s) within a circle. Each hexagon has different radius with outermost having the radius equal to the circle. The nodes or genes are being positioned on the edge of these hexagons (Fig. 3.). If *n* is the number of genes in each layer and *m* is the number of nodes in each hexagonal segment, Fig. 3. suggests, the larger the value of *m*, the more genes can be placed in one layer. To get the most compact view, the minimum value of *m* is needed to be derived that satisfy the following equation [15, 16].

$$3(m-1)\,m+1 < n \leq 3m\,(m+1)+1 \qquad (1)$$

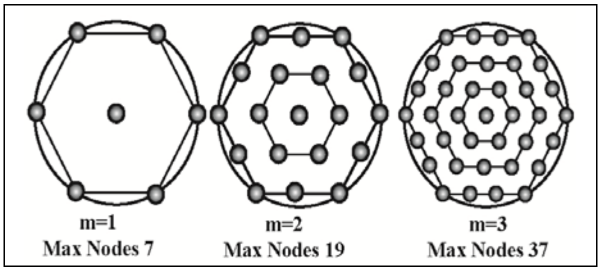


Fig. 3. Positioning of genes in different layers (redrawn from [16])

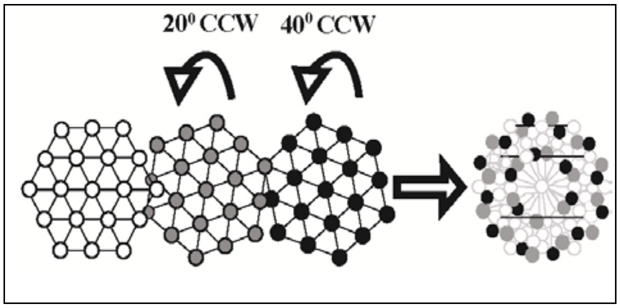


Fig. 4. Rotation of layers for easy visualization (redrawn from [16])

The three layers are overlapped and the revelation of internal layers may be hindered by the positioning of genes in the outer layers. To overcome this problem each layer is shifted by 20 degrees (Fig. 4.) to avoid overlapping nodes from a particular angle.

III. MODEL EVALUATION CRITERIA

A measurement technique is needed to evaluate the different candidate layouts which are encountered during the searching. Two entities are mainly involved in the process of generating layouts of gene regulatory networks are:

1) A mathematical model or fitness function.
2) A search method that can find the most probable outputs within the framework of the model.

A. Fitness function

For a graph $G = (V, E)$, a *layout* $L = (V, E, U, P)$ of G consists of the underling graph G and a function $P: V \rightarrow U$ with $|V| \leq |U|$ and $P(a) \neq P(b)$ for any two distinct nodes a, b in V, where V is the set of nodes, E is the set of edges, and U is the set of points. This definition does not allow overlapping of nodes in the layout. For constants *c, d < 0*, *layout cost C* (*L*) of *L* can be defined as follows [17]:

$$C(L) = \sum_{e_i,e_j \in E} EdgeCross_{e_i,e_j}(L) + \sum_{v_k \in V, e_l \in E} NodeCross_{v_k,e_l}(L) + \sum_{v_m,v_n \in V} SNode_{v_m,v_n}(L) \qquad (2)$$

Where

$$EdgeCross_{e_i,e_j}(L) = \begin{cases} c, & edge\ e_i\ crosses\ edge\ e_j \\ 0, & otherwiese \end{cases},$$

$$NodeCross_{v_k,e_l}(L) = \begin{cases} d, & edge\ e_l\ crosses\ node\ v_k \\ 0, & otherwiese \end{cases}$$

The constants *c* and *d* are called the *edge-edge crossing cost* and *node-edge crossing cost* respectively. Finally, node distance $SNode_{v_m,v_n}(L)$ is the same as S_{node}(m, n) defined in [15] for nodes v_m and v_n,

$$SNode_{v_m,v_n}(L) = \begin{cases} \left(1 - \sqrt{\left(v_{m_x} - v_{n_x}\right)^2 + \left(v_{m_y} - v_{n_y}\right)^2}\right)^2, & \left(v_{m_x} - v_{n_x}\right)^2 + \left(v_{m_y} - v_{n_y}\right)^2 \leq 1 \\ 0 & , \left(v_{m_x} - v_{n_x}\right)^2 + \left(v_{m_y} - v_{n_y}\right)^2 > 1 \end{cases} \qquad (3)$$

Basically, the problem is to minimize the distance between nodes while avoiding edge-edge and node-edge crossing. Since the node-edge cross is more serious than edge-edge cross, we have chosen d < c < 0[17]. The score remains unchanged after the upper limit of a certain distance has been exceeded, because no significant difference is observed when the viewed distance is greater than a certain level [15].

B. Genetic algorithm in layout searching

The proposed method used an enhanced *Evolutionary Algorithm (EA),* which is developed for estimating model parameters of layout searching problem. Memetic version of genetic algorithm known as *memetic Genetic Algorithm (memGA)* has been used in the core of the algorithm as the optimizer of the fitness function of (2). Like other EAs, memGA is a population-based search heuristic, where each population comprises of a certain number of individuals and each individual represents a candidate solution for the problem. A new generation is created from the current generation by using GA operators [18] and current generation is replaced by the new one. Thus memGA searches for the optimal solution by iteratively producing and replacing generations. The difference between GA and memGA is that while exploring the neighborhood, the later always preserves the best solution of each generation for the next one.

In the proposed layout searching approach, the actual 3D space is divided into three layers where each layer can contain fixed number of genes. Each layer is encoded independently using the value encoding technique [18]. Each position of a layer can be assigned to one gene and contains the information of that gene. As each individual solution consists of three layers, the encoding technique for the functions of two or more variables [18, 21] has been used to encode each candidate solution. The searching mechanism of memGA that has been used in the proposed methodology can be described as follows:

1) Generate initial population randomly P_i
2) Generate a new population P_{i+1} from P_i using crossover operator of memGA. As a result of crossover operation, more than one gene may be assigned on a single position of a layer which is not possible. To deal with this problem multi-point crossover [22] has been used in this work.
3) If the best individual of the population does not improve for G_m consecutive generations, mutation operator of memGA has been invoked to ensure that the final solution does not trap into local search space.

4) If the termination criterion is not met, the above procedure is repeated from step 2.

The algorithm terminates after a specified number of generations has been produced. Then the best individual of the final generation has been chosen as the solution of the layout search problem.

IV. EXPERIMENTS

The performance of the proposed layout searching approach has been evaluated by experimental analysis. Two different networks have been visualized by the proposed methodology, using data sets (A) and (B), given in TABLE 1. Data sets (A) and (B) are real data, inferred by an experiment using Boolean network model [15].

TABLE 1.
DATA SETS

Gene Networks	No. of genes	No. of relations	Source
A	82	84	[23]
B	552	2953	[2,24]

The proposed methodology used score functions based on equations (2) and (3) where $d = -0.8$ and $c = -0.2$ had been chosen as the model parameters in the layout searching process. The parameters for GA were chosen as follows: crossover probability, $pc = 0.8$; mutation probability, $pm = 0.1$; maximum number of generations, $G_{max} = 1000$ for data set (A) and $G_{max} = 10000$ for data set (B). Each experiment was repeated 10 times to observe the deviation of fitness scores. The small deviation (≤ 0.5) of fitness scores in the different run of the algorithm assures the soundness of the proposed layout searching technique.

A. Experimental results

Table 2 shows the comparison between the averaged values of the proposed evaluation functions and the averaged values of fitness functions used in [15]. Hosoyama et al. [15] used random search (Rnd), stochastic hill-climbing search (HC) and incremental search (Inc) in their layout searching approach. They also showed the comparative study of these three searching algorithms. This paper used memetic version of genetic algorithm as the searching technique. Although the proposed fitness function contains two penalty terms node-edge crossing, d and edge-edge crossing, c, memGA exhibits better performance than the other three algorithms used in [15], in terms of obtained fitness scores (TABLE 2). Moreover, the presence of penalty terms in our fitness functions ensure that the generated layouts will contain less node-edge and edge-edge crossing which also indicates suitability of the proposed approach.

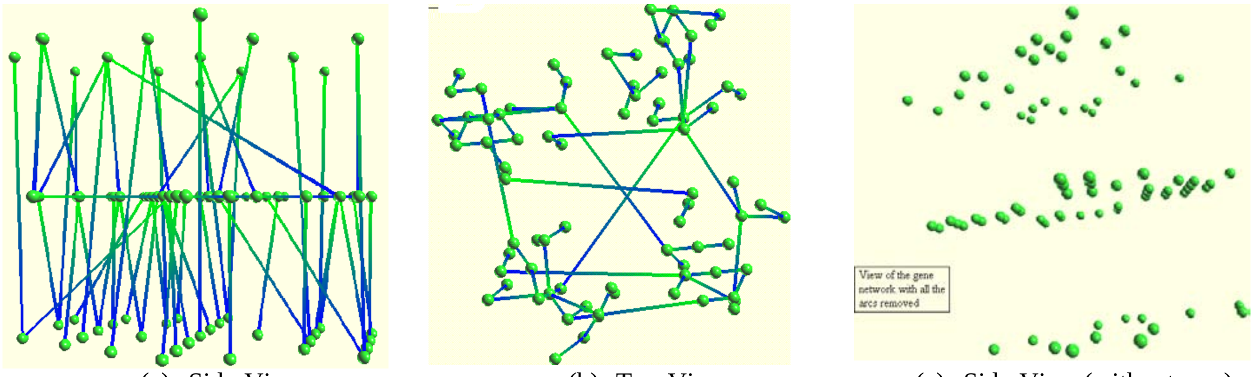
TABLE 2.
EXPERIMENTAL RESULTS

Network	Rnd	HC	Inc	memGA	
01.	A	159.4 ±1.1	256.0±1.0	257.5±1.8	259.6±0.5
02.	B	10.6±0.2	19.3±0.3	19.9±0.2	21.1±0.4

B. Visualization results

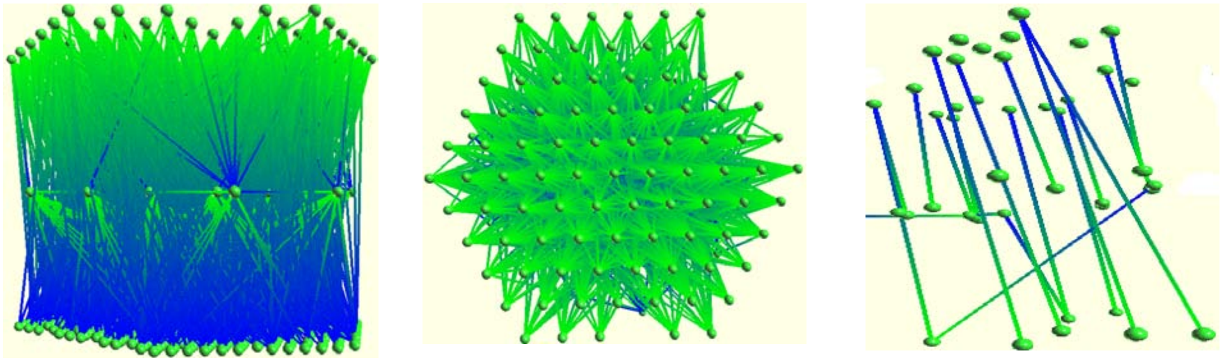
Fig. 5. shows the visualization result of the whole network for the data set (A). In the rendered graph (Fig. 5(a).), the spheres represent the genes and solid line shows the regulation relations among the genes. The regulatory direction has been shown upward and downward. Fig. 5(a). shows that the upper layer genes only regulate other genes of the middle and the bottom layers. The lower layer genes are only regulated by the genes of the top and middle layer. The middle layer genes regulate some genes as well as regulated by some other genes. Fig. 5(b). shows the top view of the network which is important for finding clusters among the genes of a network. From Fig. 5(b)., it can be observed that the gene network generated by data set (A) contains four different clusters. Moreover, the lower right quadrant of the network is connected to the upper right quadrant by one edge only. Fig. 5(c). shows the side-view of the generated network without any regulatory input. It also focuses the fact, that the whole network structure is divided into three layers.

Fig. 6. shows the generated gene network for data set (B). In this case it is very difficult to grasp the whole structure as there are 2953 relations among the genes and all the arcs have been drawn. However, if only a few genes and their adjacent area are of particular interest then the relative neighboring nodes can be crowed within a certain distance, as shown in Fig. 6(c).



(a) Side-View (b) Top-View (c) Side-View (without arcs)

Fig. 5. Visualization results of 82 genes



(a) Side-View (b) Top-View (c) Sub-network

Fig. 6. Visualization results of 552 genes

V. CONCLUSION

This paper comes with the description of conventional visualization methods and their related problems. As a solution to these problems, a new approach of visualizing gene regulatory network has been proposed which arranges the nodes in a 2-dimensional plane and combines it with a 3-dimension structure. A new fitness function has been presented that ensures the minimization of node-edge and edge-edge crossing. Memetic GA, an EA based search heuristic, has been integrated in the proposed methodology, because of its reputation of fast convergence in complex search space.

The proposed methodology has been applied on two sets of gene network data and it has successfully visualized the proper layouts of corresponding networks. The analysis of experimental results shows that in terms of fitness function values the proposed methodology outperforms some other visualization techniques such as hill-climbing, incremental search, etc. [15]. With proposed algorithms, gene networks are arranged on 3-D space so that edge-edge crossings and node-edge crossings are reduced. Nevertheless, for understanding the behavior of genes comprehensively, they should be considered in the clusters rather than individually. The visualization output of the proposed method also yields some insight of detecting biological clusters and their interrelationships.

A good layout encompasses some basic characteristics (i.e. less node-edge and edge-edge crossings) which stimulate the proposed fitness function having some user defined penalty terms. The value of these control parameters may vary from one experimentation to another and it may need to run many simulations to determine the correct parameter values. To avoid this situation, multi-objective optimization algorithm can be incorporated in the proposed method. However, the search becomes very complicated for large scale networks, as the search space increases very rapidly with the size of target network. Decoupling the original method not just creates the recognition of larger networks computationally feasible, but also facilitates the immediate parallelization or distributed implementation of the layout searching approach.

VI. REFERENCES

[1] N. Noman, and H. Iba, “Inferring gene regulatory networks using differential evolution with local search heuristics”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* vol. 4, no. 4, pp. 634-647, 2007.

[2] S. Aburatani, et al., “Discovery of novel transcription control relationships with gene regulatory networks generated from multiple-disruption full genome expression libraries”, *DNA Research*, vol. 10, pp. 1-8, 2003.

[3] H. H. McAdams, and A. Arkin, “Stochastic mechanisms in gene expression”, *Proceedings of the National Academy of Sciences,* vol. 94, no. 3, pp. 814-819, 1997.

[4] D. Sahoo, D. L. Dill, A. J. Gentles, R. Tibshirani, and S. K. Plevritis, “Boolean implication networks derived from large scale, whole genome microarray datasets”, *Genome Biology*, vol. 9, no. 10, R157, 2008.

[5] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, “Linear modeling of mRNA expression levels during CNS development and injury”, In *Pacific symposium on biocomputing,* vol. 4, no. 1, pp. 41-52, 1999.

[6] J. Mazur, and L. Kaderali, “The importance and challenges of Bayesian parameter learning in systems biology”, In *Model Based Parameter Estimation*. Springer Berlin Heidelberg, pp. 145-156, 2013.

[7] N. Noman, L. Palafox, and H. Iba, “Reconstruction of gene regulatory networks from gene expression data using decoupled recurrent neural network model”, In *Natural Computing and Beyond*. Springer Japan, pp. 93-103, 2013.

[8] T. Chen, H. L. He, and G. M. Church, “Modeling gene expression with differential equations”, In *Pacific symposium on biocomputing,* vol. 4, no. 29, 1999.

[9] M. Kabir, N. Noman, and H. Iba, “Reverse engineering gene regulatory network from microarray data using linear time-variant model”, *BMC bioinformatics*, vol. 11, no. 1, S56, 2010.

[10] A. R. Chowdhury, M. Chetty, and N. X. Vinh, “Incorporating time-delays in S-System model for reverse engineering genetic networks”, *BMC bioinformatics,* vol. 14, no. 1, 196, 2013.

[11] M. A. Savageau, *Biochemical Systems analysis: a study of function and design in molecular biology.* Addison Wesley Reading, 1976.

[12] T. Itoh, K. Inoue, J. Doi, Y. Kajinaga, and Y. Ikehata, “An Improvement of force-directed graph layout method”, *Information Processing Society of Japan,* CG103:2, 2001.

[13] T. Toyoda, Y. Mochizuki, and A. Konagaya, "GSCope: a clipped fisheye viewer effective for highly complicated biomolecular network graphs." *Bioinformatics,* vol. 19, no. 3, pp. 437-438, 2003.

[14] N. Hosoyama, and H. Iba, "3-D Visualization of a gene regulatory network: stochastic search for layouts", *Proc. of IEEE Conference on Electronic Commerce*, 2003.

[15] N. Hosoyama, N. Nasimul, and H. Iba, "Layout search of a gene regulatory network for 3-D visualization", *GENOME INFORMATICS SERIES*, pp. 104-113, 2003.

[16] N. Noman, K. Okada, N. Hosoyama, and H. Iba, "Use of Clustering to improve layout of gene network for visualization", *Congress on Evolutionary Computation,* (CEC2004), vol. 2, pp. 2068 – 2075, 2004.

[17] M. Kato, M. Nagasaki, A. Doi, and S. Miyano, "Automatic drawing of biological networks using cross cost and subcomponent data", *Genome Informatics*, vol. 16, no. 2, pp. 22-31, 2005.

[18] M. Negnevitsky, *Artificial Intelligence: A guide to Intelligent Systems.* Pearson Education Limited, England, 2002.

[19] L. D. (Ed.) Chambers, *Practical handbook of genetic algorithms: complex coding systems,* vol. 3, CRC press, 2010.

[20] D. B. West, *Introduction to graph theory* (Vol. 2). Upper Saddle River: Prentice hall, 2001.

[21] http://www.civil.iitb.ac.in/tvm/2701_dga/2701-ga-notes/gadoc/gadoc.html, accessed on 15/03/2014.

[22] K. A. De Jong, and W. M. Spears, "A formal analysis of the role of multi-point crossover in genetic algorithms", *Annals of Mathematics and Artificial Intelligence*, vol. 5, no. 1, pp. 1-26, 1992.

[23] T. Akutsu, S. Miyano, and S. Kuhara, "Algorithms for inferring qualitative models of biological Networks", *The Pacific Symposium on Biocomputing*, vol. 5, pp. 290-301, 2000.

[24] C. J. Savoie, et al., "Use of gene networks from full genome microarray libraries to identify functionality relevant drug-affected genes and gene regulation cascades", *DNA Research*, vol. 10, pp. 19-25, 2003.

Sumon Ahmed, completed his bachelor and masters degree of science from the department of Computer science and Engineering, University of Dhaka. His research interest covers machine learning, computational biology and bio-informatics. Currently he is working as a Lecturer in the Institute of Information Technology, University of Dhaka.

Mohammad Shahedul Islam, completed his bachelor and masters degree of science from the department of Computer science and Engineering, University of Dhaka. His research interest includes computational geometry and bio-informatics. His current affiliation is with the Department of Computer Science, University of Texas at San Antonio, USA as a PhD student.

Nasimul Noman achieved his bachelor and masters degree of science from the department of Computer science, University of Dhaka. He received his PhD from the University of Tokyo, Japan. His research interest covers in the field of Bioinformatics and Evolutionary Algorithms. Currently he is working with the School of Electrical Engineering and Computer Science, University of NEWCASTLE, Australia.

Kazi Sakib, completed his bachelor and masters degree of science from the department of Computer science, University of Dhaka. He achieved his PhD from RMIT University, Australia. His research interest includes distributed systems and software engineering. Currently he is working as an associate professor, Institute of Information Technology, University of Dhaka.