Dhaka Univ. J. Eng. & Tech. Vol. 1(2) 125-129, 2011 (January)

Reconstructing Gene Regulatory Network Using Linear Time-Variant Model

Sumon Ahmed, Md. Mahmudul Hasan and Nasimul Noman

Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh

Received on 21. 03. 2010. Accepted for publication on 29. 01. 2011

Abstract

With the advent of high throughput DNA microarray technology, the field of functional genomics has been revolutionized by the large amount of gene expression data generated in recent years. The analysis of these large-scale data has become very useful for investigating gene functions and the interactions among the genes. However, there are few known data analysis techniques capable of fully exploiting this new class of data. In this research work, we have presented a multi-objective evolutionary strategy for efficiently attaining the skeletal structure of the biomolecular networks and estimating the effective regulatory parameters from the gene expression time-series data using the linear time-variant formalism. Here, *Elitist Differential Evolution for Multi-objective Optimization*, a versatile, robust and well-known Multi-Objective Evolutionary Algorithm has been used. The suitability of the proposed method has been verified in gene network reconstruction experiments, varying the noise level present in gene expression profiles. And finally, we have applied the methodology for analyzing the real expression dataset of SOS DNA repair system in *Escherichia coli* and succeeded to reconstruct the network of some key regulators.

Keywords: Gene Regulatory Network, Reverse Engineering, Multi-objective Optimization

1. Introduction

Gene Regulatory Networks (GRNs) can be defined as the functioning circuitry in living organisms at the gene level which are abstract mapping of the more complicated biochemical networks and represent the regulatory relationships among genes in a cellular system.

By using gene expression arrays, it is possible to measure mRNA expression levels of thousands of genes simultaneously. In previous studies [1, 2, 3, 4, 5, 6, 7], the proper analysis of these gene expression data in a time series paradigm have been proved very useful for investigating regulatory interactions among genes. The task is challenging because of the noise presented in the microarray data and gene networks are typically hidden within the thousands of genes found in the genomes. Therefore, identifying gene regulatory networks from gene-expression data is now an extremely active research field in System Biology [1].

Many different methods for inferring bio-molecular networks from time-series microarray data have been proposed in recent literature, such as Boolean Network [2], Linear Model [3], Bayesian Network [4], Neural Network [5], Differential Equations [6] and models including stochastic components on the molecular level [7]. The common problem related with all of these models is scarcity of data, that is the number of genes far exceeds the number of time points for which data are available, making the problem of inferring GRN structure a difficult and ill-posed one.

Two major challenges in reconstructing GRNs are 1) detecting the sparse topological architecture of biological networks and 2) estimating the regulatory parameters from a limited amount of gene expression data corrupted with a significant level of noise. To cope with these problems, we have developed an Evolutionary Algorithm (EA) based inference method using linear time-variant formalism. Among several linear formalisms, the linear time-variant model is of particular interest because of its capability of discovering non-linear interactions among genes (a very common phenomenon in biochemical networks) even with noisy gene expression profiles requiring much less time than the non-linear formalisms. A multi-objective evolutionary algorithm (elitist DEMO) has been introduced for optimizing the parameters of regulation in gene networks with the aim of providing a method that can fulfill the experimental requirements. The primary contributions of this paper are as follows:

- design of a new objective function, for identifying the skeletal network structure more precisely,
- development of an efficient, effective, and generalized algorithm that does not require any user-defined parameter,
- verification of the proposed method by reconstructing GRNs from both synthetic and real gene expression data to show its efficacy in estimating the correct network architecture and the actual regulatory parameter values.

Moreover, the results of various experiments demonstrate that our proposed method requires much less time compared to other existing methods for reconstructing GRN and estimating regulatory parameters.

2. Linear Time-Variant Model for Gene Network

The biochemical systems are inherently non-linear in nature. If the model is assumed as linear time-variant, then the total regulatory input to gene-*i*, can be expressed as,

$$Z_i(t) = \sum_{j=0}^{n} W_{i,j}(t) X_j(t), \forall i \quad (1)$$

where, $Z_i(t)$ is the total regulatory input to gene-*i*, X_i is the expression level of gene-*i* at time *t*. W is a matrix of time-varying coefficients providing information about the relationships among genes and can be used to construct underlying GRN. The weight coefficient, $W_{i,j}$ indicates the type and strength of the influence of gene-*j* on gene-*i*. Since $W_{i,j}(t)$ is a time-varying function, it can be written as a finite sum of Fourier Series [8] as follows:

$$W_{i,j} = \alpha_{i,j} \sin(\omega_i t + \varphi_{i,j}) + \beta_{i,j}, \forall_{i,j} \quad (2)$$

Here, $\alpha_{i,j}$, $\beta_{i,j}$, $\Phi_{i,j}$ and ω_i are the constants to be determined, the linear part of the interactions is represented by $\beta_{i,j}$ and the sinusoidal term approximates any non-linear terms in the interactions. Thus the linear time-variant model can be defined by the parameter set, $\{\alpha, \beta, \Phi, \omega\}$.

The response of gene-*i* to the regulatory inputs is the expression level of gene-*i* at time $t + 1$, i.e, $X_i(t + 1)$. For biological realism, the value of $X_i(t + 1)$ is obtained by normalizing Z_i using a sigmoid squashing function:

$$X_i(t+1) = \frac{1}{1+e^{-z_i(t)}} \quad (3)$$

For reconstructing a gene network modeled by linear time-variant system, usually the inference method tries to estimate $N(3N + 1)$ parameters that can mimic the experimentally obtained gene expression data.

3. Model Evaluation Criteria

3.1 Generic Fitness Evaluation Function

The generic fitness evalualtion function, *Mean Squared Error (MSE)* is used to find the gene regulatory network that best fits the experimental data. The smaller the value of MSE, the better the match between observed and calculated expression dynamics, the better the approximation.

$$f1 = \sum_{k=1}^{M} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(\frac{X_{k,i}^{cal}(t) - X_{k,i}^{exp}(t)}{X_{k,i}^{exp}(t)} \right)^2 \quad (4)$$

Here, $X_{k,i}^{exp}(t)$ represents the experimentally observed expression level of gene-*i* at time *t* in the k^{th} data set. Whereas, $X_{k,i}^{cal}(t)$ is the numerically calculated expression level of gene-*i*, at sampling time *t* in the same data set which is acquired by solving Equations (2) and (3). Here *M* is the number of experimental data sets used, *T* is the number of sampling time points and *N* represents the number of genes in the regulatory system.

3.2 Attaining the Skeletal Network Structures

In a biological system very few genes or proteins interact with a particular gene. One of the major difficulties of linear time-variant model is that the large parameter set makes the detection of the underlying skeletal system architecture difficult. Because of the high degree of freedom of the model, there exist many local minima in the search space that can also mimic the time courses very closely. Therefore, the method may get stuck on some locally optimum solution and fail to obtain the true skeletal network structure.

Here we have designed a new fitness function for generating true skeletal network structure from experimental time courses and used this fitness function as the second objective in our multi-objective inference algorithm. The value of this fitness function is calculated by summing up the number regulatory inputs of all the genes in the system. The smaller the value of this fitness function, the sparser the underlying skeletal network structure, closer approximation of the biological reality. Thus for each set of parameters representing regulation networks in linear time-variant system, the fitness function for obtaining globally optimal gene network structure has been defined as:

$$f_2 \sum_{i=1}^{N} \quad (5)$$

Here, I_i is the number of regulatory inputs to gene-*i* and *N* is the number of genes in the regulatory system.

4. Inference Method

The aim of search is to find a set of parameters $\{\alpha, \beta, \varphi, \omega\}$ that minimizes both f_1 and f_2. Steps of the proposed algorithm are similar as in NSGA-II [9] which are described below:

1) Generate initial population randomly P_t.
2) Use the mutation and crossover operation of DE [10] to generate a new offspring population Q_t.
3) Generate a combined population $R_t = P_t \cup Q_t$.
4) Evaluate the individuals of R_t using Equation (4) and (5).
5) Conduct a fast non-dominated sorting [9] to order the individuals of R_t into non-dominated fronts F_0, ...F_k, where the members of one front are non-dominated by each other, and the best non-dominated solutions are in F_0.
6) The next generation P_{t+1} is filled beginning with members of F_0 and subsequently adding the members of following fronts. If not all members of a front can be added because

otherwise *NP* (number of population) would be exceeded, it is decided based on crowding distance [9] which solutions should be kept.

7) If the fitness values (i.e., f_1 and f_2) of the best compromise individual does not improve for G_m consecutive generations, then the mutation operation is evoked, which mutates all the other individuals in the current generation. The α and β parameters of an individual are mutated by adding random numbers drawn from Gaussian distribution with mean $\mu_r = 0$ and standard deviation σ_α and σ_β, respectively, where the ϕ and ω parameters are mutated using random numbers drawn from a distribution with mean $\mu_\alpha = 0$ and standard deviation σ_α.
8) If the termination criterion is not met then the above procedure is repeated from step 2.

The output generated by any MOEA is the non-dominated set of solutions known as the *Pareto-optimal* solutions. However the decision maker may have imprecise or fuzzy goals for each objective function. Thus, upon having the Pareto-optimal set, we have used, a fuzzy based mechanism described in [11], for extracting a Pareto-optimal solution as the best compromise solution.

5. Reconstruction Experiments and Results

To see how successfully the proposed method can reconstruct the network topology and estimate the regulatory parameters, we have first applied it on artificial target networks and then for analyzing real microarray time-series data.

5.1 Artificial Network Inference

The target has been generated according to equation (6).

$$\tilde{x}_i(t) = x_i(t) \times (1+\eta), (-R \le \eta \le R) \qquad (6)$$

where N indicates the number of nodes in the network, connectivity k is the maximum number of inputs per gene and the noise percentage R indicates the maximum amount of randomly added noise to the expression level for generating the expression pattern from the target network [12].

The generated artificial network of 5 genes is presented in Figure 1. In all the figures, we have maintained the convention that → represents activation and a ⊣ represents suppression. The network contains both positive and negative regulations along with feedback loop. We conducted 10 runs for each condition using 10 sets of data and the result is shown in Figure 2. The time required for reconstructing 5-gene target network using 10 sets of data, is approximately 7 minutes on a Intel(R)Core(TM)2Duo 2.80 GHz, 2GB RAM - personal computer. The average *Sensitivity SN, Specificity SP* and MSE are given in the Table 1.

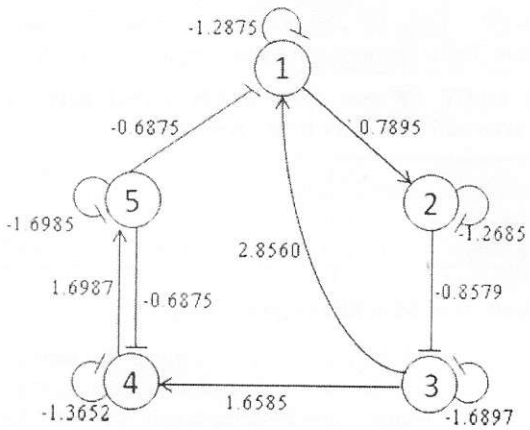
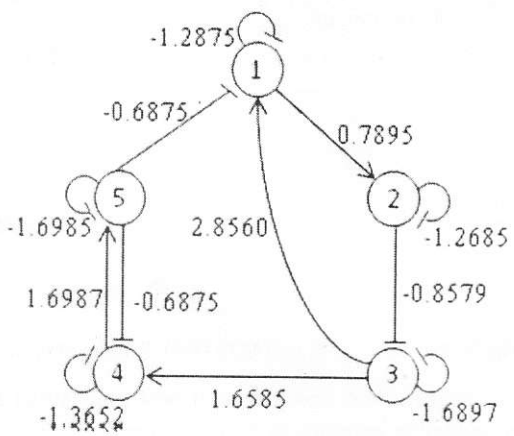
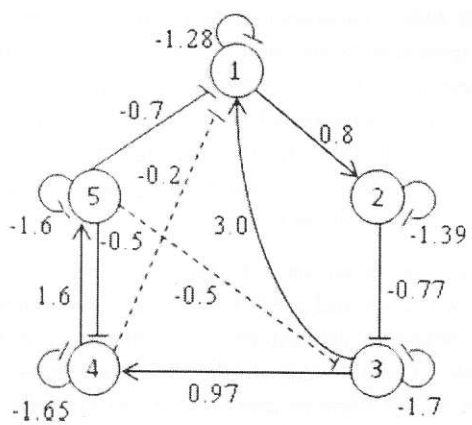


Fig. 1: Target Network.



(a)



(b)

Fig. 2: Estimated Networks from (a) noise-free and (b) 5% noisy gene expression profiles. Dashed line indicates false positive regulations.

The sensitivity and specificity is defined by following equations,

$$SP = \frac{TP}{TP + FN'} SN = \frac{TN}{TN + FN} \quad (7)$$

where TP , TN , FP and FN denote True Positive, True Negative, False Positive and False Negative, respectively.

Table 1: *SN, SP* and MSE of the target network for noise-free and 5% noisy time-series data

	SN	*SP*	MSE
Noise-free	1.0	1.0	10^{-12}
5% Noisy	1.0	0.96	1.13

5.2 Analysis of Real Microarray Data

We have applied the algorithm to analyze the well-known SOS DNA repair network in *E. coli* as shown in Figure 3. Genes are in lowercase and Proteins are in uppercase letter.

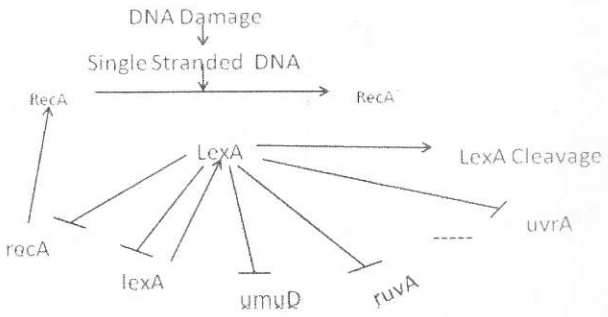


Fig. 3: The bacterial *E. coli* SOS DNA Repair network.

It is the longest, most complex and best understood DNA damage-inducible network to be characterized to date. The expression data of the SOS DNA repair system have been downloaded from the homepage of Uri Alon Lab [13]. Although Alon's experimental data contains 4 sets of time-series expression levels, here in this study only the data from experiment 3 and 4 have been considered. The data corresponding to each gene have been normalized within the range (0, 1] against their maximum value. Also a very small value (~ 10^{-4}) has been used to replace all the zero expression levels in these two data sets.

We have considered only 6 genes namely *uvrD, lexA, umuD, recA, uvrA* and *polB*. Being actual microarray data, there is unknown amount of noise inherently present in these data. These noises in the data may have had an influence on the inference method. So, the generated results have been much dispersed. The results have been generated based on the different runs of the algorithm.

The known regulations and the predicted regulations for all the 6 genes in the SOS repair network identified by the proposed algorithm have been summarized in Table 2.

Table 2: Estimated regulations for SOS DNA repair system

Gene	Predicted Regulations	References
uvrD	uvrD ⊣ uvrD, lexA → uvrD, uvrA → uvrD	[14, 15, 16]
lexA	uvrD lexA, lexA ⊣ lexA, umuD → lexA,recA ⊣ lexA	[14, 15, 17]
umuD	lexA ⊣ umuD, recA ⊣ umuD	[16, 17, 18]
recA	lexA → recA, umuD → recA, recA ⊣ recA, polB → recA	[14, 17]
uvrA	lexA → uvrA, recA ⊣ uvrA, uvrA ⊣ uvrA	[4, 16, 19]
polB	LexA ⊣ polB, uvrA → polB	[16, 19]

The computational time for the algorithm in predicting the SOS repair network ~35 minutes on a Intel(R)Core(TM)2Duo 2.80 GHz, 2GB RAM - personal computer which is much shorter than some previous works. Whereas the S-tree based system [20] running on the computer system Athlon Mp2800+ took about 35 hour for inferring this network. The method proposed by Gardner in [18] took approximately 1hour for reconstructing this network.

6. Conclusion

In this research work, an inference methodology has been developed for addressing the challenge of reconstructing molecular pathways of gene regulation from gene expression time-series data. The performance of the proposed framework makes it more applicable to the problem of reconstructing gene regulatory networks. The method has been verified by both synthetic and real expression data and it is easily extendible on larger networks. For synthetic network, the method finds the true regulations even in the presence of 5% noise. In actual SOS DNA repair network data, the proposed method outperforms some other methods [18, 20] as it finds the regulations in comparatively faster execution time. For dealing with the problem of high dimensionality and for parallel execution of algorithm, the original model of this research work may be decoupled into N sub problems, where N is the number of genes in the system.

References

1. Segal, E., M. Shapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman, "Module networks: Discovering regulatory modules and their condition specific regulators from gene expression data", Nature genetics, vol. 34, no. 2, pp. 166 - 176, 2003.

2. Sahoo, D., D. Dill, A. Gentles, R. Tibshirani, and S. Plevritis, "Boolean implication networks derived from large scale, whole genome microarray datasets", Genome Biology, vol. 9, no. 10, p. R157, 2008.

3. Dhaeseleer, P., 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, pp. 41 - 52, 1999.

4. Perrin, B., L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc, "Gene networks inference using dynamic Bayesian networks", Bioinformatics, vol. 19, no. 2, pp. 138 - 148, 2003.

5. Vohradsky, J., "Neural model of the genetic network", Journal of Biological Chemistry, vol. 276, no. 39, p. 36168, 2001.

6. Chen, T., H. He, and G. Church, "Modeling gene expression with differential equations", in Pacific Symposium on Biocomputing, vol. 4, pp. 29 - 40, 1999.

7. McAdams, H. and A. Arkin, "Stochastic mechanisms in gene expression", Proceedings of the National Academy of Sciences, vol. 94, no. 3, p. 814, 1997.

8. Proakis, J. and D. Manolakis, "Digital Signal Processing: Principles, Algorithms, and Applications", 1996.

9. Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II", IEEE TRANS EVOL COMPUT, vol. 6, no. 2, pp. 182 - 197, 2002.

10. Storn, R. and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces", Journal of global optimization, vol. 11, no. 4, pp. 341 - 359, 1997.

11. Abido, M., "Multiobjective evolutionary algorithms for electric power dispatch problem", IEEE transactions on evolutionary computation, vol. 10, no. 3, pp. 315 - 329, 2006.

12. Ando, S. and H. Iba, "Inference of gene regulatory model by genetic algorithms", in Proc. Congress on Evolutionary Computation (CEC 2001), vol. 1, pp. 712 - 719, 2001.

13. http://www.weizmann.ac.il/mcb/UriAlon/, accessed on January 18, 2010.

14. Shuhei, K., S. Katsuki, Y. Soichiro, M. Hideki, M. Koki, and H. Mariko, "Function approximation approach to the inference of reduced NGnet models of genetic networks", BMC Bioinformatics, vol. 9, no. 1, p. 23, 2008.

15. Cho, D., K. Cho, and B. Zhang, "Identification of biochemical networks by S-tree based genetic programming", Bioinformatics, vol. 22, no. 13, p. 1631, 2006.

16. Shuhei, K., N. Satoshi, and H. Mariko, "Genetic network inference as a series of discrimination tasks", Bioinformatics, vol. 25, no. 7, pp. 918-925, 2009.

17. Bansal, M., G. Gatta, and D. Di Bernardo, "Inference of gene regulatory networks and compound mode of action from time course gene expression profiles", Bioinformatics, vol. 22, no. 7, p. 815, 2006.

18. Gardner T., D. di Bernardo, D. Lorenz, and J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling", Science, vol. 301, no. 5629, p. 102, 2003.

19. Kabir, M., N. Noman, and H. Iba, "Reverse engineering gene regulatory network from microarray data using linear time-variant model", BMC Bioinformatics, vol. 11, no. Suppl 1, p. S56, 2010.

20. Cho, R., M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, *et al.*, "A genomewide transcriptional analysis of the mitotic cell cycle", Molecular cell, vol. 2, no. 1, pp. 65 - 73, 1998.