A Novel Genetic Programming Approach for Inferring Gene Regulatory Network

M.N.Vamsi Thalatam#
Associate Professor
Department of CSE
GVP College for Degree & PG Courses,
Visakhapatnam, AP, India-530045,
#Correspondence author, Research Scholar,
JNTUH, Hyderabad.

Allam Appa Rao, PhD
Director, CR Rao Advanced Institution for
Mathematics, Statistics &Computer Science
University of Hyderabad Campus,
Hyderabad, AP, India.

ABSTRACT

Computational intelligence (CI) techniques are well suited to many of the problems arising in biology as they have flexible information processing capabilities for handling huge volume of real life data with noise, ambiguity, missing values, and so on. To process the executed knowledge through CI approaches needs techniques from computer science & engineering, mathematics and statistics. It involves in- depth study with in the areas of genomic signals, gene regulation and homeostatic regulation. The evolutionary computing techniques like genetic programming plays vital role to effectively resolve several crucial problems related to genomics. The core objective of this work is to study and model the different regulation mechanisms involved in the living organisms and propose accurate evolutionary algorithms for inferring Gene Regulatory Networks.

Keywords: Computational intelligence, genetic programming, Gene Regulatory Networks.

1. INTRODUCTION

The ultimate goal of the genomic revolution is to understand the genetic causes behind phenotypic characteristics of organisms. The availability of genome-wide gene expression technologies helps to identify the interactions between genes in a living system, or gene regulatory networks. A gene regulatory network is a blue print for understanding the functional properties of genes. These networks can be used for simulating and even predicting future behavior of the system. Studying the interactions between these genes or proteins can help in finding molecular targets of specific drugs or drugs for specific targets.

Underlying concept of Gene Regulatory Network, gene expression and meaning of the short time gene expression data were studied. In this present work the problem solving approach is come under the concept of reverse engineering, because the gene regulatory network is inferred from the time series gene expression data through the novel genetic programming algorithm and differential equation models [1].

Several evolutionary methods are available in literature for inferring gene regulatory networks from time series data [2]. The summary of these approaches and their contributions towards the problem were presented below.

HIDDE DE JONG [3] published a literature review on modeling and simulation of genetic regulatory networks. In this review he quoted that to understand the functionality of organisms at molecular level, one should know which genes are expressed and which genes control the expression of the others.

The reviewer [3], described the GRN in terms of directed and undirected graph, similar definition is described by Das et al in the research book computational methodologies in GRN [4].Ordinary differential equations (ODEs) have been widely used to analyze genetic regulatory systems[Smolen et al., 2000][5].

Christopher and Wild (2011) [6] presented a review with title "How to infer gene networks from expression profiles, revisited". In this review the authors described different approaches for inferring GRNs from time series data, in which the author focused on the IRMA network evaluation.

F.He, et.al described the process of verification of gene networks through reverse engineering [7]. The well established reverse engineering approaches BANJO (Bayesian networks) [8], NIR (Network Identification by multiple Regression) [9] and TSNI (Time Series Network Identification) [10] tested on the IRMA time series data set.

A Ferinaccio et. al (2013) [11] reported that GRNGen model has shows better accuracy compared to BANJO,NIR and TSNI approaches.

I.Cantone et.al [12] presented the details of IRMA network.

1.1 Gene expression data and GRN Models:

The gene expression data is derived from microarray technology. The gene expression is the mRNA concentrations of genes measured in microarray experiments. A specific gene is expressed means the corresponding state of the gene is 'ON' otherwise 'OFF. These state ON can be represents with binary bit '1' and the state OFF can be represented with binary bit '0', for the sake of understanding. These gene expressions were measured during a discrete time intervals. Such gene expression data is called time series gene expression data. Basically gene expression data were measured in short time samples.

Models of differential equations (DE) are taken as gene regulatory networks and infer these networks through NGP_GRN. In this work the time series gene expression data of genes involved in benchmark network called IRMA network is taken as an input. IRMA means In-vivo Reverse-engineering and Modeling Assessment yeast network.

This network consists of five genes as shown in Fig (1). It is called Gold Standard Network. The gene expression data of

genes related to this network is defined in two conditions. These conditions are "Switch on" data and "Switch off" data. (Cantone et. al[12]).

With this proposed approach the IRMA network was inferred and compared the result with Gold standard and other state of the art approaches like GRNGen, BENJO, NIR and TSIN.

This method was tested not only on IRMA network but also on the standard gene expression data set of Rat Central Nervous System (CNS). This Rat data set consists of 112 genes and for each gene 9 time points are presented.

GABA family of genes is a part of these 112 genes of Rat Central Nervous System which forms a network of size 17 genes (Patrik D'haeseleer et. al)[13]. This GABA network is shown in fig (2).

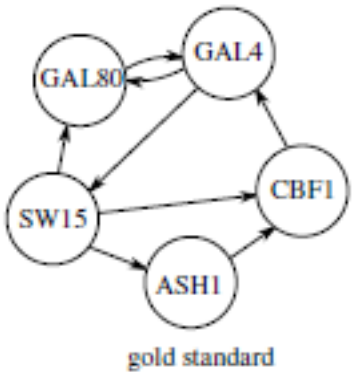


Fig (1): IRMA Gold standard Network.

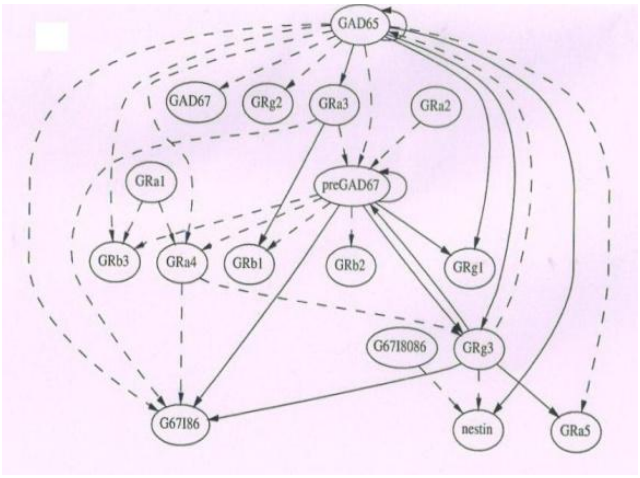


Fig 2: GABA Signaling family of Rat CNS [13]

2. METHODOLOGY

In this present work, an attempt is made to inferring the gene regulatory network by proposing a Novel Genetic Programming (NGP) on time series gene expression data. Models of differential equations (DE) are taken as gene regulatory networks and infer these networks through NGP_GRN. The time series gene expression data of genes involved in IRMA five gene of gold standard network is taken as input of this method. The genes are chosen from the standard benchmark network models available in the literature. Once the genes are selected then apply the NGP_GRN approach on IRMA gene expression data set of both Switch ON data as well as switch OFF data. Similarly the NGP_GRN is applied on GABA family of genes belonging to Rat CNS data set. Of course, the exact regulation of individual genes can't be explained by these models.

2.1 NGP_GRN Approach

In this work the ordinary differential equation as a system of gene regulatory network is taken as

$$\frac{d}{dt}E_i(t) = f\{E_1(t), E_2(t), E_3(t), \dots\dots\dots\dots E_n(t)\} \quad \text{Eq(1)}$$

Where Ei(t) is gene expression value of the ith gene at time t and n is the number of genes in the network and their respective gene expression values.

The fitness function applied to the gene regulatory network shown in the Eq. (1) is given below

$$Fitness(F) = \sum_{i=1}^{N}\sum_{j=0}^{T-1}[E_i'(t_0 + j\delta t) - E_i(t_0 + j\delta t)]^2 \quad \text{……..Eq. (2)}$$

t0 : the starting time

δt : the step size

N : the number of nodes (genes) in the network

T : the number of data points

The value of $E_i'(t_0 + j\delta t)$ is found by using Ranga Kutta method, according to this method,

$$E_i'(t_0 + j\delta t) = [E_i(t_0 + j\delta t) + \frac{1}{6(k1+2k2+2k3+k4]}$$

Where

k1= $f\{ E_1, E_2, E_3 \dots\dots\dots\dots E_i\}$

k2= $f\{ (E_1 + \delta t/2), (E_i + k1\ \delta t/2) \}$

for every i=2,3,4,…….

k3= $f\{ (E_1 + \delta t/2), (E_i + k2\ \delta t/2) \}$

for every i=2,3,4,…….

k4=$f\{ (E_1 + \delta t/2), (E_i + k3\ \delta t/2) \}$

for every i=2,3,4,…….

Now apply the Cross Over and Mutation on the Chosen differential equations and again applied the above mentioned methodology find the Fitness values. Now take the average of Fitness values before and after the NGP_GRN methodology.

2.2 Proposed Algorithm

Step:1 Input: Gene Expression Data of N×M order

N- Number of genes with Expression value Ei, ; i=1,2,3……N

M-Number of time points at which genes are expressed.

Step: 2 Select any two differential equation models of GRN in terms of gene expression Ei as shown in Eq(1).

Step 3: Now apply NGP_GRN for finding Fitness (F) value as in Eq(1) of Eith gene on Ejth gene at time tk, where k=1,2,3…..M.

Step 4: Take the average fitness Avg(F) of each of the ith gene with remaining (N-1) genes and store the values in an array A[a].

Where A[a]= { Avg(F)[EiEj]} for (i<j) for j= 1,2,3…N

Step 5: Do the step 4 until all nC2 combinations exhausted.

Step 6: Now sort the all combinations into an array A[a] in descending order.

Step 7: Choose a threshold p э for (a=0; a<=p, a++)

A[a]-{ Array of Max.Avg(F)}, where p<= nC2

Step 8: Then establish an edge between Ei to Ej for the values in step 7

Step 9: Connect all the nodes within the values up to p.

Step 10: End

2.3 Network Evaluation

The network accuracy is estimated using F-Score parameter. Which is a measure of overall model accuracy.

F-Score value varies between zero (0) and one (1) where zero is worst and one is best value. F-score is defined with the PPV (Positive Predictive Value or Precision) and Sensitivity (Se) (also called recall or true positive rate) parameters. These two are used to estimate the performance of algorithm (Bansal et al., 2007)[14].

F-Score= (2*PPV*Se)/(PPV + Se)

PPV and Se:

The PPV and Se parameters re defined as given below.

PPV=TP/(TP+FP) and Se= TP/(TP+FN)

TP-Number of True Positives= number of edges in the real network that are correctly inferred.

FP-Number of False Positives= number of inferred edges that are not in the real network.

FN-Number of False Negatives=number of edges in the real network that are not inferred.

3. RESULTS & DISCUSSIONS

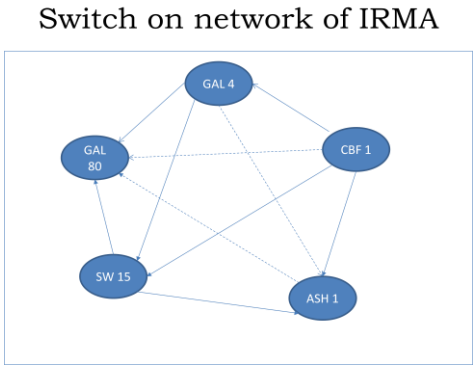
First the IRMA five gene data is taken as input with two conditions as:

1) IRMA Switch ON data, which consist of 16 time points of all five genes.

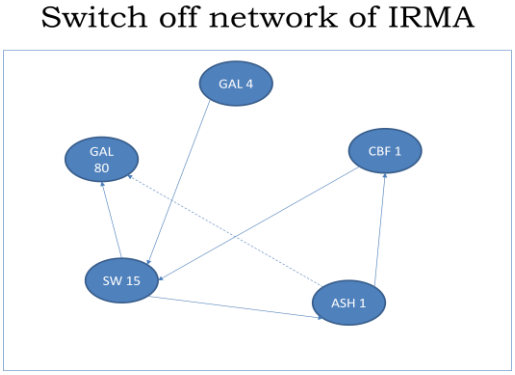
2) IRMA Switch OFF data, which consist of 21 time points

The proposed approach is applied to the first condition, i.e Switch ON data and the predicted network result is shown in Fig (3). Compared this result with the existing methods reported in literature and the comparative results are shown in table (1).

As per the IRMA network of Switch ON network is concern the PPV is relatively very less comparative to other methods. But the sensitivity is far better than the other methods so that the overall F- Score is better than the existing methods. Similarly the proposed method is applied to Switch OFF data and calculated the PPV and Se values. The result revealed that the proposed method accuracy is better comparing to the existing state of the art approaches .These obtained better results were indicated in table (2) and Fig (4).



Fig(3): Network inferred by NGP_GRN



Fig(4):Network inferred by NGP_GRN

	BANJO	NIR &TSNI	GR NGen	GeNet	NGP_GRN
PPV	0.33	0.75	0.80	0.60	0.70
Se	0.25	0.42	0.75	1.00	1.00
F-Score	0.28	0.54	0.77	0.75	**0.82**

Table[1]: IRMA Switch ON data result

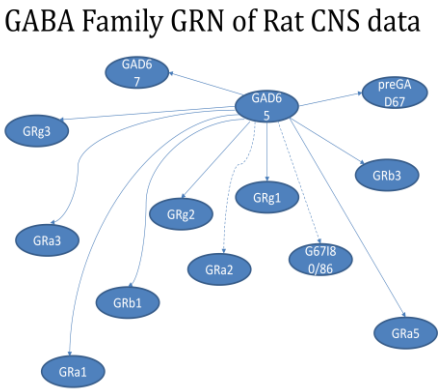
	BANJO	NIR&TSNI	GRN Gen	GeNet	NGP_GRN
PPV	0.60	0.60	0.80	0.62	0.83
Se	0.42	0.42	0.75	1.00	0.71
F-Score	0.49	0.49	0.77	0.76	**0.77**

Table [2]: IRMA Switch OFF data result

The result focused on the two data sets of IRMA (both Switch ON and Switch OFF) network shows a clear trend that the

NGP_GRN out performed by the state-of-the-art methods in terms of positive predicted value (PPV) and Sensitivity (Se). Since both cases of IRMA network shows better results as compared to the other existing methods ,this proposed method is straight away applied to the third data set belong to Rat CNS GABA family.

The inferred Gene Regulatory Network with good F-score is shown in Fig(5) and table[3].



Fig(5): Network inferred by NGP_GRN method.

F-Score of GABA family network:

PARAMETER	VALUE
TP	**8**
FP	**2**
TN	**3**
PPV	**0.80**
Se	**0.72**
F-Score	**0.76**

Table [3]: Result of GABA network.

As shown in the above table the PPV and Sensitivity are above 70% and the corresponding F-Score is 76%, so the inferred network from GABA family is considerable one and need to verify by the laboratory tests.

4. CONCLUSION

The proposed work in this paper is a novel Genetic Programming approach for inferring Gene Regulatory Network which was expressed in terms of ordinary differential model. This novel method is tested on IRMA standard time series gene expression data in both Switch ON and Switch OFF cases. The inferred network that obtained through the proposed method is validated with references to PPV, Sensitivity and F-Score Values. The results reveal that the proposed method shows better results compared to the existing four methods. With this confidence this method is applied on gene expression data of Rat CNS and inferred the Gene Regulatory Network of GABA family of Rat CNS. Hence this paper could conclude that with this novel method the bio-informaticians and biologists can infer any Gene Regulatory Network of their interest and from which they can understand the regulatory mechanism of specific genes which causes the combat diseases.

5. REFERENCES

[1] T.M.N Vamsi, et.al "Inferring gene regulatory network through genetic programming and polish notation", Proc. of Int. Conf. on Advances in Communication, Network, and Computing, CNC, Elsevier, pp: 807-814. ISBN: 978-81-910691-7-8. (2014).

[2] Khalid raza, Rafat parveen, Evolutionary algorithms in Genetic regulatory networks model, ISSN 0976-2604 ,Vol 3, Issue 1, 2012, pp 271-280, Journal of Advanced Bioinformatics Applications and Research.

[3] Hidde De Jong, " Modelling and Simulation of Genetic Regulatory Systems: A Literature Review", Journal of Computational Biology, Volume 9, number 1,Pp. 67-103, 2002.

[4] Sanjoy Das,Doina C, S M Welch, William.H Hsu, "Handbook of Research on computational methodologies in Gene Regulatory Networks", Pp. 1-5,Medical Information Science

[5] Smolen, P., Baxter, D.A., and Byrne, J.H. 2000. Modeling transcriptional control in gene networks: Methods, recent results, and future directions. Bull. Math. Biol. 62, 247–292.

[6] Christopher, David Wild Review on " How to infer Gene Networks from expression profiles, revisited, Interface Focus(2011), 1,857-870,doi:10.1098.

[7] F.He.R.Balling & A. Zeng, " Reverse engineering and verification of gene networks: Priniples, assumptions and limitations of present methods and future perspectives," vol.144, pp.190-203,2009.

[8] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational Biological data," Bioinformatics, vol. 20, pp. 3594–3603,(2004).

[9] Della Gatta et.al "Direct targets of the TRP63 TFs revealed by a combination of gene expression profiling and reverse engineering". GenomeRes.18,939-948.(2008)

[10] Bansal et.al, Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics 22, 815-822,doi: 10.1093(2006).

[11] Leonardo Vanneschi et. al, "Gene regulatory networks reconstruction from time series datasets using genetic programming: a comparison between tree-based and graph-based approaches", Genetic Programming and Evolvable Machines, Volume 14, Issue 4, pp 431-455,(2013)..

[12] Cantone,I. et.al. A Yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell 137,172-181,Elsevier Inc,April 2009.

[13] D'haeseleer, P., et al. "Linear Modeling of mRNA Expression Levels During CNS Development and Injury" Pacific Symposium on Biocomputing 4: 41-52, (1999).

[14] Bansal et.al," How to infer gene networks from expression profiles", Mol.Syst. biology.3,78 (2007).