A METHOD FOR MODELLING GENETIC REGULATORY NETWORKS BY USING EVOLVING CONNECTIONIST SYSTEMS AND MICROARRAY GENE EXPRESSION DATA

Nikola K. Kasabov[1] and Dimiter S. Dimitrov[2]

1: Knowledge Engineering and Discovery Research Institute, Auckland University of Technology,
Private Bag 92006, Auckland 1020, New Zealand; nkasabov@aut.ac.nz
2: National Cancer Institute at Frederick, NIH, Frederick, MD,USA, dimitrov@helix.nih.gov

ABSTRACT

The paper describes the problem of discovering genetic networks from time course gene expression data (the reverse engineering approach) and introduces a novel method for using evolving connectionist systems (ECOS) for this task. A case study is used to illustrate the approach. Genetic regulatory networks, once constructed, can be potentially used to model the behaviour of a cell or an organism from initial conditions.

1. EVOLVING PROCESSES IN MOLECULAR BIOLOGY AND THE PROBLEM OF GENETIC NETWORK DISCOVERY

In a single cell, the DNA, the RNA and the protein molecules evolve and interact in a continuous way. At the cell level evolving are all the metabolic processes, the cell growing, the cell division, etc. [1,2,3,17]. This interaction can be represented as a complex genetic regulatory network (GRN) of genes connected to each other so that the connections represent this interaction [4]. Genes can trigger other genes to over-express, or to become down-expressed, or may not have a direct relation at all.

The following issues are related to the problem:

- It is assumed that a GRN describes the regulatory interaction between genes;
- It is assumed that reverse engineering – from gene expression data to GRN, is appropriate to apply;
- It is assumed that gene expression data reflect the underlying GRN;
- If there are co-expressed genes over time – either one regulates the other, or both are regulated by same other genes;
- The time unit of interaction needs to be defined;
- Appropriate data need to be obtained;
- A validation procedure needs to be used;
- A correct interpretation of the models may generate new biological knowledge.

Several approaches have been introduced so far for the problem of genetic network discovery and modeling as presented briefly in the next section.

2. GRN MODELS – A BRIEF REVIEW

An extended review of the literature on the existing models for modelling GN is presented in [4].

There are several types of GN representation, some of them listed below:

- Boolean GRN (using Kauffman boolean networks), where boolean vectors represent the state of the genes at every time point, i.e. values of 1 or 0; this representation is too simplistic and is imprecise [5];
- Bayesian and regression networks - transitional probabilities are represented in the model [13,14];
- Connectionist networks (genes are represented as neurons and the interaction between them – as weighted connections [20,21,24];
- Fuzzy connectionist networks - fuzzy representation is used to represent the transition in a connectionist GRN network [24];

Several methods have been introduced for reverse engineering in order to detect a GN from manifestation of data:

- Detecting gene relations from MEDLINE abstracts [19];
- Analytical modeling – formulas are derived from gene data [10,15];
- Correlation analysis of gene data to find correlations between gene expression over time [12].
- Cluster analysis – genes are clustered based on their expression [7,8,9];
- Evolutionary computation – GRN are evolved from gene data based on a fitness function [11,16];
- Connectionist techniques (neural networks) are used to learn a GRN from data [20,21].

Despite of the existence of these methods, the problem of the genetic network discovery has not been solved so far. One of the reasons is that the processes are too complex for the existing computational models. Generally speaking, modeling genetic networks requires that the model evolves both its structure and functionality in time. A potential approach to apply to this task is the *evolving connectionist systems (ECOS)* approach as presented and applied in this paper.

3. EVOLVING CONNECTIONIST SYSTEMS

Evolving connectionist systems are multi-modular, connectionist architectures that facilitate modelling of evolving processes and knowledge discovery [24]. An evolving connectionist system may consist of many evolving connectionist modules.

An evolving connectionist system is a neural network that operates continuously in time and adapts its structure and functionality through a continuous interaction with the environment and with other systems according to: (i) a set of parameters P that are subject to change during the system operation; (ii) an incoming continuous flow of information with unknown distribution; (iii) a goal (rationale) criteria (also subject to modification) that is applied to optimise the performance of the system over time.

The set of parameters P of an ECOS can be regarded as a chromosome of ”genes” of the evolving system and evolutionary computation can be applied for their optimisation.

The evolving connectionist systems presented in [22-25] have the following specific characteristics: (1) they evolve in an open space, not necessarily of fixed dimensions; (2) they learn in on-line, pattern mode, incremental learning, fast learning - possibly by one pass of data propagation; (3) they learn in a life-long learning mode; (4) they learn as both individual systems, and evolutionary population systems; (5) they have evolving structures and use constructive learning; (6) they learn locally and locally partition the problem space, thus allowing for a fast adaptation and tracing the evolving processes over time; (7) they facilitate different kinds of knowledge extraction, mostly combined memory based, statistical and symbolic rule knowledge.

Some ECOS models, such as ZISC [26] and EFuNN [22] have been patented and used widely [24]. Some of the evolving connectionist models presented in [22-26] are knowledge-based models, facilitating Zadeh-Mamdani fuzzy rules (EFuNN, HyFIS), Takagi-Sugeno fuzzy rules (DENFIS), on-line fuzzy clustering (ECM).

Fig.1 shows a simplified version of an evolving fuzzy neural network (EFuNN) [22] that facilitates the extraction of rules of the type of Zadeh-Mamdani, as an example is given below:

IF x1 is High (0.7) and x2 is Low (0.8) THEN y is Medium (0.9), number of examples accommodated in the rule is 45; radius of the cluster covered by the rule is 0.5.

Each rule node captures one fuzzy rule that can be extracted at any time of the operation of the system. A rule links a cluster of data from the input space to a cluster of data from the output space and can be interpreted as knowledge.

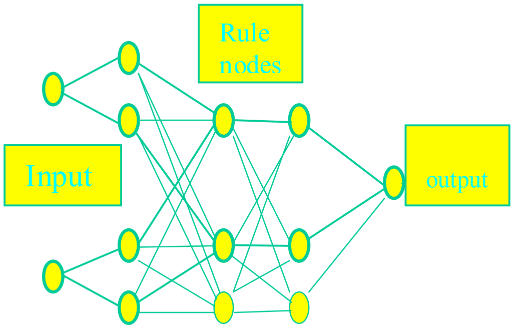

Fig.1. A simplified version of EFuNN (from [22,24])

Another type of ECOS – DENFIS [25] deals with Takagi-Sugeno fuzzy rules of the form of:

IF x1 is High (0.7) and x2 is Low (0.8) THEN y=0.5 +3.7x1 + 4.5x2, number of examples accommodated in the rule is 15; the area of the cluster covered by the rule is [0.5, 0.7].

Each evolving connectionist system consists of three main parts:

(1) Pre-processing and feature evaluation part
(2) Connectionist modelling part
(3) Knowledge acquisition part

4. A METHOD FOR GRN MODELING AND DISCOVERY USING EVOLVING CONNECTIONIST SYSTEMS AND MICROARRAY GENE EXPRESSION DATA

Genes are complex structures and they cause dynamic transformation of one substance into another during the whole life of an individual, as well as the life of the human population over many generations. When genes are “in action”, the dynamics of the processes in which a single gene is involved are complex, as this gene interacts with many other genes, proteins, and is influenced by many environmental and developmental factors.

Modelling these interactions, learning about them and extracting knowledge, is a major goal for the scientific area of computational molecular biology and bio-informatics. The whole process of the expression of genes and the production of proteins, and back to the genes, *evolves* over time.

The method proposed here and illustrated in the next section consists of the following steps:

1) Micro-array data is collected from cells in a time course manner at time moments t=0,1,2,…p.

2) A number of genes that are relevant to the process of modelling are selected, that include genes that change over the time of the cell development.

3) The genes may be grouped into grouped based on their correlation with time (or another variable) into two groups – the group of positive correlation, and the group of negative correlation, each group represented as a collective integral "gene" having the average expression level of all genes in the group.

4) The microarray data is then used to evolve a clustering-based ECOS (regardless of its type, e.g. EfuNN [22], ZISC [26]) with inputs being the expression level of a certain number of selected genes (e.g.100) and the outputs being the expression level of the same genes at the next time moment as recorded in the data.

5) After the ECOS is trained on time course gene expression data, rules that express transitions of gene states over time are extracted from it. The rule nodes in an ECOS capture clusters of input genes that are related to the output genes at next time moment.

6) The rules are linked to each other in terms of time-arrows of their creation, thus representing the GRN.

7) The extracted rules are fuzzy rules of Zadeh-Mamdani type, they represent the relationship between the gene expression of a group of genes G(t) at a time moment t and the expression of the genes at the next time moment G(t+dt), e.g. the following is a Zadeh-Mamdani type of fuzzy rule:

IF g13(t) is High (0.87) and g23(t) is Low (0.9)
THEN g87 (t+dt) is High (0.6) and g103(t+dt) is Low

8) Through modifying a threshold for rule extraction (see [24]) stronger or weaker patterns of relationship are extracted.

9) New data are added to the model in an on-line mode continuously and incrementally, so the ECOS allows for learning dynamic GRN, so that on-line, incremental learning of a GRN is possible as well as adding new inputs/outputs (new genes) to the GRN.

In another implementation of the same method from above, a Takagi-Sugeno type of fuzzy inference systems such as DENFIS [25] are applied as follows:

1) A set of DENFISi,1=1,2,..n (number of genes) models will be trained, one for each gene g_i so that input vector is the expression vector G(t) and the output is g_i(t+dt). DENFIS allows for a dynamic partitioning of the input space.

2) Takagi-Sugeno fuzzy rules, that represent the relationship between gene g_i with the rest of the genes, are extracted from each DENFISi model, e.g.:

If g1 is (0.63 0.70 0.76) and
g2 is (0.71 0.77 0.84) and
g3 is (0.71 0.77 0.84) and
g4 is (0.59 0.66 0.72) and
then g5 = 1.84 - 1.26 g1 - 1.22g2
+ 0.58g3 - 0.03 g4

5. A CASE~STUDY OF A SMALL GRN MODELING WITH THE USE OF ECOS

In a particular implementation of the method presented in the previous section, a small GRN of a leukemic cell line U937 [27] is modeled with the use EfuNN [22,24].

Retinoic acid and other reagents can induce differentiation of cancer cells leading to gradual loss of proliferation activity and in many cases death by apoptosis. Elucidation of the mechanisms of these processes may have important implications not only for our understanding of the fundamental mechanisms of cell differentiation but also for treatment of cancer. We studied differentiation of two subclones of the leukemic cell line U937 induced by retinoic acid [27]. These subclones exhibited highly differential expression of a number of genes including c-Myc, Id1 and Id2 that were correlated with their telomerase activity – the PLUS clones had about 100-fold higher telomerase activity than the MINUS clones [27]. It appears that the MINUS clones are in a more "differentiated" state. The two subclones were treated with retinoic acid and samples were taken before treatment (time 0) and then at 6 h, 1, 2, 4, 7 and 9 days for the plus clones and until day 2 for the minus clones because of their apoptotic death. The gene expression in these samples was measured by Affymetrix gene chips that contain probes for 12,600 genes. To specifically address the question of telomerase regulation we selected a subset of those genes that were implicated in the telomerase regulation and used ECOS for their analysis.

The task is to find the gene regulatory network G={g1,g2,g3,$g_{rest-,}g_{rest+}$} of three genes g1=c-Myc, g2=Id1, g3=Id2 while taking into account the integrated influence of the rest of the changing genes over time denoted as g_{rest-} and g $_{rest+}$ representing respectively the integrated group of genes which expression level decreases over time (negative correlation with time), and the group of genes which expression increases over time (positive correlation with time).

Groups of genes $g_{rest-,}$ g_{rest+} were formed for each experiment of PLUS and MINUS cell line, forming all together four group of genes. For each group of genes, the average gene expression level of all genes at each time moment was calculated to form a single aggregated variable $g_{rest.}$

Two EfuNN models, one for the PLUS cell, and one – for the MINUS cell, were trained on 5 input vector data, the expression level of the genes G(t) at time moment t, and five output vectors – the expression level G(t+1) of the same genes recorded at the next time moment. Rules were extracted from the trained structure that describe the transition between the gene states in the problem space. The rules are given in appendix and their transition in time is :

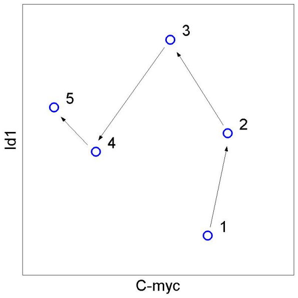

Fig.2a. The genetic regulatory network extracted from a trained EfuNN on time course gene expression data of genes related to telomerase of the PLUS leukemic cell line U937. Each point represents a state of the 5 genes used in the model, the arrows representing (rules) transitions of the states.

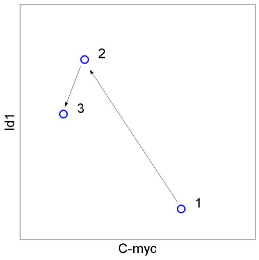

Fig.2b.The regulatory network of three time steps for the MINUS cell line represented in the 2D space of the expression level of the first two genes – c-Myc and Id1.

6. CONCLUSIONS AND FUTURE DIRECTIONS

Using the extracted rules that form a gene regulatory network, one can simulate the development of the cell from initial state G(t=0), through time moments in the future, thus predicting a final state of the cell.

Future directions include a more rigorous analysis of the theoretical limits of ECOS, building multi-modular systems of multiple sources of information, building large ECOS to model complex gene/protein complexes, building large scale adaptive decision support systems that consists of hundreds and thousands of adaptive modules.

7. REFERENCES

[1] Baldi, Bioinformarics – A Machine Learning Approach, 2001.

[2] L. Hunter, Artificial intelligence and molecular biology. Canadian Artificial Intelligence, No 35, Autumn 1994.

[3] B. Sobral, Bioinformatics and the future role of computing in biology. In: From Jay Lush to Genomics: Visions for animal breeding and genetics, 1999

[4] H. De Jong, Modeling and simulation of genetic regulatory systems: a literature review, Journal of Computational Biology, vol.9, No.1, 67-102, 2002

[5] T. Akutsu, S. Miyano, and S. Kuhara, “Identification of genetic networks from a small number of gene expression patterns under the boolean network model,” Pacific Symposium on Biocomputing, vol. 4, pp.17-28, 1999

[6] S. Ando, and E. Sakamoto, and H. Iba, “Evolutionary Modelling and Inference of Genetic Network,” Proceedings of the 6th Joint Conference on Information Sciences, March 8-12, pp.1249-1256, 2002.

[7] P. D'Haeseleer, S. Liang, and R. Somogyi, “Genetic network inference: from co-expression clustering to reverse engineering,”, Bioinformatics, vol. 16, no. 8, pp.707-726, 2000.

[8] P. D'Haeseleer, S. Liang, and R. Somogyi, “Gene expression data analysis and modelling,” Session on Gene Expression and Genetic Networks, Pacific Symposium on Biocomputing, Hawaii, Jan 4-9, 1999.

[9] S. Kauffman, “The large scale structure and dynamics of gene control circuits: an ensemble approach,” Journal of Theoretical Biology, vol. 44, pp.167-190, 1974.

[10] K. W. Kohn, and D. S. Dimitrov, “Mathematical Models of Cell Cycles,” Computer Modeling and Simulation of Complex Biological Systems}, 1999

[11] J. R. Koza, W. Mydlowec, G. Lanza, J. Yu, M. A. Keane, “Reverse Engineering of Metabolic Pathways from Observed Data using Genetic Programming,” Pacific Symposium on Biocomputing, vol. 6, pp.434-445, 2001

[12] A. Lindlof, and B. Olsson, "Could Correlation-based Methods be used to Derive Genetic Association Networks?," Proceedings of the 6thJoint Conference on Information Sciences, March 8-12, pp.1237-1242, 2002.

[13] M. Kato, T.Tsunoda, T.Takagi, Inferring genetic networks from DNA microarray data by multiple regression analysis, Genome Informatics, 11, 118-128, 2000

[14] .Gomez, S.Lo, A.Rzhetsky, Probabilistic prediction of muknown metabolic and signal-transduction networks, Genetics 159, 1291-1298, November 2001

[15] Liang, S. Fuhrman, and R. Somogyi, REVEAL: A general reverse engineering algorithm for inference of genetic network architectures," Pacific Symposium on Biocomputing, vol. 3, pp.18-29, 1998.

[16] Mimura, and H. Iba, "Inference of a Gene Regulatory Network by Means of Interactive Evolutionary Computing," Proceedings of the 6th Joint Conference on Information Sciences, March 8-12, pp.1243-1248,2002.

[17] P. A. Pevzner, Computational Molecular Biology: An Algorithmic Approach, MIT Press, 2000.

[18] R. Somogyi, S. Fuhrman, and X. Wen, "Genetic network inference in computational models and applications to large-scale gene expression data," Computational Modeling of Genetic and Biochemical Networks, in: J. Bower and H. Bolouri (eds.), {MIT} Press, pp.119-157, 1999.

[19] Z. Szallasi, "Genetic Network Analysis in Light of Massively Parallel Biological Data Acquisition," Pacific Symposium on Biocomputing, vol. 4, pp.5-16, 1999

[20] J. Vohradsky, "Neural network model of gene expression," The FASEB Journal, vol. 15, March, pp.846-854, 2001.

[21] J. Vohradsky, "Neural model of gene network," Journal of Biological Chemistry, vol. 276, pp.36168-36173, 2001

[22] N. Kasabov, Adaptive learning system and method, WO 01/78003, patented as PCT (publication date 20.04.2001)

[23] N. Kasabov, Evolving fuzzy neural networks for on-line supervised/unsupervised, knowledge–based learning, IEEE Trans. SMC – part B, Cybernetics, vol.31, No.6, 902-918, December 2001.

[24] N. Kasabov, Evolving connectionist systems: Methods and Applications in Bioinformatics, Brain study and intelligent machines, Springer, London, New York, Heidelberg, 2002.

[25] N. Kasabov and Q. Song, DENFIS: Dynamic, evolving neural-fuzzy inference systems and its application for time-series prediction, IEEE Trans. On Fuzzy Systems, vol.10, No.2, 144-154, April 2002.

[26] ZISC Manual, Silicon recognition, http://www.silirec.com

[27] Xiao, X., Phogat, S., Sidorov, I.A., Yang, J., Horikawa, I., Prieto, D., Adelesberger, J., Lempicki, R., Barrett, J.C., and Dimitrov, D.S. Identification and characterization of rapidly dividing U937 clones with differential telomerase activity and gene expression profiles: Role of c-Myc/Mad1 and Id/Ets proteins. Leukemia, 2002, 16:1877-1880

Acknowledgement

We would like to acknowledge Dr X.Xiao, Dr I.Sidorov (NCI-Frederick), Dr R.Lempicki (SAIC, Frederick) and Dr Q.Song (KEDRI, AUT) for interesting discussions and help with data.

Appendix

A. Gene regulatory rules extracted at consecutive time moments for the PLUS cell line.

Denotation: the type of the rules is: IF G(t) THEN G(T+1); [1],[2],[3],[4],[5] denote the 5 genes used in the model; 1,2 and 3 denote Small, Medium and High expression level as a fuzzy membership function; the number attached to it is the membership degree, for example [1] (2 0.299)(3 0.701) means that gene 1 is expressed at a medium level with a membership degree of 0.299 and at a High level with a degree of 0.701.

```
Rule 1:
if      [1] (2  0.299) (3  0.701)
        [2] (1  0.909) (2  0.091)
        [3] (1  0.070) (2  0.930)
        [4] (2  0.683) (3  0.317)
        [5] (1  0.731) (2  0.269)
then     [1] (2  0.091) (3  0.909)
        [2] (1  0.798) (2  0.202)
        [3] (1  0.048) (2  0.952)
        [4] (2  0.439) (3  0.561)
        [5] (1  0.838) (2  0.162)
```

Rule 2:
if [1] (2 0.091) (3 0.909)
[2] (2 0.961) (3 0.039)
[3] (2 0.955) (3 0.045)
[4] (2 0.559) (3 0.441)
[5] (1 0.836) (2 0.164)
then [1] (2 0.622) (3 0.378)
[2] (1 0.231) (2 0.769)
[3] (1 0.909) (2 0.091)
[4] (2 0.896) (3 0.104)
[5] (1 0.355) (2 0.645)

Rule 3:
if [1] (2 0.691) (3 0.309)
[2] (2 0.091) (3 0.909)
[3] (1 0.909) (2 0.091)
[4] (1 0.174) (2 0.826)
[5] (1 0.341) (2 0.659)
then [1] (1 0.311) (2 0.689)
[2] (1 0.909) (2 0.091)
[3] (1 0.244) (2 0.756)
[4] (2 0.091) (3 0.909)
[5] (1 0.909) (2 0.091)

Rule 4:
if [1] (1 0.471) (2 0.529)
[2] (1 0.131) (2 0.869)
[3] (1 0.171) (2 0.829)
[4] (2 0.091) (3 0.909)
[5] (1 0.909) (2 0.091)
then [1] (1 0.699) (2 0.301)
[2] (1 0.641) (2 0.359)
[3] (2 0.269) (3 0.731)
[4] (1 0.443) (2 0.557)
[5] (2 0.138) (3 0.862)

Rule 5:
if [1] (1 0.909) (2 0.091)
[2] (2 0.719) (3 0.281)
[3] (2 0.091) (3 0.909)
[4] (1 0.909) (2 0.091)
[5] (2 0.091) (3 0.909)
then [1] (1 0.909) (2 0.091)
[2] (2 0.091) (3 0.909)
[3] (2 0.091) (3 0.909)
[4] (1 0.909) (2 0.091)
[5] (2 0.091) (3 0.909)

B. Gene regulatory rules for the MINUS cell module (same denotation as above is used):

Rule 1:
if [1] (2 0.091) (3 0.909)
[2] (1 0.909) (2 0.091)
[3] (2 0.091) (3 0.909)
[4] (2 0.604) (3 0.396)
[5] (2 0.983) (3 0.017)
then [1] (2 0.091) (3 0.909)
[2] (2 0.091) (3 0.909)
[3] (2 0.996)
[4] (2 0.091) (3 0.909)
[5] (1 0.909) (2 0.091)

Rule 2:
if [1] (1 0.583) (2 0.417)
[2] (2 0.091) (3 0.909)
[3] (1 0.909) (2 0.091)
[4] (2 0.091) (3 0.909)
[5] (1 0.909) (2 0.091)
then [1] (1 0.840) (2 0.160)
[2] (1 0.909) (2 0.091)
[3] (2 0.091) (3 0.909)
[4] (1 0.641) (2 0.359)
[5] (2 0.810) (3 0.190)

Rule 3:
if [1] (1 0.909) (2 0.091)
[2] (2 0.757) (3 0.243)
[3] (1 0.114) (2 0.886)
[4] (1 0.909) (2 0.091)
[5] (2 0.091) (3 0.909)
then [1] (1 0.909) (2 0.091)
[2] (1 0.508) (2 0.492)
[3] (1 0.909) (2 0.091)
[4] (1 0.909) (2 0.091)
[5] (2 0.091) (3 0.909)