Bioinformatics: A Knowledge Engineering Approach

Nikola Kasabov

Abstract—**The paper introduces the knowledge engineering (KE) approach for the modeling and the discovery of new knowledge in Bioinformatics. This approach extends the machine learning approach with various rule extraction and other knowledge representation procedures. Examples of the KE approach, and especially of one of the recently developed techniques - evolving connectionist systems (ECOS), to challenging problems in Bioinformatics are given, that include: DNA sequence analysis, microarray gene expression profiling, protein structure prediction, finding gene regulatory networks, medical prognostic systems, computational neurogenetic modeling.**

Index Terms—**Bionformatics, knowledge-based neural networks, evolving connectionist systems, DNA analysis, microarray gene expression analysis, gene regulatory networks, computational neurogenetic modeling**

I. Bioinformatics: An Area of Data Growth and Evolving Knowledge

With the completion of the sequence draft of the human genome and the genomes of other species (more to be sequenced during this century) the task is now to be able to process this vast amount of ever growing dynamic information and to create intelligent systems for prediction and knowledge discovery at different levels of life, from cells to whole organisms and species.

The central dogma of the molecular biology is that the DNA (Dioxyribonucleic Acid) present in the nucleus of each cell of an organism is transcribed into RNA, which is translated into proteins [1].

Genes are complex molecular structures that 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.

Even the static information about a particular gene is very difficult to understand (see the GenBank database www.genebank.com). When genes are "in action", the dynamics of the processes in which a single gene is involved are thousand times more 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 Bioinformatics. Bioinformatics is concerned with the *application and the development of the methods of information sciences for the analysis, modelling and knowledge discovery of biological processes in living organisms.*

The whole process of the expression of genes and the production of proteins, and back to the genes, *evolves* over time. Proteins have 3D structures that evolve over time governed by physical and chemical laws. Proteins make some genes to express and may suppress the expression of other genes. The genes in an individual may mutate, change slightly their code, and may therefore express differently at a next time. So, genes may change, may evolve in the life time of an organism.

In the evolutionary processes (evolution) genes are slowly modified through many generations of populations of individuals and selection processes (e.g. natural selection).

Evolutionary processes imply the development of a sequence of generations of populations of individuals where crossover, mutation, selection of individuals, fitness (survival) criteria are applied in addition to the developmental (learning) processes of each individual.

The more new information is made available about DNA, gene expression, protein creation, metabolic pathways, etc., the more accurate their information models will become. They should adapt to the new information in a continuous way. The process of biological knowledge discovery is also evolving in terms of data and information being created continuously.

A biological system evolves its structure and functionality through both, life-long learning of an individual, and evolution of populations of many such individuals, i.e. an individual is part of a population and is a result of evolution of many generations of populations, as well as a result of its own developmental, life-long learning process.

The whole human genome is evolving as a process of development and modification over time. Named genes are same in millions of individuals but they may be expressed differently - in different individuals, and within an individual - in different cells of their body. The expression of these genes is a dynamic process depending not only on the types of the genes, but on the interaction of the individual with the environment as well (the Nurture versus Nature issue).

Many challenging problems in Bioinformatics need to be addressed and new knowledge about them revealed, to name only some of them [1-4,9]:

The work is funded by the NERF – FRST grant AUTX0201 and the School of Business at the Auckland University of technology, New Zealand (www.aut.co.nz).

The author is the Founding Director and the Chief Scientist of the Knowledge Engineering and Discovery Research Institute KEDRI (www.kedri.info) and Chair of Knowledge Engineering at the School of Computer and Information Sciences, Auckland University of Technology, Auckland, Private Bag 920010. He is also a co-founder of the Pacific Edge Biotechnology Ltd (www.peblnz.com).

Email for contacts: nkasabov@aut.ac.nz.

0-7803-8278-1/04/$20.00 ©2004 IEEE

- Recognizing patterns from sequences of DNA, e.g. promoter recognition;
- Recognizing patterns in RNA data (e.g. splice junctions between introns and exons; micro RNA structures; non-coding regions analysis);
- Profiling gene microarray expression data from RNA in different types of tissue (cancer vs normal), different types of cells, to identify profiles of diseases;
-Predicting protein structures;
-Modeling metabolism in cells;
-Modeling entire cells;
-Modeling brain diseases;
-Creating complex medical decision support systems that deal with thousands of variables to obtain the right diagnosis and prognosis for a patient.

Modeling adequately the processes in living organisms is crucial for their understanding, for the prognosis of their development, for the prognosis of drug and environmental effects [1-4]. But the enormous complexity of life needs sophisticated methods not only to model the processes, but to enable knowledge discovery as well. The paper, first reviews briefly machine learning and knowledge engineering (KE) methods, and then applies them to solving some of the above problems. As a new and a promising technique, the use of evolving connectionist systems for modeling and rule extraction from different types of biological data is presented in more details.

II. The Methods of Machine Learning (ML) and Knowledge Engineering (KE)

The methods of ML and KE include:

- Probabilistic learning methods, e.g. Hidden Markov Models [2];
- Statistical learning methods, e.g. Support Vector Machines (SVM), Bayesian classifiers [2,3];
- Case-based reasoning (e.g. k-NN; transductive reasoning) [3];
- Decision trees [3,4];
- Rule-based systems (propositional logic dated back to Aristotel) and fuzzy systems (introduced by L.Zadeh) [5,6,8]
- Neural networks (e.g. SOM, MLP, RBF) [7,8];
- Evolutionary computation (GA, ES, EP) [8,9,10];
- Hybrid systems (e.g. knowledge-based neural networks [8,9,10]; neuro-fuzzy systems; neuro-fuzzy-genetic systems [8]; evolving connectionist systems [9]).

The machine learning approach to Bioinformatics involves learning from data [3]. There are many learning techniques developed so far as shown in fig.2.

Artificial neural networks (ANN) (connectionist systems) are ML models that mimic vaguely the nervous system in its main functions of adaptive learning and generalization. They are universal computational models [7,8,9]. ANN can implement any of the ML techniques from fig.2 and hence - the variety of the ANN architectures [7,8]. Many of these architectures are known as "black boxes" as they do not facilitate revealing internal relationships between inputs and output variables of the problem in an explicit form.

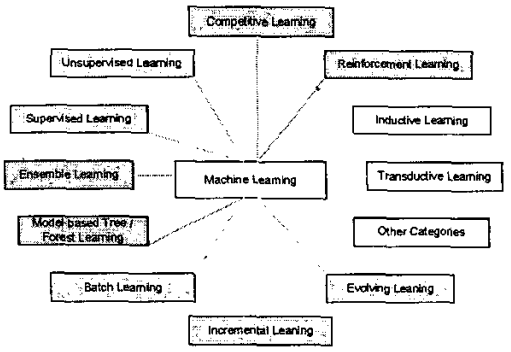

Fig. 1. A classification diagram of machine learning techniques

But for the process of knowledge discovery, having a "black box" learning machine is not sufficient. A learning system should also facilitate extracting useful information from data for the sake of a better understanding and learning of new knowledge.

The knowledge-based ANN (KBANN) have been developed for this purpose [8,9]. They combine the strengths of different AI techniques, e.g. ANN and rule-based systems, or fuzzy logic. Rules can be extracted from the KBANN as illustrated below and also – in fig.2 [8,9]:

Rule r_1: IF x1 is Small (DI11) and x2 is Small (DI21) THEN Output is Small (CF1),

Rule r_j: IF x1 is Large (DI1j and x2 is Large (DI2j) THEN Output is Large (CFj),

where: x1 and x2 are input variables, and Output is an output variable; Small and Large are fuzzy values defined by their respective fuzzy membership functions [5,6,8] and DI and CF are degree of importance (membership) and certainty factors respectively.

Evolving connectionist systems (ECOS) have been recently developed to facilitate both adaptive learning in an evolving structure and knowledge discovery [9]. ECOS are modular connectionist-based systems that evolve their structure and functionality in a continuous, self-organised, on-line, adaptive, interactive way from incoming information; they can process both data and knowledge in a supervised and/or unsupervised way. Learning is based on clustering in the input space and on function estimation for this cluster in the output space. Prototype rules can be extracted to represent the clusters and the functions associated with them.

Different types of rules are facilitated by different ECOS architectures, i.e. Zadeh-Mamdani rules [5,6] - in the evolving fuzzy neural networks EfuNN [11] – see fig.2, or Takagi-Sugeno rules - in the dynamic neuro-fuzzy inference systems DENFIS [12]. The ECOS grow and "shrink" in a continuous way from input data streams. Feed-forward and feedback connections are both used in the architecture. The ECOS are not limited in number and types of inputs, outputs, nodes, connections. All learning methods from fig.1are facilitated in different types of

ECOS that have been applied to the Bioinformatics *problems from section 1 [9].* MATLAB codes of EfuNN and DENFIS, as well as some other ECOS techniques, are available from www.kedri.info/.

III. The Knowledge Engineering Approach to Bioinformatics

The knowledge engineering approach to Bioinformatics includes the following processes:

- Data analysis and feature extraction (using statistical-, PCA-, clustering-, SNR-, and other techniques);
- Data modelling for the purpose of classification, prediction, optimisation, etc. with the use of alternative machine learning methods to chose the most suitable for the task;
- Model validation;
- Rule extraction and rule interpretation;
- Data and existing model (e.g. regression) integration;
- Multi-model system development;
- Adaptation on new data and tracing the knowledge evolution.

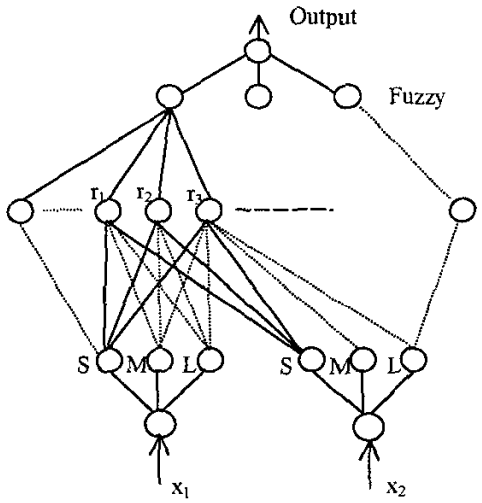

Fig. 2. A simple EFuNN structure of two inputs and one output

To facilitate the KE approach to data analysis, modelling and knowledge discovery, a software environment NeuCom has been developed and made available from www.theneucom.com, or from www.kedri.info/. NeuCom includes about 60 various techniques (see a screen snapshot in fig. 3.) that include the above discussed methods of ML and KE.

In the following sections problems from Bioinformatics are explained and solutions are shown with the use of the KE approach and the NeuCom environment in particular.

IV. A KE Approach to DNA and RNA Sequence Data Analysis and Pattern Discovery

The KE approach applied to the problem of finding informative patterns from complex sequences, such as DNA, or RNA, can generate rules that represent significant patterns of data. Such patterns can be promoter regions – regions in the DNA that bind to enzymes and cause a gene sequence to be transcribed into RNA. A promoter defines the region where transcription will begin.

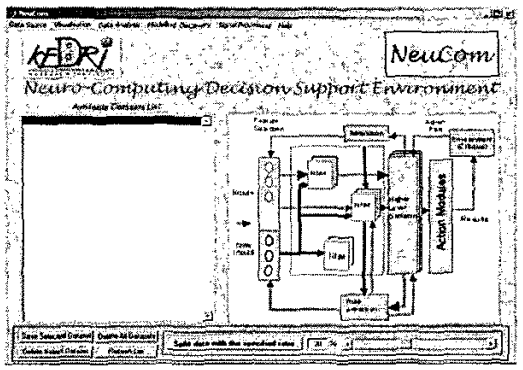

Fig. 3. A snapshot of the NeuCom environment for data analysis, modelling and knowledge discovery (from www.theneucom.com)

A KE system can be trained on promoter regions data and used to discover a new promoter sequence in either the same species or new ones [13].

A gene sequence in an RNA consists of two main parts – exons (the sections that are translated into proteins) and introns (sections that do not translate into proteins). Finding the boundaries between the two sections from an RNA sequence may help to predict what proteins and how much of them will be produced later in the cell.

In [9] an EfuNN is trained on sequence data that represent boundaries between introns and exons and then the system was used to not only predict these boundaries (junctions) on new data, but to extract rules that explain where in an RNA such boundaries are likely to occur. Two of the extracted rules are shown below:

Rule1: IF ----------*AGGT-AG*---------- ---------- *THEN [Exon Intron junction]*

Rule8: IF ----------*T*------*T-CAG*---------- ---------- *THEN [Intron Exon junction],*

where the dash positions represent unimportant information as any of the nucleic bases can appear there, but the positions where A,G,C and T bases appear are significant for the pattern [9].

V. A KE Approach to Gene Expression Data Analysis, Modelling and Knowledge Discovery

Contemporary technologies, such as gene microarrays, allow for the measurement of the level of expression of up to 30,000 genes in RNA sequences that is indicative of how much protein will be produced by each of these genes in the cell. The goal of the microarray gene expression data analysis is to identify a gene or a group of genes that are differently expressed in one state of the cell or a tissue (e.g. cancer) versus another state (normal). Generally, it is difficult to find consistent patterns of gene expression for a class of tissues.

In many research papers the machine learning approach was used to obtain a classifier or a predictor model based

on gene expression data [14-18]. A KE approach to this problem was proposed in [9,18]. Microarray data was used to evolve an EFuNN with inputs being the expression level of a certain number of selected genes (e.g.100) and the outputs being the classes (e.g. cancer and normal). After an EFuNN is trained on examples, each taken from a tissue sample, rules were extracted from the EFuNN that represent disease vs normal tissue profiles (fig.4).

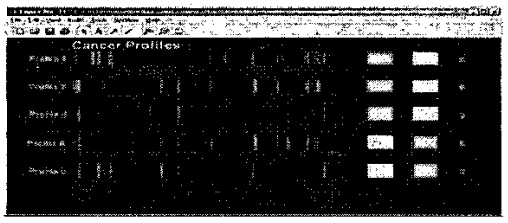

Fig. 4. Five profiles of five sub-groups of cancer tissue (the last column) versus normal tissue (the next to the left column), each of them representing a pattern of gene expression (the rest of the columns) (from [9, 18])

In this application an ECOS would adapt in time through learning from new data. The issue of data and model integration is addressed in [19] where a method for this integration is proposed and tested on Lymphoma data [14]. An EFuNN prognostic model to predict survival/fatal outcome of the disease based on gene expression of 11 genes was developed based on the available data [14]. The model was trained and tested with the leave-one-out method on the 58 data samples. 90% prognostic accuracy (88% class cured and 92% class fatal) was achieved (compared with the 75% accuracy achieved in [14]). In addition, rules that represent the gene profile of the survival versus the fatal group of patients were extracted; one of them is shown below:

Rule: IF G1 is (2: 0.8) and G2 is(2: 0.8) and X3 is (1: 0.9) and G4 is(1: 0.9) and G5 is(3: 0.9) and G6 is(1: 0.9) and G8 is(1: 0.9) and G9 is (1: 0.9) and G10 is (3: 0.9) and G11 is (1: 0.8)
THEN Class is Fatal,

where: G1,..., G11 are the selected 11 genes; 1, 2 and 3 denote Small, Medium, and Large expression value respectively, and the numbers after these values represent the fuzzy membership degrees.

A specialized gene expression profiling software that implements the KE approach called SIFTWARE, has been developed (www.peblnz.co).

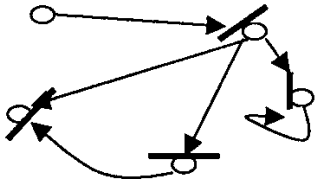

Fig. 5. A simple gene regulatory network of 5 genes and links between them that represent the interaction – either inhibition or excitation to a certain degree.

VI. A KE Approach to Gene Regulatory Networks Modelling and Discovery

In a living cell genes interact in a complex, dynamic way and this interaction is crucial for the cell behaviour. This interaction can be represented in an approximate way as a gene regulatory network (GRN) – as example shown in fig. 5.

GRN models can be derived from time course gene expression data of many genes measured over a period of time. Some of these genes have similar expressions to each other as shown in fig. 6 [9].

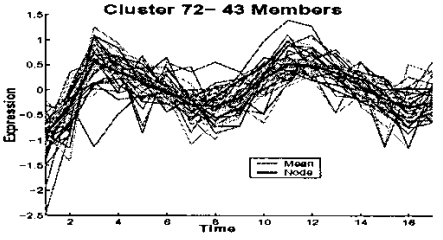

Fig. 6. A cluster of gene expression time course data of 72 genes [9]

Genes that share similar functions usually show similar gene expression profiles and cluster together. In a GRN clusters can be used and represented by nodes instead of genes or proteins. A GRN model can be used to predict the expression of genes and proteins in a future time and to predict the development of a cell or an organism [9, 20-31]. The process of deriving GRN from data is called reverse engineering [20-31]. Many methods of computational intelligence and machine learning have been used so far for the problem, that include: correlation and regression analysis, Boolean networks, graph theory, differential equations, evolutionary computation, neural networks, ect [20-31].

In [31] a KE approach, based on EfuNN and DENFIS, was used on a small data set of Leukemia cell line data to extract a GRN and to represent it as a set of rules associating the expression of the genes at time *t*, with the level of their expression in the next time moment *(t + dt)*. The rules are of Zadeh-Mamdany form when EFuNN was used , e.g.:

IF g1(t) is High (0.87) and g2(t) is Low (0.9)
THEN g3 (t+dt) is High (0.6) and g5(t+dt) is Low,

or of the Takagi Sugeno form (when DENFIS was used), e.g.:

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

where the gene expression values in the condition part are defined as triangular membership functions specified by the center, the left and the right arms of the triangles.

Extracting rules from a model gives an insight of the gene regulatory processes and helps understand better inter relationships.

VII. A *KE* Approach to Protein Data Analysis and Structure Prediction

Proteins are complex molecular 3D structures and one of their important feature, still not well known, is how they bind to each other and when – fig.7. A protein is a sequences of amino-acids, each of them defined by a group of 3 nucleotides (codons). There are 20 amino acids all together denoted by capital letters (A,C-H,I,K-N,P-T,V,W,Y). Proteins have complex 3D structures defined as: Primary (linear); Secondary (3D, defining functionality); Tertiary (energy minimization packs); Quaternary (interaction between molecules). Segments from a protein can have different shapes: Helix; Sheet; Coil (loop). The Protein Data Bank – www.rcsb.org contains a large amount of information on protein structures.

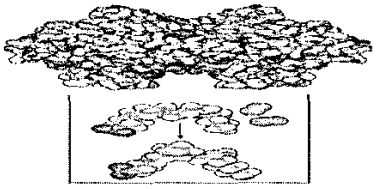

Fig. 7. Proteins are complex molecular structures that bind to each other

The primary structure of a protein can be derived from the RNA code. The prediction of the secondary structure from the primary one has been investigated in many publications. In [9] an EfuNN was used not only to evolve a protein structure prediction model, but to extract rules that are behind these structures.

VIII. Integrating Bioinformatics and Brain Study - Towards Computational Neurogenetic Modelling

GRN are important for the processes in living cells including neurons. Some genes and proteins relate directly to the activity of the neurons and thus – to the behavior of the whole ensemble of neural networks. The integration of neural networks and gene networks in computational models that are biologically plausible and thus applicable to model brain functions and diseases is called *computational neurogenetic modeling* [32-34] and constitutes a new discipline that brings the areas of molecular biology, brain science, and computer science together.

IX. Conclusions and Future Directions

The problems in Bioinformatics are too complex to be adequately modelled with the use of the current methods of ML and KE. New methods are needed in the future for: the integration of biological data – both molecular and clinical; for a personalised drug design and personalised medicine; for building embedded KE systems and implementing them into biological environments; for drug design; for computational neurogenetc modelling; and for the solution of many other challenging problems in Bioinformatics.

References

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

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

[3] Baldi, *Bioinformarics – A Machine Learning Approach*, 2001

[4] L. Hunter, *Artificial intelligence and molecular biology*, Canadian Artificial Intelligence, no 35, Autumn 1994.

[5] L. Zadeh, Fuzzy Sets. *Information and Control*, vol. 8, pp. 338-353, 1965.

[6] L. Zadeh., "A Theory of Approximate Reasoning, Machine Intelligence", vol.. 9, eds. Hayes, Michie and Mikulich, Elsevier NY , pp. 149-194, 1979.

[7] C. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[8] N. Kasabov, *Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering*. MIT Press, 1996.

[9] N. Kasabov, *Evolving connectionist systems: methods and applications in bioinformatics, brain study and intelligent machines*. Springer Verlag, 2002.

[10] G. Fogel and D.Corne, *Evolutionary Computation in Bioinformatic*. Morgan Kaufmann Publ., 2003.

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

[12] N. Kasabov and Q. Song, "DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and its Application for Time Series Prediction", *IEEE Transactions on Fuzzy Systems*, 2002, April.

[13] N. Kasabov and Shaoning Pang, "Transductive Support Vector Machines And Applications In Bioinformatics For Promoter Recognition", in Proc. of IEEE International Conference on Neural Networks and Signal Processing, Nanjing, China, pp. 1-6, Dec. 2003

[14] N. Shipp, et al, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning", *Nature Medicine*, vol. 8, no. 1, pp. 68-74, January 2002.

[15] Alizadeh, et al, "Distinct types of diffuse large B-cell lymphoma identified by gene-expression profiling", *Nature*, vol.403, pp. 503-511, February 2000.

[16] J. Khan., et al., "Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays", *Cancer Res*, vol. **58**, no. 22, pp. 5009-5013, 1998.

[17] S. Ramaswamy, , et al., "Multiclass cancer diagnosis using tumor gene expression signatures", *Proceedings of the National Academy of Sciences of the United States of America*, vol. **98**, no. 26, p. 15149, 2001.

[18] M.E. Futschik., A. Reeve, and N. Kasabov, "Evolving Connectionist Systems for knowledge discovery from gene expression data of cancer tissue", *Artificial Intelligence in Medicine*, vol. 28, pp. 165-189, 2003.

[19] M.E. Futschik, A. Reeve, and N. Kasabov, "Prediction of clinical behaviour and treatment for cancers", *Applied Bioinformatics*, vol. **2**: pp. 52-58, 2003.

[20] T. S. Akutsu, 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.

[21] S. Ando, 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] S. 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.

[27] A. 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.

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

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

[30] J. Vohradsky, "Neural network model of gene expression", *The FASEB Journal*, vol. 15, March, pp. 846-854, 200

[31] N. Kasabov and D. Dimitrov, "A method for gene regulatory network modelling with the use of evolving connectionist systems", in Proc. of ICONIP'2002 - International Conference on Neuro-Information Processing, Singapore, November 2002, IEEE Press, 2002.

[32] N.Kasabov and L.Benuskova. "Theoretical and Computational Models for Neuro-,Genetic-,and Neuro-Genetic Information Processing", in *Handbook of Theoretical and Computational Nanotechnology*, M. Rieth and W. Sommers. Eds. vol. 1, no. 1, American Scientific Publisher, 2004, in print.

[33] N.Kasabov and L. Benuskova, "Computational Neurogenetics", *International Journal of Theoretical and Computational Nanoscience*,, vol. 1, no. 1, American Scientific Publisher, 2004

[34] N. Kasabov, L. Benuskova and S. Wysoski, Computational Neurogenetic Modelling: Gene Networks within Neural Networks, in Proc. Int. Joint Conference on Neural Networks, IJCNN 2004, Budapest, 26-30 June 2004, IEEE Press, in print.