F.S.A.I., Vol. 9, Nos. 1–3, pp. 15–22, 2003

Genetics, Gene Prediction, and Neuro-Fuzzy Systems – The Context and A Program Proposal

Horia-Nicolai Teodorescu

Romanian Academy and Technical University of Iasi,
Group of Laboratories on Intelligent Systems & Bio-Medical Engineering
hteodor@etc.tuiasi.ro

Abstract: I discuss the state of the bioinformatics as related to genomic research in a European region and I propose a research development program in the field. Further, I introduce a genetic sequence identifier based on a hierarchical system using fuzzy and classic (crisp) neural networks.

1. A program proposal – International and national context

Genomics and proteomics have the tendency to become – along with the nanotechnologies and computation intelligence – the principal progress vectors of this century. After having brought a small revolution in agriculture and zoo-technology based on the Mendelian discoveries, genetics became somewhat slower in its progress, until the discovery of DNA. It has been the huge effort of the Human Genome Project which visionary projected, in the 1990, genomics as the leading science and technology for our century. Highly interdisciplinary, genomics involves a newly redefined bioinformatics, together with various chapters of biochemistry (in a broader context, proteomics is, in many respects, the key to genomics), biophysics, and cellular biology.

Genomics has a tremendous economic potential. According to *Forbes* (quoted by [1]), genomics will contribute to about 20% of the gross product of the US in 2030. Computing for genomics has seen a tremendous development in its incipient years, from about $250 M (in USA alone), in 1999, to about $2.3 Billion in 2002 (an almost 10 times increase in 3 years.) It is expected that genomics will contribute to solve numerous health, agricultural, and ecological problems, moreover will contribute to the production of gas and of other essential materials, including pharmaceuticals, at a cheap price [2].

US are the undisputed leader in the genomic research. Unfortunately, not all the European countries have grasped the importance of genomics. Moreover, many countries that have a quite powerful pharmaceutical industry or an extensive agriculture (like Romania) are lingering in this field, despite the fact that genomics would be a key for the development in those fields. A brief presentation of the principal goals of countries recently included in NATO reveals that only two see genomics as an essential goal [3]. Neither is bioinformatics seen by all the administrators as an essential tool to achieve a significant industrial development.

Unfortunately, this is also the case of Romania, and it is one of the purposes of this paper to draw attention on the fact.

2. Toward a sensible Romanian program in the context of the European projects

Under the described circumstances, designing a sensible research and development program at the regional/national scale might be wise. This paper is partly devoted to such a purpose.

Our vision is to build a significant pool of competencies in bioinformatics, disseminated into a national network of laboratories and research groups in Romania, well connected to European and international projects in bioinformatics and genomics and able to significantly contribute to the betterment of the economy in this European region. I see the possibility to build, with appropriate support, a region of excellence in bioinformatics for genomics, with strong centers in the institutes of the Romanian Academy and in the main universities.

With this goal in mind, and using our expertise, I have established about one year ago some desirable scientific objectives, that are:

i) To develop fuzzy-statistical methods to better characterize the sequences, the genes, the exons and the introns.
ii) To develop methods based on feature-oriented filtering (noise removal).
iii) To develop fast methods to identify the type of dynamics associated to a genomic sequence; use those methods to further characterize the sequences and to help detecting their specific components.
iv) To determine if series of distances between aminic bases in the DNA can be used as input to various Markov models to improve their outcome.
v) To develop a set of tools able to "travel along the genomic sequence" and to identify known (learned) patterns.
vi) To investigate the possibility of using various types of coupled-map lattices, including fuzzy-coupled map lattices as developed in [4], to analyze and predict genes.
vii) To develop companion tools that, after recognition of a pattern, can fast make alignments and comparisons of key DNA segments.
viii) To develop an integrated system using the above components and other classic components developed by other groups, to increase the performance of sequencing, analyzing and recognizing specific sequences or genes.
ix) To include the above system in a more complex one able to perform diagnosis for human, animal or plant diseases.

Subsequently, I detail the description of the above tools.

Fuzzy-statistical methods. The use of fuzzy methods to characterize genes has been advocated recently [5]. Also, several researchers have addressed the use of

fuzzy methods to identify specific sequences in the genome and to perform data mining [6]. The uncertainty in the variability of the genes and the imperfections in defining the genes themselves [7] make the fuzzy approach interesting. There are numerous fuzzy approaches in genomics, and a reasonable selection of references is [8]-[18]. However, I see benefits in combining the fuzzy approach with other types of characterizations; namely, using more than a fuzzy description that combines fuzzy and statistical methods looks reasonable. This is one of our objectives in this program.

Filtering (noise removal). Genomic data comes noisy, either in the classic sense that parasitic signals occur during physical mapping, or in the more subtle sense that a genomic sequence may include "noisy mutations" that prevents recognition by perfect matching. Methods based on feature-oriented filtering [17-22] will be investigated to remove both types of noise. Results will be compared to classical filtering methods.

Dynamics associated to genomic sequences. The use of features of dynamics associated to genomic sequences have been proposed for a long time already, emphasizing that the sequences may have periodical, random or chaotic features [5]. I suggest searching for the dynamic characteristics under various representations. Among others, I suggest using a four-sequence representation for a genomic sequence, namely for each of the bases, to represent the sequence of distances between successive occurrences. (I first used this method in analyzing the dynamics of words in natural languages, [23].) New, fast methods to characterize the dynamics will be applied, including those (like "time-in-a-region" parameter) I have proposed in other contexts [24].

Markov models. Prediction results using homogeneous or hidden Markov models have proved to be among the best and the method of Markov models is widely spread [25-29]. I aim to further develop the method, using various new representation of the genomic sequence, like the ones of distances between the successive bases.

Neuro-fuzzy tools for sequence identification. I suggest using various neuro-fuzzy tools able to recognize specific sequences. Such a tool I propose is a neuro-fuzzy predictor that learns a specific sequence. When the prediction score is high, the sequence is recognized; when the score is low, the sequence is determined to be of another type. This predictor is described in a companion paper in this issue.

Subsequently, I detail one of the components of the proposed research program.

3. A neuro-fuzzy tool for sequence identification

A method to identify changes of the statistics is to determine if some "agent" able to predict (and thus recognize) a segment becomes "unaware" of another segment that it cannot recognize (predict). A NN or any mean recognizing that a specific species genome is investigated, in a pool of genomes, or that a genome is *not* a specific genome, or that a mutation has occurred – and where – on a genome,

will probably be in demand soon. Indeed, such means will be needed to perform extensive research on large genomic databases for thousands of organisms.

The trend today is to learn the role of any genome. However, to do so, a large number of genomes from the diseased cells have to be investigated and compared to "normal" genomes. Direct comparison might be a way to determine differences. However, this requires a large number of accesses to the memory, and a large, fast memory to store the information of the "normal" and the investigated genomes. The space complexity is large for direct comparisons (at least twice that of the genome). Moreover, because the registers used to store sections of the compared strings are rather small, the number of accesses to the memory is proportional to the genome length, too.

One of the systems currently used is like in fig. 1 (after [25]) and its operation basically consists of using a set of templates to which the actual genome sections are contrasted.

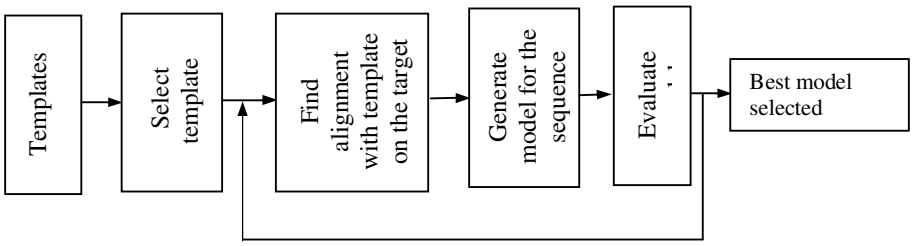

Fig. 1 – Using templates to identify genomic sequences (after [25])

An alternate system, which operation would involve a similar time complexity, but a smaller space complexity, might be achieved. Indeed, if a system or a set of systems is highly trained to "learn" normal genome sections and to be able to pinpoint that the analyzed genome section is not the one learned. Such a system may use a NN, or a fuzzy NN (FNN), knowing that NNs and FNNs are able to learn a specific context and to flag when they operate in a different context.

Instead of using FNN-based systems, other predictors could be considered. However, linear predictors and simple NN predictors can be expected to require a larger number of inputs, to efficiently learn the series. This is a major drawback when small DNA sequences are available. Moreover, a FNN has the advantage of having a significantly larger number of parameters available for training, for the same order of the predictor, compared to linear or NN predictors. This is an intrinsic advantage for the available equivalent memory per unit of order of the predictor (per leg), to store the information on the learned sequence.

In this paper, I report on the first step in this research, namely the architecture of a hybrid NN able to learn a sequence of bases and to predict it. The next step, reported in a companion paper, is to train and test the NNs to the identification of known and unknown sequences and to flag them correspondingly.

The suggested structure for the hybrid, hierarchical Fuzzy-NN to perform this task is shown in Fig. 2. The decision block is itself a (crisp) NN. Its role is to best determine the situations when a recognition score is high enough to decide the sequence has been recognized. This NN should be trained on a large number of sequences, to insure that the sequences not of the desired type are not falsely recognized as "good" sequences. For reusability reasons, it is conceivable to use an identical decision block for several FNNs-based agents.

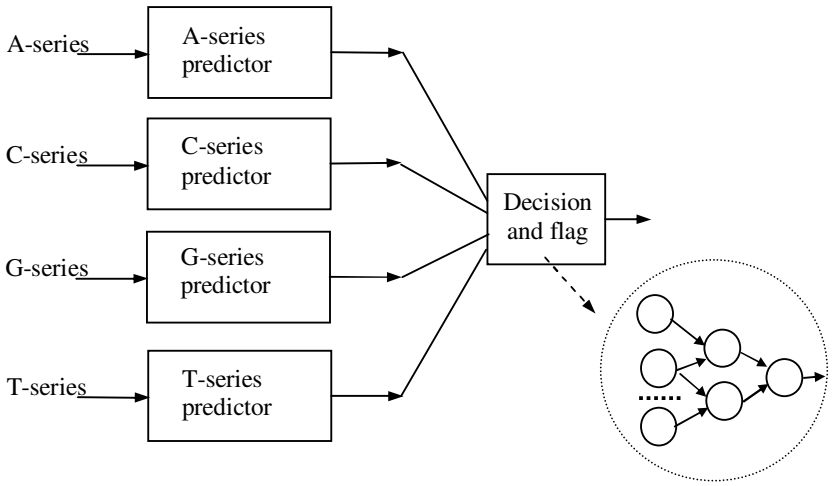

Fig. 2 – The multi-fuzzy-predictor and decision making block for pattern recognition in genomic sequences. The individual predictors may be of any type, but FNN predictors have several advantages

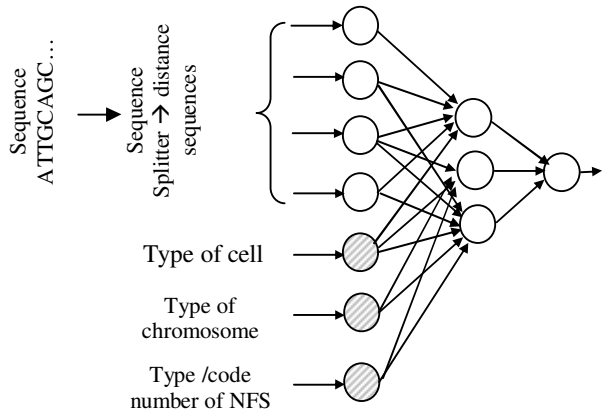

Fig. 3 – The overall structure of the genomic sequence identification agent

To achieve this, the decision-making NN should receive extra inputs, one of them for the type of sequence recognized, or, equivalently, for the type of FNN to which it applies. A code can be assigned to each FNN, to be used as input to the decision-making NN (DM-NN). Moreover, the DM-NN might receive input information on the organism and chromosome under investigation. Taking into account these requirements, the structure of the suggested DM-NN is like in Figure 3. Because of the extra information, the overall FNN-based agent has the structure of a hierarchical NN-based system.

The fuzzy-NN has been adopted from the literature and I developed the basic gradient-descent training algorithm as an example for my class in a master-degree course several years ago. With a few improvements, the algorithm reflects the classic ones used in NNs.

A convenient FNN is composed of several Sugeno fuzzy logic systems with derivable input RBF membership functions. Sugeno-type fuzzy systems are simple enough and do not require complex computations, like the Mamdani-type systems. The Gauss functions are preferred for the RBF membership functions.

The sequence identification procedure using FNNs based system is shown in Figure 4.

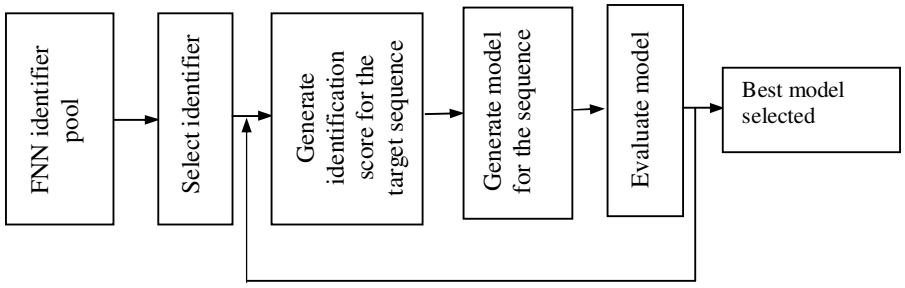

Fig. 4 – Using FNN-based agents to identify genomic sequences

Notice that there is a simple relationship relating the time series of distances (better-said, distance series). This means that the four series (one per basis) are not independent. Indeed, if we denote by $d\ [n]$ the distance series for the A, C, G, and T bases, respectively, then, for any *n*, we have:

$$\sum\ d_A[n_A]+d_C[n_C]+d_G[n_G]+d_T[n_T]=D$$

where *D* is the constant representing the sum. Therefore, knowing the maximal numbers n_A, n_C, n_G, n_T corresponding to a given *D*, any three series determine the fourth one. In other words, filling in the positions for three of the four bases, the

unoccupied positions are by rule occupied by the fourth basis. This information may help making corrections to the prediction algorithm.

Notice that the use of NNs is some kind of "indirect homology" that is not sensitive to too-close resemblance or sensitive to frame shift errors. On the other hand, the method will operate poorer on short exons. The NN method is close in effect with HMMs, as both carry statistical information on the global sequence and use information in the vicinity of the current point (several steps around or behind.)

The FNN method may be used in conjunction with other methods, either as the basic recognition method, or as a complementary method. Notice that the FNN method proposed differs significantly from the NN methods presented in the literature [5] in several respects, beyond the use of fuzzy NNs.

4. Conclusions

In this paper, a proposal for a research program in genomics-oriented bioinformatics has been presented. This program might be suitable as a basis for establishing a larger, regional or European program, to prepare for the next stage to come in genomics applications. Subsequently, a section of the research has been detailed and a tool to identify genomic sequences has been introduced. The tool is believed suitable for mass analysis of genomic sequences. The proposed genomic identification system is based on a set of fuzzy neural network predictors and a decision-making neural network. Further details on the proposed system are presented in a companion paper in this issue.

References

[1] COMPAQ INC.: http://www.sanger.ac.uk/Info/Events/Compaq/Compaq/sld016.htm

[2] DOE-GENOMES, US: *Human Genome Project Information*, http://www.ornl.gov/TechResources /Human_Genome/research/informatics.html

[3] NATO Newsletter nr. 62, March 2003

[4] H.N. TEODORESCU: *Self-organizing Uncertainty-based Networks*. Pp. 131-160. In: P. Melo-Pinto, H.N. Teodorescu and T. Fukuda (Editors), Systematic Organization of Information in Fuzzy Systems. Volume 184 NATO Science Series III: Computer & Systems Sciences. 2003

[5] E. UBERBACHER: *Computing the Genome*, http://www.ornl.gov/ORNLReview/v30n3-4/genome.htm

[6] A. KANDEL, A. KLEIN: *Fuzzy data mining.* Chapter 5, pp. 131-151, in H.N. Teodorescu, D. Mlynek, A. Kandel, H.J. Zimmermann (Eds.): Intelligent Systems and Interfaces. ISBN: 079237763X, Kluwer Academic Press, Boston. 2000

[7] GENES READING GROUP: *Minutes 2*. (Nov 13). www.pitt.edu/~kstotz/genes/minutes2.doc

[8] R. REYNOLDS, H. RESSOM, M. MUSAVI, C. DOMNISORU: *Improving Robustness of Fuzzy Gene Modeling*, ESANN'2002 Proceedings - European Symposium on Artificial Neural Networks Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 51-56

[9] A.P. GASCH, M.B. EISEN: *Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering*. Genome Biology 2002, 3(11): research0059.1–0059.22. http://rana.lbl.gov/papers/Gasch_GB_2002.pdf

[10] S. TOMIDA, T. HANAI, H. HONDA, T. KOBAYASHI: *Gene Expression Analysis Using Fuzzy ART.* Genome Informatics 12: 245–246 (2001) 245.

[11] http://www.jsbi.org/journal/GIW01/GIW01P003.pdf
[12] T.A. PASANEN, M. VIHINEN: *Formulating Gene Regulatory Patterns with Fuzzy Logic*, http://www.ki.se/icsb2002/pdf/ICSB_179.pdf
[13] H. DELALIN, J. LEGER, G. RAMSTEIN: *A Fuzzy Algorithm for Gene Expression Analysis.* http://www.lri.fr/~sebag/gafo/puces.pdf
[14] P.J. WOOLF, Y. WANG: *A Fuzzy Logic Approach to Analyzing Gene Expression Data.* Physiol Genomics, 3: 9–15, 2000. http://www.biostat.wisc.edu/geda/literature/woolf /woolf1.pdf
[15] R. GUTHKE, W. SCHMIDT-HECK, D. HAHN, M. PFAFF: *Gene Expression Data Mining for Functional Genomics using Fuzzy Technology.*
http://www.biochem.oulu.fi /BioStat/Guthke_Kluwer2002.pdf
[16] GASCH, A.P., B. MICHAEL: *Eisen{PRIVATE "TYPE=PICT;ALT="} Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering*. Genome Biol. 2002; 3 (11): research0059.1–research0059.22. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12429058
[17] B.A. SOKHANSANJ, G.R. RODRIGUE, J.P. FITCH: *Building and Testing Scalable Fuzzy Models of Bacterial Regulation: Analysis of Genomic Data for the Virulence Pathway of Yersinia pestis*. http://www.cr.org/publications/ICCN2002/pdf/264.pdf
[18] B. A. SOKHANSANJ, G. H. RODRIGUE, J. P. FITCH: *Applying URC Fuzzy Logic to Model Complex Biological Systems in the Language of Biologists.*
http://www.icsb2001.org/Posters/092_sokhansanj.pdf
[19] H. N. TEODORESCU, C. BONCIU: *Features-Oriented Filtering of Biological Signals*, Chapter 8, pp. 231-284, in H.N. Teodorescu, A. Kandel, and L.C. Jain (Eds.*):* Soft-Computing in Human-Related Sciences. ISBN: 0-8493-1635-9. CRC Press, Florida, May 1999
[20] H.N. TEODORESCU, C. BONCIU: *Features-Oriented Hybrid Neural Adaptive Systems and Applications*, Chapter 6, pp. 153-192, in H.N. Teodorescu, D. Mlynek, A. Kandel, H.J. Zimmermann (Eds.): Intelligent Systems and Interfaces. 480pp., ISBN: 079237763X, Kluwer Academic Press, Boston. 2000
[21] H.N. TEODORESCU, C. BONCIU: *Features Space Neural Filters and Controllers.* Chapter 5, pp. 105-147, in L.C. Jain and C.W. de Silva (Eds.), Intelligent Adaptive Control Industrial Applications (ISBN: 0-8493-9805-3), CRC Press, Fl, USA, 1998
[22] H.N. TEODORESCU, A. KANDEL, M. ANGHELESCU, C. BONCIU: *Feature-Oriented Filtering Hybrid Systems.* J. of Knowledge-Based Engineering Systems, Vol. 2, no. 2, April 1998, pp. 72-77
[23] H.N. TEODORESCU: *The Dynamics of the Words.* Invited Plenary Lecture, The 11th Conference on Applied and Industrial Mathematics (CAIM 2003): 29 - 31 May, 2003. University of Oradea, Romania, http://caim2003.rdsor.ro/
[24] H.N. TEODORESCU, A. KANDEL, F. GRIGORAS, D. MLYNEK: *Measuring with chaos: Sensorial systems and A-/t-ganglions.* Proc. Romanian Academy, Vol. 3, No. 1-2/2002, pp.55-62
[25] M. A. MARTÝ-RENOM, A. C. STUART, A. FISER, R. SANCHEZ, F. MELO, A. ALI: *Comparative Protein Structuremodeling of Genes and Genomes*. Annu. Rev. Biophys. Biomol. Struct. 2000. 29:291–325
[26] GENEMARK: http://opal.biology.gatech.edu/GeneMark/
[27] HMMGENE (v. 1.1), http://www.cbs.dtu.dk/services/HMMgene/
[28] R. HUGHEYGIF, A. KROGHGIF: *Hidden Markov models for sequence analysis: extension and analysis of the basic method*. CABIOS 12(2):95-107, 1996. Reprint: http://www.cse.ucsc.edu/research/compbio/html_format_papers/hughkrogh96/cabios.html
[29] R. KARCHIN: *Hidden Markov Models and Protein Sequence Analysis.* http://www.cse.ucsc.edu/research/compbio/ismb99.handouts/KK185FP.html