DNA Sequence Pattern Identification using A Combination of Neuro-Fuzzy Predictors

Horia-Nicolai Teodorescu [1, 2], and Lucian Iulian Fira [1]

[1] Faculty of Electronics and Communications, Technical University of Iasi,
B-dul Carol I, no. 11, Iasi, Romania
{hteodor, lfira}@etc.tuiasi.ro
[2] Romanian Academy, B-dul Carol I, no. 7, Iasi, Romania

Abstract. We address the prediction of the gene structure using a new method and tools, involving the sequence of distances between bases and neuro-fuzzy predictors. The method is tested on the HIV virus genome and the results look promising compared to other methods. We suggest that new, global prediction methods based on implicit, not explicit knowledge, may be as strong as the current, largely explicit knowledge based prediction methods.

1 Introduction

DNA analysis has seen last years a tremendous development. DNA includes huge amounts of data that must be interpreted. The operations to be performed include the recognition of the type of organism to which the gene belongs – when the genetic material comes from different sources, the identification of specific segments in the DNA sequence – sections that represent the genes – and the identification of the proteins the genes code – the prediction of the genes expression. Because of the huge amount of data in the genetic material, the operations must be automated. The first two tasks above are somewhat similar and may use similar tools. They are both essentially related to the identification of patterns in the genetic code. Several tools have been developed for these purposes, including hidden Markov models (HMM) [1], statistical methods, and fuzzy [2-4] and neural models that inherently reflect the statistics. Here, continuing [5-9], we present results using a novel approach in dealing with the genetic sequence prediction, based on its decomposition into four "distance series" and on the use of neuro-fuzzy predictors in a hierarchical pattern identification system.

The syntagma "gene prediction" has several meanings, relating to the object of prediction and depending on the research context. One meaning is to predict the splice sites [10]; another one is to determine what the gene expression result would be (the synthesized proteins).

2 The Pattern Detector

We approach the prediction task from a fresh point of view and propose a new type of gene prediction with a view to classify the genetic sequences, to determine the right splice sites and to classify the information they carry. For this purpose, we preprocess the original, raw base sequence and first produce four base sequences, for the four bases, A, C, G, and T respectively (Fig. 1).

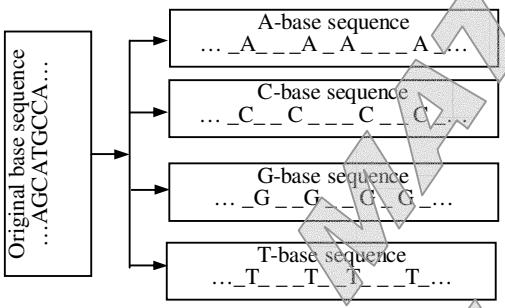

Fig. 1. Processing the base sequence

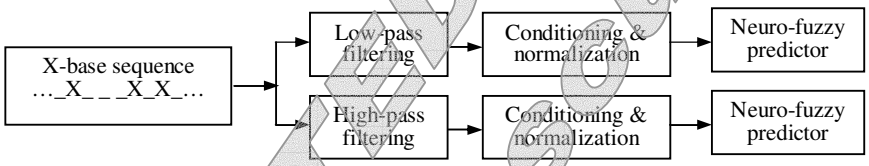

Fig. 2. Details of the processing of the individual sequences

Each of the four series is then preprocessed to derive the "trend" and the "fast varying" components. The low-pass filtering is performed by a 3-step MA filter, according to the formula:

$$y[n] = (x[n-1] + x[n] + x[n+1])/3. \quad (1)$$

Each of the two components is then predicted and the results are summed to generate the final prediction. Because the neuro-fuzzy system used in the prediction needs normalized values (values in the [-1, 1] interval), the distance series are normalized according to

$$y[n] = 2 \cdot [x[n] - (x_{max} + x_{min})/2]/(x_{max} - x_{min}). \quad (2)$$

where x_{max}, x_{min} are the maximal and minimal elements in the series.

To improve the prediction outcome, exceptional events (large distances, far out of the spreading range), if any, may first be eliminated. The sketch of the processing of the individual base sequence prediction is shown in Fig. 2.

We have used a neuro-fuzzy predictor, which has been developed and tested in our group in previous years. The architecture for one step predictive system is explained

below (see [2]). The predictor is a multi-fuzzy system network with inputs represented by the delayed samples, and the fuzzy cells are Sugeno type 0, with Gauss input membership functions. The formula $\mu(x) = exp\left(-(x-a)^2/\sigma\right)$ represents the membership degree μ of the input x; a is the center and σ is the spreading of the input membership function. Equation (3) stands for the input-output function of the neuro fuzzy predictor.

$$Y = \sum_{k=0}^{M} w_k \cdot \left[\sum_{l=1}^{N} \beta_{kl} \cdot e^{-\frac{(x_{n-k}-a_{kl})^2}{\sigma}} \Big/ \sum_{l=1}^{N} e^{-\frac{(x_{n-k}-a_{kl})^2}{\sigma}} \right] \quad (3)$$

We denote by M the number of Sugeno fuzzy systems, N the number of membership functions for each Sugeno fuzzy system, index $k = 0 \div M$, index $l = 1 \div N$, a_{kl} – the centers of the Gauss type membership functions, β_{kl} – the singletones and w_k – the weights.

Notice that the individual predictors may be of any type, but FNN predictors have several advantages. The rational for using this type of predictor is based on the complexity of the input-output function of this predictor, on its large number of parameters, and on the possibility to use various algorithms to train it, as explained below.

The class of predictor is given by the input-output function of the predicting system. This function is a ratio with sums of exponentials at the nominator and the denominator. Therefore, the capabilities of this type of predictor are higher than for simple fuzzy logic systems with triangular or other piecewise input and output membership functions. Also, the characteristic function of this predictor is more intricate than the sum of sigmoidal functions, as in the case of a single layer perceptron. Compared to a MLP using sigmoidal neurons, which has the characteristic function represented essentially by composed sigmoidal functions, still the characteristic function of this predictor is still more intricate, because it is a ratio of nonlinear functions. Compared to a MLP using Gaussian RBF neurons, which has the characteristic function represented essentially by composed Gaussian RBF functions, the characteristic function of this predictor is still more intricate, for similar reasons as above.

Regarding the number of parameters involved, showing the adaptation flexibility of the predictor, this predictor is similar to the ones discussed above. Moreover, by adding belief degrees to the rules of the TSK-fuzzy systems representing the neurons, the number of parameters is easily increased.

Regarding the possibility to use various learning algorithms, including gradient-type algorithms, gradient algorithms are easily adaptable to this predictor, in contrast to classic fuzzy systems with piecewise input and output membership functions, which do not accept classic gradient algorithms (because of the non-derivability of the function).

The training of the predictors refer to the adaptation of several sets of parameters, namely of the weights w_i, output singletons, and parameters of the input membership functions, a_k, σ_k. Two basic versions of the training algorithm can be used. According to the first, separate loops are used to adapt each of the above mentioned sets of parameters. We will name this type of algorithm "internal loops algorithm." The second algorithm uses a single loop to adapt all the parameters.

The overall multi-predictor system includes four individual predictors and a decision block (see [9]). This block fuses the results provided by the individual predictors to generate the recognition decision. Notice that, while current prediction methods use explicit knowledge on the gene sequences, system described here uses implicit information, in the sense that the information is included in the predictors, after appropriate training.

3 Results

In this section, we briefly present several results obtained by training the individual fuzzy predictors on the ENV gene (data source [11]), for the distance series corresponding to the A, C, G and T series. The summary of the results are shown in Figures 3 and 4, including the train results on the ENV gene, as well as the test results, both on ENV and on other HIV genes, moreover on genes from other viruses. In these figures, MSE denotes the mean square error and NMSE the normalized MSE.

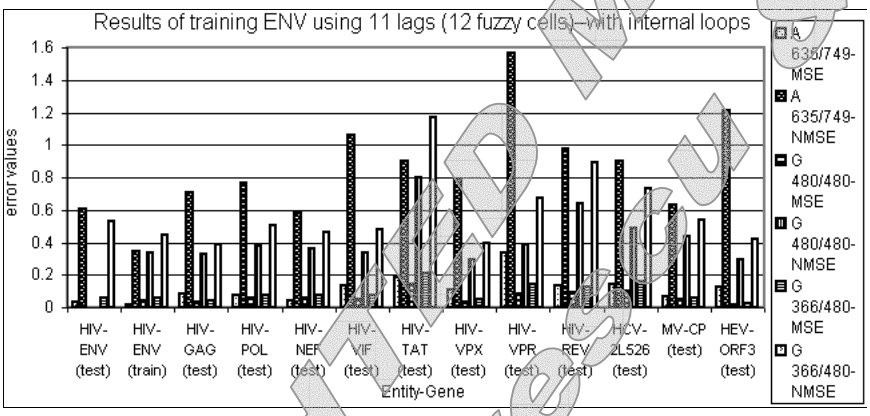

Fig. 3. Results obtained by training on the ENV sequence, using an algorithm with inner (partial) loops. The test period includes samples with indexes 636 to 749 for A basis, respectively 367-480 for G basis. The train period: samples #1 to 635 (A basis), respectively 1-480 for G, 1-366 for G basis

The results obtained by the training with either algorithm indicate that most genes are identified by the predictor for the A-base series. In the Fig. 3, this predictor confuses the ENV with the NEF gene when only the NMSE is used, but behaves correctly for the MSE (MSE for NEF > MSE for ENV–test.) Also, there is an almost miss for the CP gene of the Mosaic virus (NMSE 0.634, vs. 0.609 for ENV–HIV, but MSE-CP=0.071 vs. MSE-ENV=0.034.)

All the other genes are correctly rejected by the NMSE values alone for the A-series predictor, and all are correctly rejected by a combination of MSE and NMSE of this predictor. Excellent results are also obtained with the C-series predictor, with the exception, again, for the CP gene of the Mosaic virus. Poorer results are obtained with

the predictors for the T-series, which, however, correctly rejects the CP gene of the Mosaic virus. These results show that a combination of predictors is able to differentiate and even identify the genes.

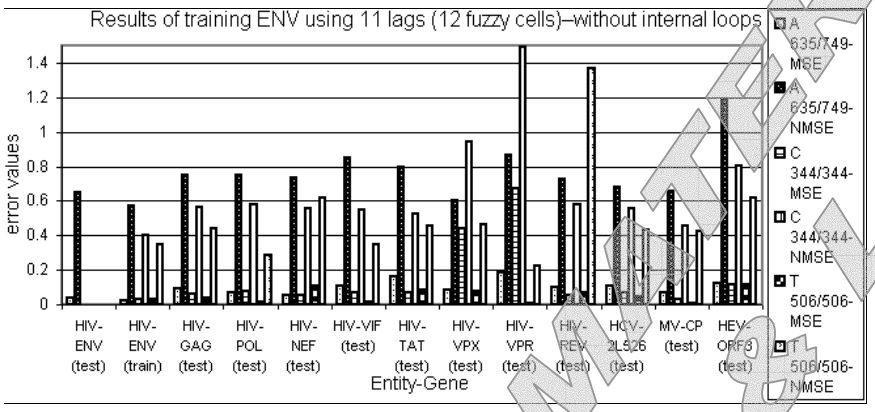

Fig. 4. Results obtained by training on the ENV sequence, using an algorithm without inner loops. The test period includes samples with indexes between 636 and 749 for A basis. The train period includes samples with indexes between 1 and 635 for A basis, respectively 1-344 for C basis, 1-506 for T basis

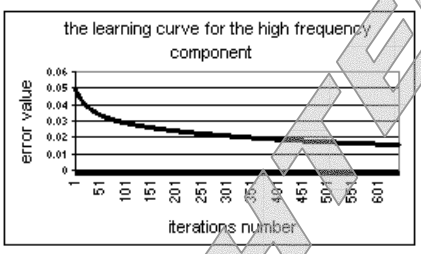

Fig. 5. Error evolution during training on the A series (ENV gene)

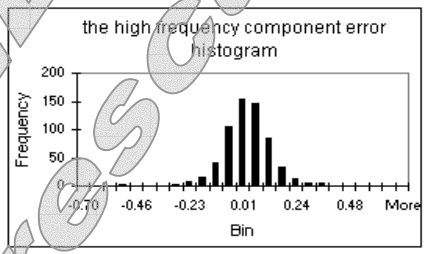

Fig. 6. Error histogram for the training results on the ENV gene, A basis, for the high-frequency component

The error evolution during training shows a correct learning process with steadily decreasing error (see example in Fig. 5). The error histogram show almost-Gauss distributions, meaning that the information in the series has been well extracted (see an example in Fig. 6).

4 Comments and Conclusions

The fact that several base sequences yield similar results in prediction, when a predictor trained on a specific sequence is used, may have several interpretations. The first

and easiest interpretation is that the predictor has not been trained well enough. We have to reject this hypothesis, because the training leads to small enough normalized errors, while the error values histogram shows a Gauss-like distribution.

The Gauss-like histogram is not, of course, an indication that the signal (remaining error) is pure white noise, but it is an indication that it might be and a necessary condition for being white noise. Moreover, it is an indication that there is no strong global correlation in the remaining signal. A Gauss histogram, however, is no guarantee that the signal is a noise. Therefore, we need other tools to determine the degree of information extraction from the signal. Such a tool is the self-correlation function.

The second possible interpretation might be that the series carry similar information and represent the almost same generation process. If true, then some of the base sequences from different genes might belong to the same class and then, we could determine classes of genes that, according to the prediction criterion are similar in the same class and dissimilar in different classes. The classes would be information-specific, while members of the same class still may look quite different. This finding may shed new light on the genetic processes and may have important consequences both in biology and bio-informatics.

References

1. Genie: Gene Finder Based on Generalized Hidden Markov Models. www.fruitfly.org/seq_tools/genie.html
2. Gasch, A.P., Eisen, M.B.: 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
3. Guthke, R., Schmidt-Heck, W., Hahn, D., Pfaff, M.: Gene Expression Data Mining for Functional Genomics using Fuzzy Technology. www.biochem.oulu.fi/BioStat/Guthke_Kluwer2002.pdf
4. Pasanen, T.A., Vihinen, M.: Formulating Gene Regulatory Patterns with Fuzzy Logic, http://www.ki.se/icsb2002/pdf/ICSB_179.pdf
5. Fira, L.I., Teodorescu, H.N.: Genome Bases Sequences Characterization by a Neuro-Fuzzy Predictor, Proc. IEEE-EMBS 2003 Conference, Cancun, Mexico, 3555-3558
6. Teodorescu, H.N.: The Dynamics of the Words. Invited Plenary Lecture, 11th Conf. Applied and Industrial Mathematics, 29-31 May, 2003. University of Oradea, Romania, http://caim2003.rdsor.ro/
7. Teodorescu, H.N., Fira, L.I.: Predicting the Genome Bases Sequences by Means of Distance Sequences and a Neuro-Fuzzy Predictor. *F.S.A.I.*, Vol. 9, Nos. 1–3, (2003), 23-33
8. Teodorescu, H.N.: Genetics, Gene Prediction, and Neuro-Fuzzy Systems–The Context and a Program Proposal. *F.S.A.I.*, Vol. 9, Nos. 1–3, (2003), 15–22
9. Teodorescu, H.N., Fira, L.I.: A Hybrid Data-Mining Approach in Genomics and Text Structures, Proc. The Third IEEE International Conference on Data Mining ICDM '03, Melbourne, Florida, USA, November 19 - 22, (2003), pp. 649-652
10. Thanaraj, T.A.: A Clean Data Set of EST-Confirmed Splice Sites from Homo Sapiens and Standards for Clean-up Procedures. Nucleic Acids Research 1999, vol. 27, no. 13, 2627-2637
11. Los Alamos National Laboratory: http://hiv-web.lanl.gov/content/hiv-db/align_current/align-index.html