International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:2, No:7, 2008

Neural Network Based Determination of Splice Junctions by ROC Analysis

S. Makal, L. Ozyilmaz, S. Palavaroglu

***Abstract*—**Gene, principal unit of inheritance, is an ordered sequence of nucleotides. The genes of eukaryotic organisms include alternating segments of exons and introns. The region of Deoxyribonucleic acid (DNA) within a gene containing instructions for coding a protein is called exon. On the other hand, non-coding regions called introns are another part of DNA that regulates gene expression by removing from the messenger Ribonucleic acid (RNA) in a splicing process. This paper proposes to determine splice junctions that are exon-intron boundaries by analyzing DNA sequences. A splice junction can be either exon-intron (EI) or intron exon (IE). Because of the popularity and compatibility of the artificial neural network (ANN) in genetic fields; various ANN models are applied in this research. Multi-layer Perceptron (MLP), Radial Basis Function (RBF) and Generalized Regression Neural Networks (GRNN) are used to analyze and detect the splice junctions of gene sequences. 10-fold cross validation is used to demonstrate the accuracy of networks. The real performances of these networks are found by applying Receiver Operating Characteristic (ROC) analysis.

***Keywords*—**Gene, neural networks, ROC analysis, splice junctions.

I. Introduction

A gene determines one inherited feature of an organism. The set of genes interacts to direct physical development and behavior of an organism. Most genes encode proteins, but some are transcribed into non-coding RNA molecules that function in protein biosynthesis and gene regulation. In all organisms, there are two major steps of producing a functional molecule of RNA or protein called gene expression. The first step is transcription that produces a single-stranded RNA molecule known as messenger RNA (mRNA) whose nucleotide sequence is complementary to the DNA from which it was transcribed. The second step is translation in which a mature mRNA molecule is used as a template for synthesizing a new protein [1].

An exon is a nucleotide sequence that is expressed or translated into protein, whereas an intron is an intervening sequence that is transcribed (into RNA) but later eliminated from the transcription by splicing its adjacent exons. Therefore, only exons represent the mature gene. The splice junctions refer to the points in which splicing takes place that connect exon and intron regions. The DNA sequence is an ordered structure, so a splice junction can be either exon-intron (EI) or intron-exon (IE) [2-5] (See Fig. 1).

Determination of exon and intron regions is crucial in diagnosis of genetic diseases. Genetic information is generated through seperation of exons and introns, and rejoining of exon regions. This process is called splicing. To provide accurate splicing, splice junctions should be obtained.

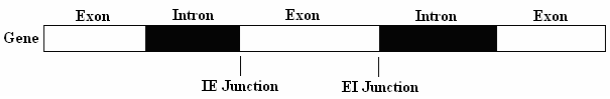

Fig. 1 Splice junctions of a gene

%15 of the mutations causing genetic diseases are originated from splicing mistakes. Most of these mutations are the changes of one nucleotid on the intronic and exonic regions of splice junctions [6].

Therefore, in a DNA sequence, the fundamental gene identification issue is to determine the presence and location of exons and introns in the sequence. Searching for special signal regions such as promoters (the initiation sites of transcription) or splice junctions is one approach. Measuring the splice characteristic of protein coding from segment to segment is another. In either case, exon identification is an essential step for gene modeling. A DNA sequence belongs to one of three classes [7, 8]according to the center of 60 nucleotides at the boundary between position 30-31: EI, an exon-intron boundary; IE, an intron-exon boundary; N, neither type of boundary (refer to Fig. 2). In this paper, two categories of IE and EI are chosen for ROC analysis. Thus, it is determined that in which category a DNA sequence belongs to.

II. Method

In this paper, ANN is chosen for determination of splice junctions because of its learning and generalization capabilities [9-11]. MLP, RBF and GRNN are applied to this work. The data set used in ANN application, is taken from Genbank [12], and it contains 1535 instances.

Manuscript received June 26, 2008.

S. M. is with the Electronics and Communications Engineering Department, Yildiz Technical University, 34349, Istanbul, Turkey (e-mail: smakal@yildiz.edu.tr).

L. O. is with the Electronics and Communications Engineering Department, Yildiz Technical University, 34349, Istanbul, Turkey (e-mail: ozyilmaz@yildiz.edu.tr).

S. P. is with the Electronics and Communications Engineering Department, Yildiz Technical University, 34349, Istanbul, Turkey (e-mail: spalavar@yahoo.co.uk).

Each instance consists of a sequence of 60 DNA nucleotides of four base types: A (Adenine), G (Guanine), T (Thymine), C (Cytosine) and a label indicated one of two possible classes: IE (an intron-exon boundary, called an acceptor), EI (an exon-intron boundary, called donor). There are 768 instances in the IE category, 767 instances in the IE category.

III. The Way of Improving Solution

Cross validation and ROC analysis are applied to find the real performances of the networks used to classify the dataset. Cross validation is used to find the generalization ability of ANN classification. ROC analysis has widely been used in medical data analysis to study the effect of varying the threshold on the numerical outcome of a diagnostic test.

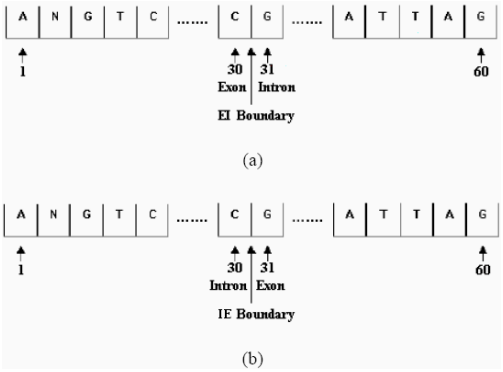

Fig. 2 Categories of a) EI; b) IE

A. Cross Validation

Cross validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. In this method, some of the data is removed before training begins. The initial subset of data is called the training set; while the other subset(s) are called testing sets. In the hold-out method that is the simple kind of cross validation, the data set is separated into two sets, called the training set and the testing set. In k-fold cross validation, the original sample is partitioned into k subset. A single subset is retained as the testing data and the remaining k − 1 subsets are used as training data. Then, the k results of testing and training can be averaged to produce a single estimation. Namely, in k-fold cross validation, the data set is divided into k subsets, and the holdout method is repeated k times [13]. In this work, k is 10.

B. Receiver Operating Chaarcteristic (ROC) Analysis

ROC Analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. It is originated from signal detection theory, as a model of how well a receiver is able to detect a signal in the presence of noise. There are four possible outcomes from a binary classifier. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP); however if the actual value is n, then it is said a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are n, and false negative (FN) is when the prediction outcome is n while the actual value is p [14].In this work, p and n is defined as EI and IE respectively shown in Table I. The limitations of diagnostic "accuracy" as a measure of decision performance require introduction of the concepts of the "sensitivity" and "specificity" of a diagnostic test. The equations of these measures can be given by (1) and (2) [15]:

$$Sensitivit\ y = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ negatives} \quad (1)$$

$$Specificit\ y = \frac{true\ negatives}{true\ \ negatives\ +\ false\ \ positives} \quad (2)$$

TABLE I
DECISION TABLE FOR EI-IE CLASSIFICATION

		Actual Value	
		EI	IE
Prediction	EI	True Positive	False Positive
	IE	False Negative	True Negative

The key feature of ROC analysis is the distinction between hit rate (or true positive rate) and false alarm rate (or false positive rate) as two separate performance measures.

IV. Results

MLP network has 60 input, one hidden layer of 6 hidden unit and one output layer of one unit. Each input DNA sequence for processing consists of 60 nucleotides. In the network, the nucleotide in each position is encoded by four input units designated by A, G, T and C. In hidden and output layers, “logaritmic sigmoid’’ (logsig) and “saturating linear transfer function’’ (satlin) are used. Data should be numeric for artificial neural network [16], so numerical values instead of A,T,G,C are applied to input layer of the network. MLP network was trained 50 epochs. While one subset is used for testing, nine of them is used for training. Each of the subsets is tested ten times in MLP and the average of them is calculated as success rate. GRNN and RBF have 60 inputs and an output layer of one unit. The spread value is chosen 0.5 for both of them. Classification accuracies of every subsets in training and testing processes and the average success rates of MLP, RBF and GRNN are shown in Table II.

TABLE II
RESULTS OF MLP, RBF AND GRNN

	MLP	RBF	GRNN
Testing (%)	91.23	89.35	91.14
Training (%)	98.88	100	91.99

There are applications using 10-fold cross validation for recognition splice junctions in literature. They are shown in Table III [12]. In this work, it has been shown that MLP, RBF and GRNN have a higher testing rate than previous works have.

TABLE III
PREVIOUS WORKS

Previous Works	Accuracies (%)
KBANN	83.37
Backprob	83.51
Pebls	84.27
Perceptron	67.27
ID3	75.43
COBWEB	75.5
Neigreast-Neigbour	79.26

The sensitivity and specificity values of MLP, RBF and GRNN are demonstrated in Table IV.

TABLE IV
THE SENSITIVITY AND SPECIFICITY VALUES OF MLP, RBF AND GRNN

		Sensitivity	Specificity
MLP	Testing	0.9	0.92
	Training	0.99	0.99
RBF	Testing	0.95	0.85
	Training	1	1
GRNN	Testing	0.93	0.89
	Training	0.94	0.9

According to obtained results of MLP, GRNN and RBF, training accuracies of all networks are higher than 90 %. Results of testing sets including data that is not used in training sets are more suitable for evaluation of network performances. Thus, all of obtained success rates by using test sets, ROC analysis of test sets and ROC curves drawn for test sets are considered for three networks. Success rates for test sets are acquired as 91.23 % for MLP, 91.14 % for GRNN and 89.35 % for RBF. So, MLP and GRNN are more superior in accuracies. However, ROC analysis results are taken into consideration for a more detailed review because these values alone are not expressive (they do not demonstrate network performances accurately.). Sensitivity values that give correctly detected EI junctions are found as 0.9 for MLP, 0.93 for GRNN and 0.95 for RBF. Specificity values which give correctly detected IE junctions are found as 0.92 for MLP, 0.89 for GRNN and 0.85 for RBF. MLP gives best results for specificity and worst results in sensitivity. On the other hand, RBF gives best results for sensitivity and worst results in specificity. According to these results, GRNN has the most correct result supplying network for both specificity and sensitivity values. In addition, according to ROC curves drawn for different decision borders used in test sets, GRNN supplies best result (Fig. 3). Thus, it is tellable that GRNN is more successful than other networks which are used in this work to determine EI_IE and IE_EI junctions. Determination of exon regions that supplies genetic information on gene is planned as a next step of this work by appliying GRNN.

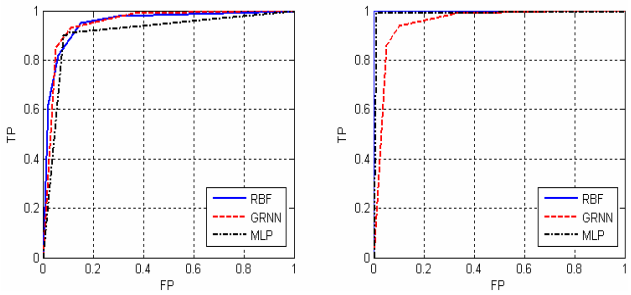

Fig. 3 ROC Comparison of neural networks for (a) testing results (b) training results

REFERENCES

[1] W.S. Klug, M.R. Cummings, Concepts of Genetics, Prentice Hall, 2000.
[2] S. Makal, L. Ozyilmaz, "Determination of splice junctions on DNA by neural Networks," International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, 2007, pp. 234–237.
[3] T. Naenna, R.A. Embrechts, "A modified Kohonen network for DNA splicejunction classification," IEEE Region 10 Conference, Chiang Mai, 2004, pp. 215–218.
[4] S. Mereuta, V. Munteanu, "A New Information Theoretic Approach to Exon-Intron Classification," International Symposium on Signals, Circuits and Systems, Lasi, 2007, Vol.2, pp. 1–4, 2007.
[5] M. Sarkar, T.Y. Leong, "Splice junction classification problems for DNA sequences," 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, İstanbul, 2001, pp. 2895–2898.
[6] L. Ozyilmaz, "Determination of exon and intron regions on DNA sequences by artificial neural Networks," Advances in Molecular Medicine International Journal of MolecularBiology, Biochemistry and Gene Technology, Istanbul, 2005, pp. 452–453.
[7] S. Rampone, "Splice-junction recognition on gene sequences (DNA) by BRAIN learning algorithm," IEEE World Congress on Computational Intelligence Neural Networks Proceedings , *Anchorage, 1998, Vol.1,* pp. 774–779.
[8] L. Fu, "An Expert Network For DNA Sequence Analysis," IEEE Intelligent Systems, *Vol.14, Issue 1,* pp. 65–71.
[9] Y. Xu, G. Helt, J.R. Einstein, G. Rubin, E.C. Uberbacher, "Drosophila GRAIL: an intelligent system for gene recognition in Drosophila DNA sequences," First International Symposium on Intelligence in Neural and Biological Systems, Herndon, 1995, pp. 128–135.
[10] J.J. Li, D.S. Huang, R.M. MacCallum, X.R. Wu, "Characterizing human gene splice sites using evolved regular expressions," IEEE International Joint Conference on Neural Networks, Montreal, 2005, pp. 493–498.
[11] T. Naenna, R.A. Bress, M.J. Embrechts, "DNA classifications with self-organizing maps (SOMs)," IEEE International Workshop on Soft Computing in Industrial Applications, Binghamton, 2003, pp. 151–154.
[12] Available: http://www.ics.uci.edu/~mlearn/databases/moleculer-biology/
[13] C.E. Vasios, G.K. Matsapoulos, E.M. Ventauras, K.S. Nikita, N. Uzunoglu, "Cross validation and neural network architecture selection for the classification of intracranial current sources," 7th Seminar on Neural Network Applications in Electrical, Serbia, 2004, pp. 151–158.
[14] T.J. Downey, D.J. Meyer, R.K. Price, E.L. Spitznagel, "Using the receiver operating characteristic to asses the performance of neural classifiers," International Joint Conference on Neural Networks, Washington, 1999, pp. 3642–3646.
[15] S. Wang, C.I. Chang, S.C. Yang, G.C. Hsu, H.H. Hsu, P.C. Chung, "3D ROC analysis for medical imaging diagnosis," IEEE Engineering in Medicine and Biology, Shanghai, 2008, pp. 7545–7548.
[16] C.H. Wu, J.W. MacLarty, Neural Networks and Genome Informatics, Elsevier Science Ltd., 2000.