Optimized Numerical Mapping Scheme for Filter-Based Exon Location in DNA Using a Quasi-Newton Algorithm

Parameswaran Ramachandran, Wu-Sheng Lu, and Andreas Antoniou
Department of Electrical and Computer Engineering
University of Victoria, BC, Canada, V8W 3P6
Email: rpara26@ieee.org, wslu@ece.uvic.ca, aantoniou@ieee.org

***Abstract*—An optimized numerical mapping scheme for achieving improved location of exons in DNA sequences using digital filters is proposed. Characteristic numerical values for the four nucleotides, referred to as pseudo-EIIP values, are obtained using a training procedure where the location accuracy is maximized using a quasi-Newton algorithm based on the Broyden-Fletcher-Goldfarb-Shanno updating formula. A training set of 80 DNA sequences is chosen from the HMR195 database. The objective function for the optimization procedure is formulated using the so-called receiver operating characteristic (ROC) technique and the procedure is initialized using electron-ion interaction potential (EIIP) values. Unbiased testing of the optimized characteristic values is carried out using a set of DNA sequences that has no overlap with the training set. Simulation results show that the pseudo-EIIP values yield more accurate exon locations than those obtained using the actual EIIP values.**

***Index Terms*—DNA, exons, period-3 property, electron-ion interaction potential, bandpass notch digital filters, BFGS updating formula, ROC plots.**

I. Introduction

DNA encodes the set of instructions to build and maintain a living organism [1]. It is composed of *nucleotides* that can be of four possible types, namely, adenine, thymine, guanine, and cytosine denoted by the letters A, T, G, and C, respectively. A DNA sequence can thus be represented by a string of characters. Specific regions in DNA known as *genes* contain the instructions for making proteins. The genes of higher organisms are split into coding regions called *exons* and noncoding regions called *introns*. Accurate location of exons is very important for understanding the functions of DNA and proteins and their interrelationships.

It turns out that the power spectra of DNA segments corresponding to exons exhibit a strong component at frequency $2\pi/3$, known as the *period-3 frequency*, while those corresponding to introns do not. The strength of the period-3 frequency component along the length of a DNA sequence can thus be used to distinguish between introns and exons and thereby locate the exons present in the DNA sequence [2]–[6]. Notable among the various approaches proposed in the past is an approach based on the use of digital filters owing to its high accuracy and computational efficiency [5], [6].

In addition to the choice of the filtering approach, the choice of the character-to-numerical mapping scheme has a critical influence on accuracy and computational efficiency. The electron-ion interaction potentials (EIIPs) used in [4] and [6], given in Table I, yield accuracies comparable to those obtained using binary sequences [2], [3], [5] and lead, in addition, to a significantly improved computational efficiency. This is due to the fact that the use of EIIP values involves the processing of only a single numerical sequence while the use of binary sequences involves the processing of four numerical sequences, one for each nucleotide. Thus if a set of alternative characteristic numerical values for the four nucleotides that would yield improved accuracy could be found to replace the EIIP values, then the tremendous computational efficiency associated with the processing of a single numerical sequence can be achieved while achieving improved accuracy.

In this paper, we propose an optimization-based training procedure for finding an alternative set of characteristic numerical values to replace the EIIP values. Simulations have shown that the numerical values obtained tend to yield improved accuracy in the exon-location process. Furthermore, they correlate well with the actual EIIP values and for this reason we refer to the new values as *pseudo-EIIP values*. The optimization is performed using a quasi-Newton algorithm based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) updating formula using the EIIP values as the initial point. The objective function is formulated using the so-called *receiver operating characteristic* (ROC) technique. To assure unbiased testing, two separate data sets from the HMR195 database are employed for the training and testing procedures.

The paper is organized as follows. Section II briefly describes exon location using digital filters, while Section III discusses the ROC technique. The optimization-based training procedure is described in Section IV and simulation results are presented in Section V.

II. Filter-Based Exon Location

Exons can be located by tracking the strength of the period-3 frequency component along the length of a DNA sequence. This can be effectively achieved through the following procedure: (1) The DNA sequence is converted into a numerical sequence of the form

$$x[n] = w_A x_A[n] + w_T x_T[n] + w_G x_G[n] + w_C x_C[n] \quad (1)$$

TABLE I
EIIP VALUES FOR THE NUCLEOTIDES

Nucleotide	EIIP
Adenine (A)	0.1260
Thymine (T)	0.1335
Guanine (G)	0.0806
Cytosine (C)	0.1340

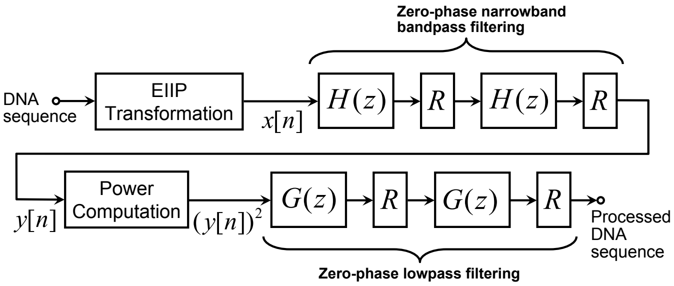

Fig. 1. Filter-based exon location system.

where w_A, w_T, w_G, and w_C are the four EIIP values and $x_A[n]$, $x_T[n]$, $x_G[n]$, and $x_C[n]$ are the corresponding binary sequences. (2) The numerical sequence $x[n]$ is filtered using a narrowband bandpass digital filter in order to select the period-3 frequency. (3) The output of the bandpass filter, $y[n]$, which turns out to be an amplitude-modulated signal, is demodulated by filtering its power $(y[n])^2$ using a lowpass filter. (4) The exon locations are identified as well-defined segments of $x[n]$ for which $(y[n])^2$ is equal to or exceeds a specified threshold. A block diagram of the system used is illustrated in Figure 1. Zero-phase filtering [7] is employed in order to avoid computing the phase response and to eliminate phase distortion. See [6] for further details.

III. ROC TECHNIQUE

The ROC technique is a tool for evaluating prediction techniques in terms of their performance [8]. It is based on evaluation metrics known as the *true positive rate* (TPR) and the *false positive rate* (FPR) defined by

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{and} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2}$$

respectively, where TP, TN, FP, and FN denote the number of *true positives*, *true negatives*, *false positives*, and *false negatives*, respectively, of the predicted exon locations relative to a set of known true locations. The TPR is plotted versus the FPR to obtain a point in the ROC plane as illustrated in Figure 2. Since the TPR and FPR can assume values in the range 0 to 1, the total area of the ROC plane is unity. The northwest pole, $(0, 1)$, represents perfect prediction, i.e., FPR $= 0$ and TPR $= 1$. The goal of any prediction technique is to reach this point. Informally, a point in the ROC plane is better than another if it is to the northwest of the latter. The diagonal line $y = x$ represents random predictions.

The construction of an ROC plot for the classification of predictions based on the above filter-based exon-location technique is carried out as follows: (1) For a given threshold value, the numerical values of $(y[n])^2$ are sorted into true positives, true negatives, false positives, and false negatives relative to a set of known true exon locations. (2) The numerical values of TP, TN, FP, and FN are obtained and, in turn, the metrics TPR and FPR are evaluated and used to plot a point in the ROC plane. (3) The preceding steps are repeated for different threshold values in the range of 0 to 1 and new points are plotted in the ROC plane to obtain an ROC curve as illustrated in Figure 2. The area under this curve (AUC) is a good indicator of the overall performance of an exon-location technique. The greater the AUC, the better would be the performance. Thus, for a given range on the x axis, the AUCs corresponding to two different exon location techniques can be compared and their performance relative to each other can be evaluated. ROC curves for two techniques A and B are shown in Figure 2 where technique A is deemed to be better than technique B.

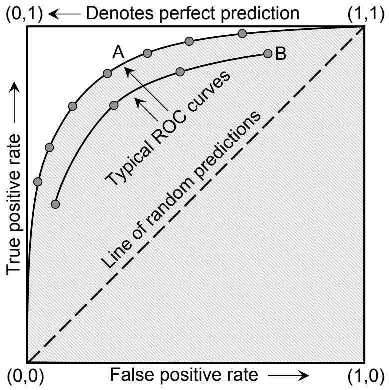

Fig. 2. The ROC plane illustrating typical ROC curves. The shaded region represents the area under curve A.

IV. PROPOSED TRAINING PROCEDURE

A. Optimization

A better set of numerical constants associated with the four nucleotides, designated as $\hat{w}_A$, $\hat{w}_T$, $\hat{w}_G$, and $\hat{w}_C$, can be obtained by maximizing the AUC corresponding to a training set of DNA sequences or, equivalently, by minimizing the quantity $1 - \text{AUC}$ since the total area of the ROC plane is unity.

A variety of algorithms can be used for the optimization problem under consideration such as algorithms of the quasi-Newton family which are both very efficient as well as robust [9]. A quasi-Newton algorithm based on the BFGS updating formula was found to give good results.

The objective function for the minimization involves several interdependent steps including bandpass and lowpass filtering of the numerical sequence, squaring the filtered output, and computing the AUC. Hence, deriving a closed-form expression for the objective function is not feasible. Instead, the optimization is carried out by numerically evaluating the objective function and the gradient in each iteration.

For the sake of consistency between the optimized numerical constants and the EIIP values, we need to ensure that (1) the four variables are always positive and (2) their numerical values are normalized at the end of each iteration such that their sum is always equal to the sum of the EIIP values. Positive numerical values can be easily achieved by replacing each variable by its square in the objective function. The normalization can be achieved by using the scaling factor

$$\mu = \sqrt{\frac{0.4741}{\hat{w}_A^2 + \hat{w}_T^2 + \hat{w}_G^2 + \hat{w}_C^2}} \tag{3}$$

where the constant 0.4741 is the sum of the four EIIP values. On the basis of extensive simulation results, the above adjustments in the variables do not seem to impede our ability to obtain optimized characteristic values that yield improved exon-location predictions.

B. Model for ROC Curves

ROC curves are inherently not continuous due to the fact that the number of thresholds used is finite. This poses a problem for the optimization procedure because the objective function to be minimized, $1 - \text{AUC}$, would not be continuous. To overcome this problem, the ROC curve can be approximated using an exponential model of the form

$$y = \alpha \left(1 - e^{-\left[\beta_1 \sqrt{x} + \beta_2 x\right]}\right) \tag{4}$$

where α, β_1, and β_2 are appropriate constants. These parameters can be determined by minimizing the error function

$$\text{E}(\mathbf{p}) = \sum_{i=1}^{n} \left[\alpha \left(1 - e^{-\left[\beta_1 \sqrt{x_i} + \beta_2 x_i\right]}\right) - y_i\right]^2 \tag{5}$$

where $\mathbf{p} = [\alpha \;\; \beta_1 \;\; \beta_2]^T$ and $\{x_i, y_i\}$ denotes the n pairs of FPRs and TPRs forming the ROC curve that is being modeled. $\text{E}(\mathbf{p})$ can be minimized in a straightforward manner using a quasi-Newton algorithm incorporating the BFGS updating formula, as before. For this application, closed-form expressions can be used for the gradient. A sample ROC curve and the approximation obtained using the above approach are illustrated in Figure 3. Using a termination tolerance of 10^{-6}, the algorithm required 19 iterations to converge. The values obtained for the model parameters in this specific example are

$$\alpha = 0.9291, \quad \beta_1 = 1.0611, \quad \beta_2 = 2.8443$$

As can be seen from Figure 3, the model closely approximates the ROC curve and thus can be effectively used to compute the AUC.

C. Training Procedure

The training procedure adopted is as follows: (1) A set of DNA character sequences is chosen and the parameter vector $\mathbf{x}$ is initialized using the positive square roots of the known EIIP values. (2) The character sequences are converted into numerical sequences using the squares of the current values in $\mathbf{x}$. (3) The sequences obtained are arranged consecutively to form a single cumulative contiguous sequence. (4) The cumulative sequence is processed using the exon-location system in Figure 1 and the amplitudes of the processed signals are normalized with respect to the interval $[0, 1]$. (5) Using the ROC technique as detailed in Section III, a curve in the ROC plane is obtained. (6) The ROC curve is approximated using the exponential model of Eq. (4). (7) The objective function $1 - \text{AUC}$ is evaluated. (8) The gradient of the objective function is evaluated by perturbing each variable, one at a time, by 10^{-4}. (9) An approximation to the Newton direction is generated and the variables are normalized using Eq. (3) and are then updated. (10) The procedure is repeated from Step (4) until convergence is achieved. The squares of the optimized parameters are the pseudo-EIIP values.

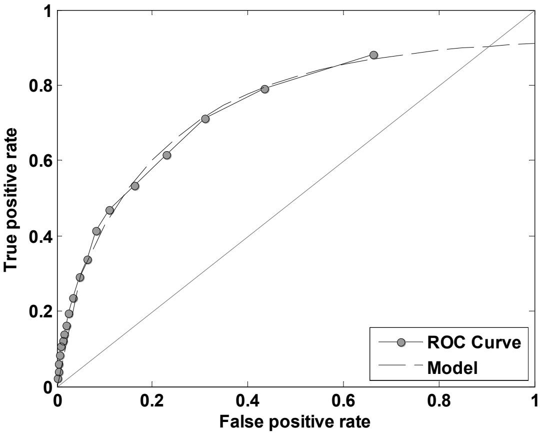

Fig. 3. An ROC curve and its exponential model.

The quasi-Newton algorithm used in the above training procedure was Algorithm 7.3 in [9].

V. Results

We now present simulation results obtained by training the numerical values using a specific data set and then testing the pseudo-EIIP values on another data set that has no overlap with the training set. This prevents the occurrence of any type of training bias in the test results. The data sets were chosen from the popular HMR195 database [10] since it provides the true exon locations for each sequence in addition to the sequence itself. Of the 195 sequences in the database, some were found to have *ambiguous nucleotides*, i.e., nucleotides whose identities have not as yet been experimentally validated. Avoiding such sequences, we selected 160 unambiguous sequences and divided them into a training and a test data set of 80 sequences each. The training procedure was carried out as described in Section IV. The DNA sequences were processed using a bandpass notch digital filter that was designed as described in [11]. The execution of the quasi-Newton algorithm was terminated when the 2-norm of the change in $\mathbf{x}_k$ and the change in the value of the objective function were both simultaneously less than a termination tolerance of 10^{-6}. The procedure took a total of 42 iterations for the main objective function. The average number of iterations taken for the exponential model was 20. The optimized nucleotide parameters are compared

TABLE II
INITIAL AND OPTIMIZED VALUES

Nucleotide	Initial (EIIP)	Pseudo-EIIP
Adenine (A)	0.1260	0.1994
Thymine (T)	0.1335	0.1933
Guanine (G)	0.0806	0.0123
Cytosine (C)	0.1340	0.0692

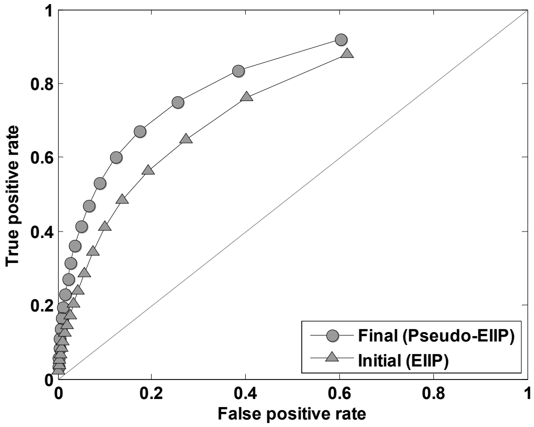

Fig. 4. ROC curves corresponding to the initial and the optimized values, obtained using the training set.

with the corresponding EIIP values in Table II. The initial and the final values of the objective function were 0.2661 and 0.1954, respectively.

The ROC curves obtained using the training set before and after the training procedure are illustrated in Figure 4. As can be seen, the pseudo-EIIP values yield a better measurement. Unbiased testing was then carried out on the test set and the resulting ROC curves obtained are shown in Figure 5. From this figure, it is clear that in addition to the expected improvements demonstrated on the training set, the pseudo-EIIP values also perform well on a set of DNA sequences that were not used for training. The overall accuracy, in terms of ROC curves, is significantly improved. The optimum operating threshold based on Figure 5 was found to be 0.15. It can be used for quick preliminary exon-location studies.

The results obtained from the simulations are encouraging since they strongly indicate that the nucleotide parameters can be optimized for improved accuracy. Consequently, it is possible to achieve the tremendous computational advantage associated with the processing of a single numerical sequence instead of four binary sequences. Extensive testing is currently underway in order to fully explore the benefits of the training procedure.

VI. CONCLUSION

A numerical mapping scheme that assigns optimized characteristic numerical values, referred to as pseudo-EIIP values, to the four nucleotides was proposed for use in filter-based exon location in DNA sequences. The scheme employs a training procedure executed using ROC plots. Unbiased testing was carried out and simulation results obtained indicate that the pseudo-EIIP values yield more accurate exon locations than those obtained using the actual EIIP values.

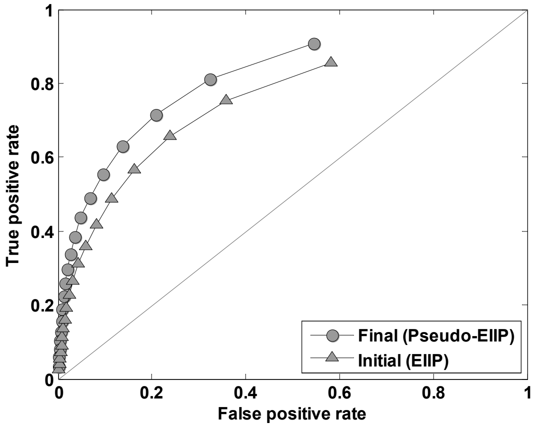

Fig. 5. ROC curves corresponding to the initial and the optimized values, obtained using a test set with no overlap with the training set.

The ROC technique is a powerful tool that can be employed for reliably evaluating the accuracies of prediction techniques using large data sets. Furthermore, as demonstrated here, it can be used as a training methodology for adjusting the parameters of a prediction system for optimum performance. In the future, we will investigate potential improvements to the training procedure as well as the optimization of prediction techniques for features in DNA other than exon location.

ACKNOWLEDGMENT

The authors are grateful to the Natural Sciences and Engineering Research Council of Canada for supporting this research.

REFERENCES

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology*. New York: Garland Publishing, 1998.

[2] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992.

[3] D. Anastassiou, "Genomic signal processing," *IEEE Signal Process. Mag.*, pp. 8–20, Jul. 2001.

[4] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation*, vol. 1, no. 6, pp. 197–202, 2006.

[5] P. P. Vaidyanathan and B.-J. Yoon, "Digital filters for gene prediction applications," in *Proc. 36th Asilomar Conference on Signals, Systems, and Computers*, Monterey, Nov. 2002, pp. 306–310.

[6] P. Ramachandran, W.-S. Lu, and A. Antoniou, "Location of exons in DNA sequences using digital filters," in *Proc. IEEE Int. Symp. Circuits Syst.*, Taipei, May 2009.

[7] A. Antoniou, *Digital Signal Processing: Signals, Systems, and Filters*. New York: McGraw-Hill, 2005.

[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[9] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. New York: Springer, 2007.

[10] S. Rogic, A. K. Mackworth, and F. B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome Research*, vol. 11, pp. 817–832, 2001.

[11] P. Ramachandran, W.-S. Lu, and A. Antoniou, "Improved hot-spot location technique for proteins using a bandpass notch digital filter," in *Proc. IEEE Int. Symp. Circuits Syst.*, Seattle, May 2008, pp. 2673–2676.