Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA

Parameswaran Ramachandran, *Member, IEEE,* Wu-Sheng Lu, *Fellow, IEEE,* and Andreas Antoniou, *Life Fellow, IEEE*

***Abstract*—The so-called *receiver operating characteristic technique* is used as a tool in an optimization procedure for the improvement and assessment of a filter-based methodology for the location of hot spots in protein sequences and exons in DNA sequences. By optimizing the characteristic values of the nucleotides, high efficiency as well as improved accuracy can be achieved relative to results obtained with the electron-ion interaction potentials. On the other hand, by using the proposed filter-based methodology with binary sequences, improved accuracy can be achieved although the efficiency is somewhat compromised relative to that achieved using the optimized characteristic values. Extensive experimental results, evaluated using measures such as the *g*-mean, the Matthews correlation coefficient, and the chi-square statistic, show that the filter-based methodology performs much better than existing techniques using the short-time discrete Fourier transform, particularly in applications where short exons are involved.**

***Index Terms*—Exons in DNA, hot spots in proteins, receiver operating characteristic technique, chi-square statistic, optimization, digital filtering, genomic signal processing.**

I. Introduction

HOT SPOTS and exons are regions in proteins and DNA sequences that play a crucial role in the functioning of proteins and DNA molecules, respectively. Hot spots contribute most of the binding energy for protein-target interactions while exons contain DNA code for making proteins. Locating hot spots and exons accurately is, as a consequence, an important fundamental problem the solution of which would lead to a better understanding of the functioning of proteins and DNA and the interdependencies between them [1], [2].

Several experimental techniques exist for locating hot spots and exons [3]. However, they need to be carried out in wet laboratories, which makes them both time-consuming and costly. Moreover, experimental data are available only for a limited number of protein complexes [4]. Therefore, to reduce costs, improve efficiency, and facilitate large-scale application, various computational hot-spot and exon location techniques have been developed in recent years. An early approach to the exon-location problem involves a general probabilistic model of the gene structure of human genomic sequences [5]. The model takes into account a variety of genetic markers and sequence patterns found along the genome in order to make biologically meaningful predictions. Subsequently, a number of improved prediction models as well as homology-based techniques have been developed [6], [7].

Computational hot-spot location techniques can be categorized as structure-based or sequence-based. *Structure-based* techniques make use of various types of complex information such as a protein's three-dimensional structure combined with statistical models to make predictions [8]–[15]. One of the first attempts at modeling the free energy of protein-protein interactions was made in [8] where an energy function involving solvation interactions and hydrogen bonding was used to predict hot-spot locations. The results obtained agree reasonably well with experimental data. In [9], non-covalent interactions were employed to estimate the energy contributions of residues that take part in binding. A support vector machine (SVM)-based prediction model was introduced in [10], which was further improved in [11] to achieve better accuracy by separately treating the predictions involving certain amino acids. SVMs have also been employed in [12] and [13] to develop feature-based hot-spot prediction techniques. In [14], the all-atom free-energy force field was used to locate hot spots in two specific protein complexes. Features such as conservation and solvent accessibility have also been employed for studying their effects on hot-spot detection [15]. A couple of Web servers, one involving a knowledge-based machine learning approach and the other involving an empirical model based on a set of simple physical properties, have been described in [16] and [17], respectively. A special class of structure-based techniques involve molecular dynamics (MD) simulations of the movements of atoms and molecules. MD simulations have been employed in [18] and [19] to study a set of protein complexes and determine their hot-spot locations. Although structure-based techniques yield good predictions, they are not suitable for large-scale applications due to their high computational cost and certain associated operational difficulties. Their applicability is further limited by the fact that the complex structural information they require is usually unavailable for newly-discovered proteins.

Sequence-based techniques, on the other hand, predict hot spots based solely on the linear amino-acid sequences, making them suitable for preliminary studies when the only type of data available are the linear sequences. Although such techniques can be quite useful, only a few attempts to develop them have been made in the literature. A notable example of this type of technique is the neural network based technique described in [20], which uses features such as sequence environment profiles, solvent accessibility, and evolutionary conservation.

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

The authors are with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6 Canada (e-mail: rpara26@gmail.com, wslu@ece.uvic.ca, aantoniou@ieee.org).

Recently, there has been a surging interest in the application of the principles of digital signal processing (DSP) to develop efficient sequence-based hot-spot and exon location techniques. Such techniques are easy to implement, entail reduced computational complexity, and yield fairly accurate hot-spot and exon predictions [21]–[31].

In DSP-based techniques, simple numerical mapping schemes in conjunction with well-established powerful DSP methods are employed. In [21], a DNA sequence is represented using four indicator binary sequences, one for each of the four types of nucleotides. The presence or absence of a nucleotide is indicated by the digit '1' or '0'. An alternative and more efficient mapping scheme based on certain characteristic values of the nucleotides known as *electron-ion interaction potentials* (EIIPs) involves a single numerical sequence obtained by representing the DNA or protein characters by their EIIP values [22], [24].

Binary sequences have been used to establish 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*, whereas segments corresponding to introns do not [21]. Exons can thus be located by tracking the strength of the period-3 frequency component along the length of a DNA sequence. In [23], this is achieved using the *short-time discrete Fourier transform* (STDFT) in conjunction with the rectangular window to compute a so-called *spectral content* (SC) measure of a binary sequence. An alternative, the so-called *optimized spectral content* (OSC) measure, was proposed in [26] where the frequency spectrum for each indicator sequence is multiplied by a complex coefficient to maximize the discriminatory capability between a coding region and synthetic DNA.

A measure based on the phase of the DFT, termed the *spectral rotation* (SR), was proposed in [28] where the complex coefficients of the OSC measure are replaced by constants derived from the phase distributions of the DFTs of a set of known sequences. The STDFT along with the Bartlett window was used in [29]. The independence of the period-3 property with respect to the reading frame was demonstrated in [32]. A multirate DSP analysis of the period-3 property was carried out in [32], [33].

EIIP sequences have been used to establish that the spectrum of each protein sequence for a given protein functional group exhibits a unique dominant frequency known as the *characteristic frequency* of the functional group [22]. This can be identified by computing the pointwise product of the DFTs of a sufficient number of protein EIIP sequences belonging to the functional group of interest. Such a product is known as the *consensus spectrum* [24]. Regions in a protein EIIP sequence where the characteristic frequency is dominant represent hot spots. Thus, like exons, hot spots can be located by tracking the strength of the characteristic frequency along the length of a protein sequence.

In our previous work, we have investigated the application of EIIP sequences in conjunction with the STDFT and also explored the use of different types of digital filters for hot-spot location [25], [31], [34]–[37]. In the present paper, we organize our filter-based techniques into a unified filter-based methodology that can be used for the location of hot spots in protein sequences and also for the location of exons in DNA sequences. Then we use the so-called *receiver operating characteristic* (ROC) technique as a tool in an optimization procedure for the improvement and assessment of our methodology. By using the proposed optimization procedure, we obtain optimized characteristic values for the nucleotides, we refer to as *pseudo-EIIP values*, which lead to a significant improvement in the accuracy of hot-spot as well as exon location. The paper concludes with extensive comparisons of our filter-based methodology with several other known DSP techniques for exon location. Evaluation metrics such as the g-mean, the Matthews correlation coefficient, and the chi-square statistic are computed for the comparisons. The results obtained show that our filter-based methodology is both more accurate as well as more efficient than STDFT-based methods particularly for the case of short exons.

The paper is organized as follows. Section II briefly describes our filter-based methodology. The ROC technique is described in Section III. An optimization procedure based on the ROC technique is developed and is then used to obtain optimized characteristic values for the nucleotides in Section IV. Extensive experimental results and comparisons are presented in Section V.

II. Filter-Based Location of Hot Spots and Exons

Hot spots in proteins and exons in DNA can be located by employing the filter-based systems illustrated in Fig. 1. The systems essentially convert the character sequences into numerical sequences using the EIIP values of the amino acids in the case of hot spot location or the EIIP or binary values of the nucleotides in the case of exon location; they then process them using narrowband bandpass filters centered at the characteristic frequency of the protein functional group of interest for the case of hot-spot location or at the period-3 frequency for the case of exon location. Plots of the power output would reveal the hot-spot or exon locations as regions of well-defined distinct peaks. Essentially, these systems are software implementations of spectrum analyzers. To eliminate phase distortion and the need of computing the phase response of the filters, *zero-phase filtering* is employed. Here, the signal is filtered through a cascade arrangement of two IIR filters characterized by $H(z)$ and $H(z^{-1})$. Transfer function $H(z^{-1})$ is realized using a filter characterized by a transfer function $H(z)$ sandwiched between two first-in last-out registers designated as R in Fig. 1. The frequency response of the cascade arrangement is real and, as a result, the system has a *zero* phase response. See Sec. 12.5 in [38] for further details about zero-phase filtering. Since DNA sequences are much longer than protein sequences, the bandpass filtered output of the exon location system turns out to be an amplitude-modulated signal. Hence it needs to be demodulated in order to identify the exon locations and this can be done by using a lowpass filter.

Although several types of narrowband bandpass digital filters can, in theory, be used in the systems of Fig. 1a and b, extensive experimental results (see Section V-A) have shown that narrowband bandpass notch (BPNs) filters are most suitable for this application due to their high selectivity,

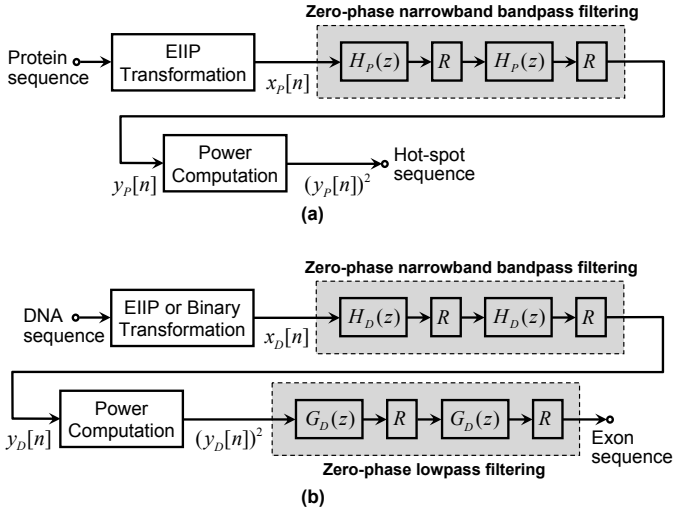

Fig. 1. Filter-based location systems for (a) hot spots and (b) exons.

reduced length of transience, and the low computational effort involved [36]. The type of lowpass filter used in Fig. 1b is less critical and any one of the standard filter types could be used such as an elliptic, Chebyshev, or inverse-Chebyshev filter.

A. Design of BPN Filter

A second-order BPN filter characterized by the transfer function

$$H_{BPN}(z) = \frac{1}{2}\left[\frac{(1-d_0)(1-z^{-2})}{1+d_1 z^{-1}+d_0 z^{-2}}\right]$$

can be designed by using an allpass filter in parallel with a bandstop notch filter. The design involves setting d_0 to $1-\tau$, where τ is a specified stability margin, and determining d_1 such that the area under the amplitude response curve is minimized. This can be readily achieved using a 1-dimensional optimization technique, such as, for example, the golden-section search [39].

B. Tuning of BPN Filter

The characteristic frequency is different for different functional groups and can be determined only as accurately as the frequency resolution of the consensus spectrum. This, in turn, is limited by the lengths of the protein sequences which are usually of the order of a couple of hundred amino acids. On the other hand, BPN filters are highly selective and to improve the performance of the hot-spot location system in Fig. 1a, the BPN filter must be tuned to ensure that the passband center frequency of the filter, ω_0, matches the characteristic frequency of the functional group.

The tuning procedure must be fast enough to facilitate the real-time observation of the results and, therefore, it must involve minimal computational effort. This can be achieved by modeling the variations of d_1 in response to small changes in the passband center frequency ω_0 using a least-squares polynomial model of the form

$$d_1(\omega_0) = d_{1U} + x_1(\omega_0 - \omega_{0U}) + x_2(\omega_0 - \omega_{0U})^2 \quad (1)$$

where d_{1U} and ω_{0U} are the values of d_1 and ω_0 in the untuned filter. Coefficients x_1 and x_2 are determined by solving an overdetermined system formed using the d_1 values of a set of known BPN filters. The values thus determined can be used to obtain d_1 for any ω_0 within the interval $\omega_l \leq \omega_0 \leq \omega_u$. For further details about the design of the BPN filter, the reader is referred to [36], [37]. A strategy for the automatic tuning of the filter is proposed below.

Note that the BPN filter need not be tuned in the case of exon location since the period-3 frequency is fixed and is accurately known. Also no tuning is necessary in the case where a bandpass inverse-Chebyshev filter is used since these filters have a flat passband over a relatively wider range of frequencies, typically, 1 to 2% of the center frequency (see Table III in [25]).

C. Automated Tuning for Hot-Spot Location

The BPN filter can be tuned by varying the center frequency from ω_l to ω_u in steps of, say, 0.001, and recording the output power for each center frequency at each of the known hot-spot locations. The sum of the recorded output power values is then computed for each center frequency and the frequency that yields the maximum value of the sum is taken to be the required center frequency for the filter.

If no hot-spot locations are known for the functional group of a given sequence of interest, then the required center frequency is taken to be the frequency that yields the maximum power output.

D. Use of Binary Sequences

Filter-based exon location can also be carried out using binary sequences. This can be done by processing each of the four binary sequences using a narrowband bandpass filter and then filtering the combined power output signal given by

$$y_B[n] = p_A\,|y_{BA}[n]|^2 + p_T\,|y_{BT}[n]|^2 + p_G\,|y_{BG}[n]|^2 + p_C\,|y_{BC}[n]|^2 \quad (2)$$

using a lowpass filter. The weights p_A, p_T, p_G, and p_C are usually assumed to have equal values, typically, 0.25.

E. Software Implementation

The above filter-based techniques for hot-spot and exon location were implemented in MATLAB employing user-friendly graphical interfaces. The software designs and tunes the required filters, reads the protein or DNA sequences, and identifies the locations of hot-spots or exons.[1] This would enable users of the proposed techniques to apply them to arbitrary protein and DNA sequences without the need to learn the considerable body of knowledge required for the design of digital filters.

[1]The reader is directed to http://www.ece.uvic.ca/~andreas/ for a copy of the software.

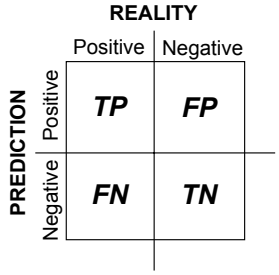

Fig. 2. Confusion matrix based on classifier outcomes.

III. Evaluation Metrics and the ROC Technique

In this section, we define several performance evaluation metrics and propose the ROC technique for use in the comparisons described in Section V. In what follows, we refer to the hot-spot and exon location techniques as *classifiers* because, by means of their predictions, they essentially classify a given location in a protein or a DNA sequence as a *positive* or *negative* instance. Thus in the *prediction* domain, a positive instance indicates that the location is predicted to be a hot spot or exon while a negative instance indicates the opposite. Similarly, in the *reality* domain representing *ground truths*, a positive instance indicates that the location truly identifies a hot spot or an exon while a negative instance indicates the opposite. The quality of the predictions can then be evaluated by comparing the number of positive and negative instances in the prediction domain with those in the reality domain. Four possibilities can arise as a result of such a comparison. They are as follows: A positive prediction is a *true positive* if it also corresponds to a positive instance in reality; otherwise, it is a *false positive*. A negative prediction is a *true negative* if it also corresponds to a negative instance in reality; otherwise, it is a *false negative*.

A. Classifier Measures

With a set of predictions obtained by running a classifier on a test data set, the number of true positives, *TP*, false positives, *FP*, true negatives, *TN*, and false negatives, *FN*, can be determined. A two-by-two matrix known as a *confusion matrix* can be constructed representing the relationship between reality and prediction, as shown in Fig. 2 [40], [41]. This matrix forms the basis for a variety of performance metrics defined as follows. *Sensitivity* (*Sn*), also known as the *true positive rate* (*TPR*), is the proportion of true positives that have been correctly predicted as positives. *Specificity* (*Sp*), also known as the *true negative rate* (*TNR*), is the proportion of true negatives that have been correctly predicted as negatives. *Sn* and *Sp* are given by

$$Sn = \frac{TP}{TP + FN} \quad \text{and} \quad Sp = \frac{TN}{TN + FP} \tag{3}$$

respectively. Due to the fact that hot spots and exons are sparsely distributed in proteins and DNA, quite often the number of negatives is much greater than the number of positives. Thus *TN* tends to be much larger than *FP* and hence *Sp*, as computed in (3), produces large uninformative values. For this reason, specificity has been defined in the gene-prediction literature as [40]

$$mSp = \frac{TP}{TP + FP}$$

and is also known as *precision* (*Prc*). It represents the proportion of predicted positives that are truly positive. *Accuracy* (*Acc*), which is the probability of correct prediction, and *false positive rate* (*FPR*) are defined by

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{and} \quad FPR = \frac{FP}{FP + TN} \tag{4}$$

respectively. Like *Sp*, *Acc* may not be an adequate measure when the number of negative cases is much greater than the number of positive cases [42]. For example, if 95 out of 100 cases are truly negative and 5 are truly positive and a classifier classifies all cases as negative, *Acc* would still be 95% although the classifier missed all positive cases. To circumvent this problem, we can define the *geometric mean* (g-mean) as [42]

$$g\text{-mean} = \sqrt{TPR \times TNR} = \sqrt{Sn \times Sp}$$

If all positive cases are classified incorrectly, the g-mean would take the value 0.

Metric g-mean has some distinct advantages. It is a meaningful measure of the overall classifier performance and takes a high value only when both *TPR* and *TNR* have close and high values. It is, in addition, robust to changes in the distributions of true positives and negatives. Furthermore, its nonlinear nature allows for a larger reduction in value as the values of *TPR* and *TNR* decrease. In other words, the penalty for misclassification is higher as the number of misclassifications increases.

Another standard classifier measure commonly used by statisticians is the Matthews correlation coefficient defined as [43]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{5}$$

In essence, *MCC* measures the correlation between the observed and predicted binary classifications and returns a value between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. Like g-mean, *MCC* uses all four numbers, namely, TP, TN, FP, and FN, and thus provides a balanced evaluation of the prediction.

B. Test of Statistical Significance

An interesting property of MCC is its direct association with the χ^2 (chi-square) distribution, which can be used to test whether the predictions are significantly more correlated with the true data than a random guess [43]. If the chi-square test is applied to a 2×2 contingency matrix formed using TP, TN, FP, and FN, then the test statistic is given by

$$\chi^2 = n \times MCC^2 \tag{6}$$

where n is the total number of observations.

For our analysis, we measure the overall performance of the techniques using g-mean and *MCC*, and test the statistical significance of the predictions using the chi-square test [44]. The χ^2 statistic can be used to obtain a p-value that denotes

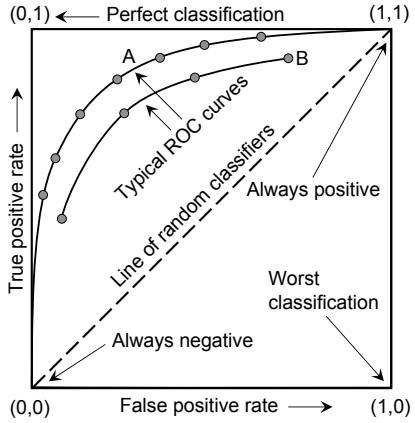

Fig. 3. Significant points in the ROC plane and typical ROC curves.

the probability that the predictions could have occurred by chance. The standard practice in statistics is to assume a level of significance of 0.05 and, consequently, the predictions are considered significant if $p \leq 0.05$. When a more stringent cutoff is required in certain applications, such as in the medical field, levels of significance of 0.01 or 0.001 are sometimes employed.

C. ROC Plots

The performance of classifiers can be evaluated by using ROC plots whereby the false positive rate is plotted versus the true positive rate as illustrated in Fig. 3 [41], [42]. A test data set classified using a given classifier thus corresponds to a point in the ROC plane. The northwest pole in the plane, $(0, 1)$, represents perfect classification with no false positives or false negatives, which would correspond to an ideal classifier. The goal of any classifier is to reach this point. The southwest pole, $(0, 0)$, represents the situation when the classifier predicts no positives, thus neither committing false positive errors nor predicting true positives. The opposite situation of unconditionally classifying all instances as positives is represented by the northeast pole, $(1, 1)$. Finally, the southeast pole, $(1, 0)$, represents the worst classification with no true positives or true negatives.

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 a random classifier. For example, a classifier randomly predicting positives half of the time would correspond to the point $(0.5, 0.5)$ and one predicting positives 90% of the time would correspond to the point $(0.9, 0.9)$. For a good classifier, its ROC point should stay as much above and to the left of the diagonal line as possible.

The classification process usually involves the use of a threshold parameter. If the system output is larger than the threshold value, then the instance is classified as positive, otherwise, it is classified as negative. Evidently, the true and false positive rates would vary with the threshold value. Thus, given a classifier and a data set, the threshold can be varied from some lower to some upper value and the resulting values of *TPR* and *FPR* can be plotted to obtain a curve in the ROC plane. By comparing curves obtained with different classifiers using the same data set, the best classifier can be determined. The optimum threshold for a given classifier would be the threshold corresponding to the point on its ROC curve closest to the northwest pole $(0, 1)$. ROC curves for two classifiers A and B are shown in Fig. 3 where classifier A is deemed to be better than classifier B.

D. Area Under an ROC Curve

ROC plots for use with the proposed filter-based methodology can be constructed as follows:

1) For a given threshold value, the predicted exon locations obtained from the demodulated output signal $(y_D[n])^2$ 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.

The area under an ROC curve (AUC) is a good indicator of the overall performance of the corresponding prediction 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.

E. Model for ROC Curves

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

$$y = \alpha \left(1 - e^{-(\beta_1 \sqrt{x} + \beta_2 x)}\right) \tag{7}$$

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

$$E(\boldsymbol{p}) = \sum_{i=1}^{n} \left[\alpha \left(1 - e^{-(\beta_1 \sqrt{x_i} + \beta_2 x_i)}\right) - y_i\right]^2 \tag{8}$$

where $\boldsymbol{p} = [\alpha \ \beta_1 \ \beta_2]^T$ and $\{x_i, y_i\}$ denote the n pairs of FPRs and TPRs used to construct the ROC curve that is being modeled.

Extensive experimentation using a variety of ROC curves has confirmed the validity of the model. A sample ROC curve and its approximation obtained using the above approach are illustrated in Fig. 4. For further details of the model, the reader is referred to [45].

F. Computational Efficiency and Overall Accuracy

The computational efficiency of various hot-spot location techniques has been compared in our previous work [25], [36], [46]. For exon-location techniques, the computational efficiency is evaluated as follows. For a chosen DNA sequence, the binary and EIIP sequences are each subjected to 1000 runs

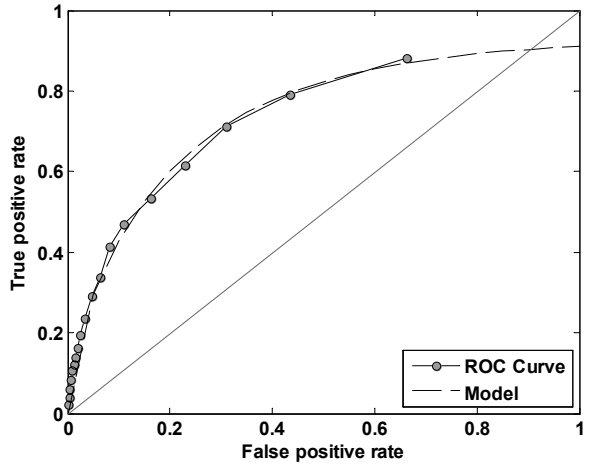

Fig. 4. An ROC curve and its exponential model.

and the required CPU times are averaged over the runs. This is referred to as the *average CPU time* (*ACT*). Its reciprocal, $1/ACT$, is a direct indicator of the computational efficiency.

An estimate of the *overall accuracy* of a technique can be obtained from its ROC curve by computing the Euclidean distance between point $(0, 1)$ and the point closest to it on the ROC curve. We refer to this measure as the *shortest Euclidean distance* (*SED*). *SED* is inversely proportional to the overall accuracy, and hence its reciprocal, $1/SED$, is a direct indicator of overall accuracy.

IV. Optimized Characteristic Values for Nucleotides

The ROC technique along with a suitable optimization procedure can be used to obtain a better set of characteristic values for the four nucleotides, as will be demonstrated below [47]. This can be achieved 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. A quasi-Newton algorithm based on the BFGS updating formula was found to give good results [39].

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.

In order to achieve consistency between the optimized characteristic values and the EIIP values, we need to ensure (1) that the four variables are always positive and (2) that 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

$$\mu = \sqrt{\frac{0.4741}{\hat{w}_A^2 + \hat{w}_T^2 + \hat{w}_G^2 + \hat{w}_C^2}} \tag{9}$$

where the constant 0.4741 is the sum of the four EIIP values and $\hat{w}_A$, $\hat{w}_T$, $\hat{w}_G$, and $\hat{w}_C$ are the optimized characteristic values for the four nucleotides at the end of each iteration.

On the basis of extensive experimental results, the above adjustments in the variables do not seem to impede our ability to obtain optimized numerical values that yield improved exon-location predictions.

In view of their consistency with the actual EIIP values, the optimized characteristic values will be referred to as the *pseudo-EIIP* values hereafter.

A. Step-By-Step Procedure

A step-by-step procedure for the optimization of the characteristic values of the nucleotides is as follows:

1) A set of DNA character sequences is chosen and the parameter vector $\boldsymbol{x}$ is initially set to the positive square roots of the EIIP values.
2) The character sequences are converted into numerical sequences using the squares of the current values in $\boldsymbol{x}$.
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 Fig. 1(b) and the amplitudes of the processed signal are normalized with respect to the interval $[0, 1]$.
5) An ROC curve is obtained using the procedure described in Section III-D and is approximated using the exponential model of Eq. (7).
6) The objective function $1 - \text{AUC}$ is evaluated using a numerical method. The gradient is then obtained by perturbing each variable, one at a time by, say, 10^{-4}.
7) An approximation to the Newton direction is generated and the variables are normalized using Eq. (9) and are then updated.
8) The procedure is repeated from Step 2 until convergence is achieved.
9) The squares of the values in $\boldsymbol{x}$ are the optimized parameters, referred to as the *pseudo-EIIP values*.

Step 2 and Step 9 ensure that the optimized nucleotide values are positive. In Step 7 an approximation to the Newton direction is generated using Algorithm 7.3 in [39] and the variables are normalized.

V. Results and Discussion

In this section, we present

- results pertaining to the choice of the type of narrowband filter for the hot-spot location system shown in Fig. 1a,
- results pertaining to the use of pseudo-EIIP values in the exon-location system of Fig. 1b, and
- a comparative analysis of our filter-based methodology with other known DSP-based techniques

as detailed below.

TABLE I
EVALUATION METRICS AT BEST OPERATING THRESHOLDS FOR HOT-SPOT LOCATION

Type	*TPR*	*FPR*	*Sp*	*Prc*	*Acc*	*g*-mean	*MCC*	χ^2
IC	0.51	0.44	0.56	0.03	0.56	0.53	0.02	8.95
tBPN	0.54	0.29	0.71	0.05	0.70	0.62	0.09	137.39

A. Bandpass Notch versus Inverse-Chebyshev Filters for Hot-Spot Location

The narrowband bandpass filter in the system of Fig. 1a should have a monotonic amplitude response with respect to the passband and two types of filters satisfy this requirement, namely, inverse-Chebyshev (IC) and BPN filters. Below we compare the performance of these two types of filters using the ROC technique.

The inverse-Chebyshev filter was designed to have a maximum passband and minimum stopband attenuation of 1 dB and 30 dB, respectively, and a passband bandwidth of 1 percent. The stability margin of the BPN filter was set to 0.03. These filters were designed using the software described in Section II-E.

Data sets comprising 66 protein sequences were used for hot-spot location. The protein sequences were obtained from the *protein data bank* (PDB) [48] and the ground truths for the hot-spot locations were obtained from the *alanine scanning energetics database* (ASEdb) [49]. If the ground truths of a protein were not available, those of the same protein from another organism were used. The set of protein sequences had a total of 482 confirmed hot spots and the mean sequence length was 277 amino acids.

The BPN filter was tuned using the automatic tuning procedure in Section II-C with 132 out of the 482 hot-spot locations in the data set. No tuning was necessary for the inverse-Chebyshev filter in view of its wider passband bandwidth.

The confusion matrix and evaluation metrics were computed cumulatively, i.e., the entire set of 66 protein sequences was regarded as a single long sequence and the values of *TP*, *TN*, *FP*, and *FN* corresponding to the individual sequences were summed to obtain a single set of four values, which was then used to compute the evaluation metrics. These computations were repeated for threshold values in the range 0.05 to 0.95 to obtain the required ROC curve. From the ROC curve shown in Fig. 5, the best operating threshold and corresponding shortest Euclidean distance from point $(0, 1)$ were found to be 0.1 and 0.66, respectively, for the inverse-Chebyshev filter and 0.25 and 0.54, respectively, for the BPN filter. The evaluation metrics as well as the χ^2 statistic for the optimum threshold are given in Table I. As can be seen, the use of a BPN filter leads to a significant improvement in our hot-spot location methodology. Both the *g*-mean and *MCC* values corresponding to the BPN filter are higher than those corresponding to the inverse-Chebyshev filter.

The precision and the *MCC* values for our filter-based methodology are moderately good at best. This is due to the large number of false positive instances that are detected based on the current set of ground truths available in the ASEdb. However, as new hot spots are added to the ASEdb by the biological community, at least some of our false positive instances will become true positive instances and, consequently, metric *FP* would decrease and *TP* would increase thereby resulting in higher precision and *MCC* values. Due to this reason, under the current circumstances, metric *g*-mean is a better indication of the overall performance of the hot-spot location techniques.

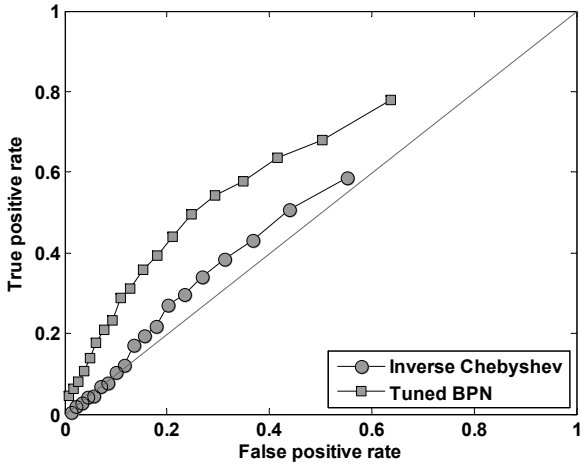

Fig. 5. ROC curves obtained with inverse-Chebyshev and BPN filters.

TABLE II
INITIAL AND OPTIMIZED NUMERICAL PARAMETERS

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

The statistical significance of the predictions was tested by obtaining *p*-values for the χ^2 statistic using the distribution table. The *p*-value obtained for the inverse-Chebyshev filter was 0.0028 while that obtained for the BPN filter was much smaller, less than 0.0001. As can be seen, both the values are far less than the standard threshold of 0.05, thereby confirming that the predictions are statistically significant.

B. Pseudo-EIIP Values for Exon Location

The pseudo-EIIP values were obtained using the procedure in Section IV-A. The execution of the quasi-Newton algorithm was terminated when the 2-norm of the change in $\boldsymbol{x}_k$ and the change in the value of the objective function were both simultaneously less than a termination tolerance of 10^{-6}.

The DNA sequences and the ground truths for the exon locations were obtained from the HMR195 data set [50]. This data set contains a sufficiently large number of DNA test sequences for three different organisms and provides the true exon locations in addition to the sequences themselves and it is, therefore, well-suited for our analysis. It has been used in several studies including recent ones such as that in [51].

We selected 160 out of 195 sequences avoiding sequences with nucleotides that have not as yet been validated experimentally. The selected subset had a total of 780 exons with a mean sequence length of 7182 nucleotides. Since the sequences have been organized as annotated coding regions beginning with the start codon and ending with one of the stop codons, there was no need to separately identify the open reading frames.

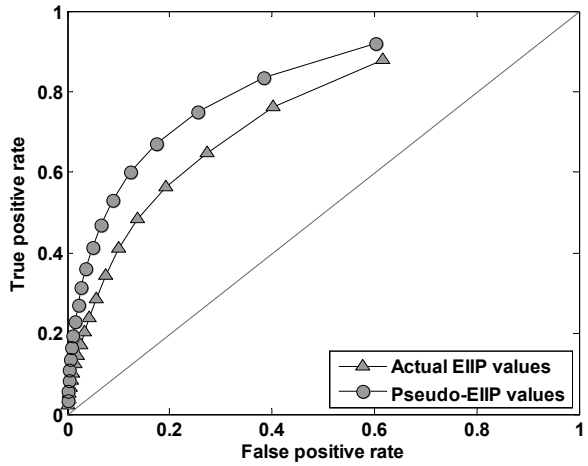

Fig. 6. ROC curves for the actual EIIP and optimized characteristic (pseudo-EIIP) values for the training set.

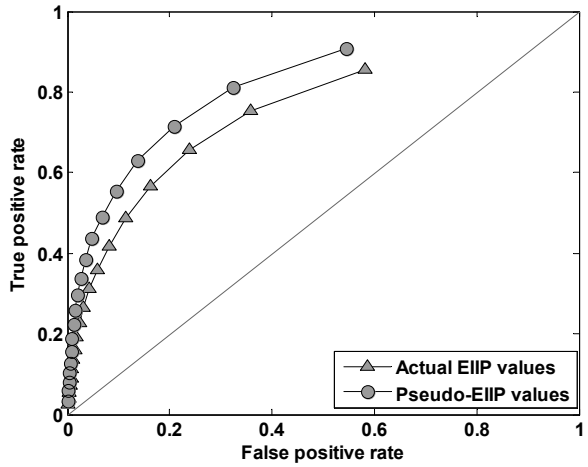

Fig. 7. ROC curves for the actual EIIP and optimized characteristic (pseudo-EIIP) values for the test set.

The selected subset of 160 DNA sequences has been already subjected to non-redundancy testing as described in [50]. Hence, to obtain our training and test sets, the subset of 160 sequences was divided into two smaller sets of 80 sequences each. One set was used for training, i.e., to obtain the optimized characteristic values, and the other was used for independent exon evaluation. The fact that the sequences in the HMR195 dataset are non-redundant prevents the occurrence of any type of training bias in the test results.

The pseudo-EIIP values obtained are compared with the actual 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 for the training set with the actual EIIP and pseudo-EIIP values are illustrated in Fig. 6. As can be seen, the pseudo-EIIP values yield better results. To test the usefulness of the pseudo-EIIP values, we obtained the ROC curves of Fig. 7 using the test set described above that is non-redundant compared to the training set. As can be seen in Fig. 7, the pseudo-EIIP values again outperformed the actual EIIP values. The optimum threshold based on Fig. 7 was found to be 0.15.

TABLE III
EXON-LOCATION TECHNIQUES FOR COMPARISON

Group	Code	Description and Reference
Binary based	TIW	Tiwari and Ramachandran [23]
	ANA	Anastassiou [26]
	KOT	Kotlar and Lavner [28]
	ASF	Datta and Asif [29]
	BPB	BPN with binary sequences
EIIP based	STE	Nair and Sreenadhan [30]
	BPE	BPN using EIIP values
	PSE	BPN using pseudo-EIIP values

C. Optimized Binary Weights for Exon Location

Essentially the same procedure as that used in Section IV was used to optimize the weights of the filtered binary sequences in (2). This work has revealed that the importance of the four binary nucleotide sequences to exon location varies significantly among the four sequences with the T sequence being the least important. On the basis of this fact, the T sequence can be ignored in binary-based techniques without introducing a significant degradation in exon-location accuracy thereby reducing the amount of computation by 25% (see [45] for details).

D. Comparative Analysis of Exon-Location Techniques

In this section, we compare our filter-based methodology with five known DSP-based exon-location techniques that make use of the period-3 property. The various techniques investigated are listed in Table III. Techniques TIW, ANA, KOT, and STE use the STDFT along with the rectangular window while technique ASF uses the STDFT along with the Bartlett window to obtain improved results. The rectangular window introduces large Gibbs oscillations which appear as noise in the amplitude spectrum of the processed sequence. Nonrectangular windows such as the Bartlett and Kaiser windows, on the other hand, have reduced ripple ratios [38, Chap. 7] and, therefore, introduce Gibbs oscillations of reduced amplitude. The Kaiser window has the additional advantage that the ripple ratio can be adjusted to suit the application by adjusting an independent window parameter α.

A sample ROC curve obtained by applying technique ANA to a set of sequences from the HMR195 data set using the three types of windows is shown in Fig. 8. As can be seen, the Bartlett and Kaiser windows give almost identical results that are significantly better than those obtained using the rectangular window. We chose the Kaiser window for our experiments with techniques TIW, ANA, KOT, and STE, because of its flexibility but retained the Bartlett window for technique ASF as used originally in [29]. We tried a range of values for parameter α of the Kaiser window in the interval $[0.5, 8.0]$ and a value of 5.0 was found to yield the best results. For all the windows, a length of 351 was used as suggested in [23], [26].

The specifications and design parameters for the BPN filters were the same as those used for hot-spot location except that the center frequency was set to the period-3 frequency for the present experiments. An inverse-Chebyshev lowpass filter was used for the demodulation of the signal. It was designed to

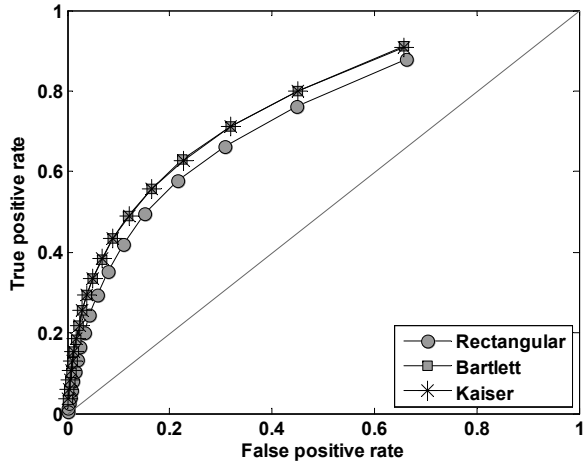

Fig. 8. ROC curves for exon location using technique ANA with the rectangular, Bartlett, or Kaiser window.

have a maximum passband attenuation of 1 dB and a minimum stopband attenuation of 80 dB. The passband and stopband edges for this filter were set to 0.4 and 0.5, respectively, as these specifications were found to yield good results. The required filter order was 14.

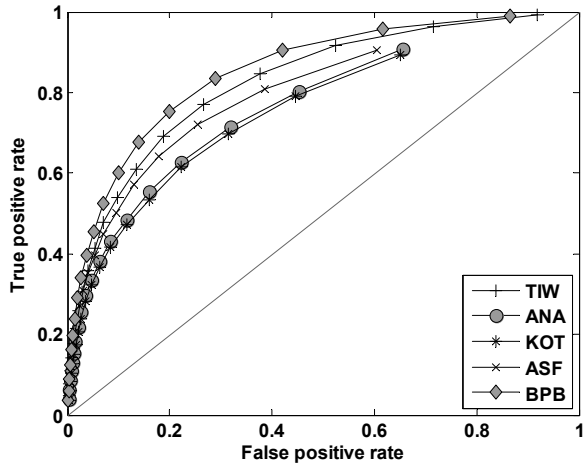

Fig. 9. ROC curves for exon location using the binary-based techniques.

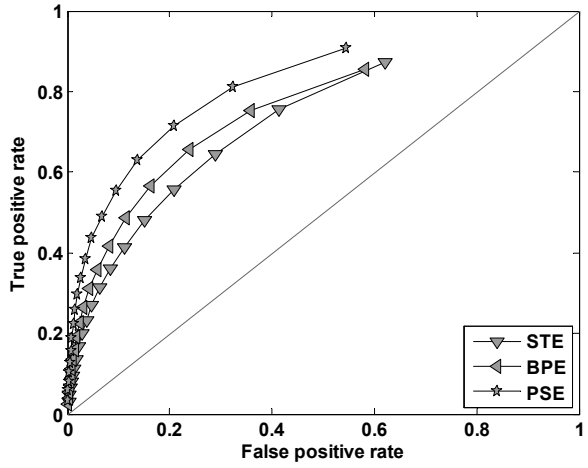

Fig. 10. ROC curves for exon location using the EIIP-based techniques.

TABLE IV
BEST OPERATING THRESHOLDS, EUCLIDEAN DISTANCES, AND χ^2 FOR EXON LOCATION

Type of technique‡	Best threshold	Euclidean distance	χ^2
BPB	0.25	0.3181	102492.2
PSE	0.15	0.3526	86313.7
TIW	0.25	0.3528	76729.5
ASF	0.15	0.3776	67738.6
BPE	0.15	0.4175	57109.4
ANA	0.15	0.4288	45314.4
KOT	0.15	0.4373	42701.7
STE	0.15	0.4578	38400.3

‡ The techniques are sorted from best to worst accuracy.

TABLE V
EVALUATION METRICS AT BEST OPERATING THRESHOLDS FOR EXON LOCATION

Type‡	*TPR*	*FPR*	*Sp*	*Prc*	*Acc*	*g*-mean	*MCC*
BPB	0.753	0.200	0.800	0.412	0.792	0.776	0.445
PSE	0.715	0.208	0.792	0.391	0.780	0.753	0.408
TIW	0.770	0.267	0.733	0.350	0.739	0.751	0.385
ASF	0.722	0.256	0.744	0.345	0.741	0.733	0.361
BPE	0.657	0.237	0.763	0.340	0.746	0.708	0.332
ANA	0.715	0.320	0.680	0.294	0.685	0.697	0.296
KOT	0.697	0.315	0.685	0.292	0.687	0.691	0.287
STE	0.646	0.290	0.710	0.293	0.700	0.677	0.272

‡ The techniques are sorted from best to worst accuracy.

The various techniques were tested using the same test set as that used in Section V-B. The curves obtained for the binary-based and EIIP-based techniques are shown in Figs. 9 and 10, respectively. The optimum operating threshold, the lowest Euclidean distance from point $(0, 1)$, and the χ^2 statistic for each case are listed in Table IV. Table V, on the other hand, lists the values of the evaluation metrics when the techniques were operated at their optimum thresholds. In both tables, the techniques have been sorted in decreasing order of accuracy, i.e., decreasing g-mean values. From these figures and tables, it can be seen that, overall, filters perform better than the STDFT. On the other hand, our optimized characteristic values for the nucleotides, i.e., the pseudo-EIIP values, yield significant improvements in the accuracy relative to that achieved with the actual EIIP values thereby advancing our method from the 5th position (BPE method) to the 2nd position (PSE method) in Table V. The pseudo-EIIP values also perform significantly better than the optimized complex values used in technique ANA. This may be due to the fact that the optimization criterion used to obtain the pseudo-EIIP values is more generic as it takes into account DNA sequences from three different organisms. On the other hand, the criterion used to obtain the complex values was based on exons from a specific chromosome of a single organism. Moreover, the pseudo-EIIP values were obtained by maximizing the prediction accuracy of coding and noncoding regions in real DNA while the complex values were obtained by maximizing the discriminatory capability between coding regions in real DNA and noncoding regions in random synthetic DNA.

The statistical significance of the predictions was tested

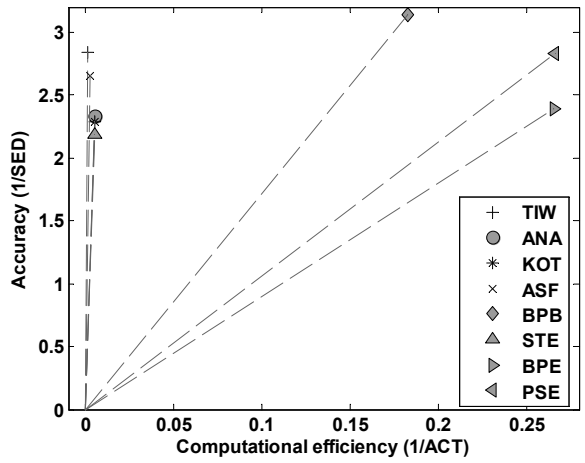

Fig. 11. Overall accuracy versus computational efficiency for exon-location techniques.

using the p-values that were obtained from the χ^2 statistic. These values were less than 0.0001 for all the techniques compared, thereby confirming that the predictions are statistically significant.

The computational efficiency of the techniques was evaluated as described in Section III-F.[2] The sequence used for the computation was 9163 nucleotides long with identifier AB018249. The results obtained, plotted in Fig. 11, show that the filter-based methodology in general requires only a small fraction of the computational effort required by the STDFT-based techniques.

Among the STDFT-based techniques, technique TIW requires the most computational effort as it requires the computation of four separate STDFTs. Technique ASF uses only two binary sequences (for reasons that have not been stated in [29]), and hence it is two times faster than TIW. Techniques KOT and ANA substitute complex values for the four nucleotides in order to obtain a single complex sequence. The STDFT of this sequence is then computed. Hence, these techniques are approximately four times faster than technique TIW.

Considering that in practice DNA sequences can typically be much longer than the ones used in our examples, the computational savings achieved by the use of our filter-based methodology would be substantial. In addition, our methodology yields better accuracy than the STDFT as seen from Table V. By using binary instead of EIIP sequences, our methodology yields the best accuracy although the computational efficiency is somewhat compromised.

DSP-based methodologies are relatively simple to use and entail minimal computational effort compared to other more sophisticated model-based approaches in that they need a minimal amount of data, namely, the protein or DNA sequences. This simplicity comes about at the cost of reduced predictive power and, in effect, there is a tradeoff between efficiency and accuracy. However, in practice, there are many situations where DSP-based methodologies would be preferable; for example, in situations where the full range of data required by the more sophisticated methodologies is not available or is difficult to obtain as would be the case when new proteins or exons are discovered.

[2]The `tic` and `toc` commands in MATLAB were used. To maximize accuracy, the following precautions were taken while computing the CPU time: all applications except MATLAB were terminated, a fresh session of MATLAB was started for the task, and MATLAB was warmed up with the code, i.e., the first run of the code was ignored.

E. Location of Short Exons

The length of the window used for the STDFT-based techniques must be sufficiently large to ensure that the period-3 frequency component dominates over the background noise in the frequency spectrum. Based on empirical studies, a window length of 351 has been found to yield a reasonably good signal-to-noise ratio for effectively identifying exon locations [23], [26]. A window of this length, however, compromises the base-domain resolution thereby limiting the capability of the STDFT to locate short exons. In [29], it has been reported that the STDFT does not perform well for the location of exons that are shorter than 150 nucleotides in length. Digital filters, on the other hand, identify localized peaks better and thereby locate short exons more accurately. This is illustrated in Fig. 12 using a couple of example DNA sequences. Three plots are shown for each example illustrating the exon locations identified by techniques ASF, TIW, and BPB, respectively. Gene AF001689 has a total of five exons out of which three are shorter than 150 nucleotides while gene AF037438 has a total of six exons out of which three are shorter than 150 nucleotides as indicated. As can be seen, for both genes our filter-based methodology identifies the locations of the short exons with higher signal amplitudes and more prominent peaks than the STDFT-based techniques.

VI. Conclusion

By using the ROC technique along with optimization, we have been able to achieve significant improvements in our filter-based methodology for the location of hot spots in protein sequences and exons in DNA sequences. By obtaining optimized characteristic values for the nucleotides, we refer to as pseudo-EIIP values, we have been able to achieve improved accuracy along with high efficiency. On the other hand, by using our methodology with binary nucleotide sequences, we have been able to achieve the best accuracy although the efficiency is somewhat compromised relative to that achieved with optimized characteristic values. Extensive experimental results, evaluated using metrics such as the g-mean, the Matthews correlation coefficient, and the chi-square statistic, have shown that our methodology performs much better than existing techniques that use the STDFT along with a rectangular, Bartlett, or Kaiser window particularly in applications where short exons less than 150 nucleotides long are involved.

References

[1] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *Journal of Molecular Biology*, vol. 280, pp. 1–9, 1998.
[2] J. W. Fickett, "The gene identification problem: An overview for developers," *Comp. & Chem.*, vol. 20, no. 1, pp. 103–118, 1996.
[3] J. A. Wells, "Systematic mutational analyses of protein-protein interfaces," *Methods in Enzymology*, vol. 202, pp. 390–411, 1991.

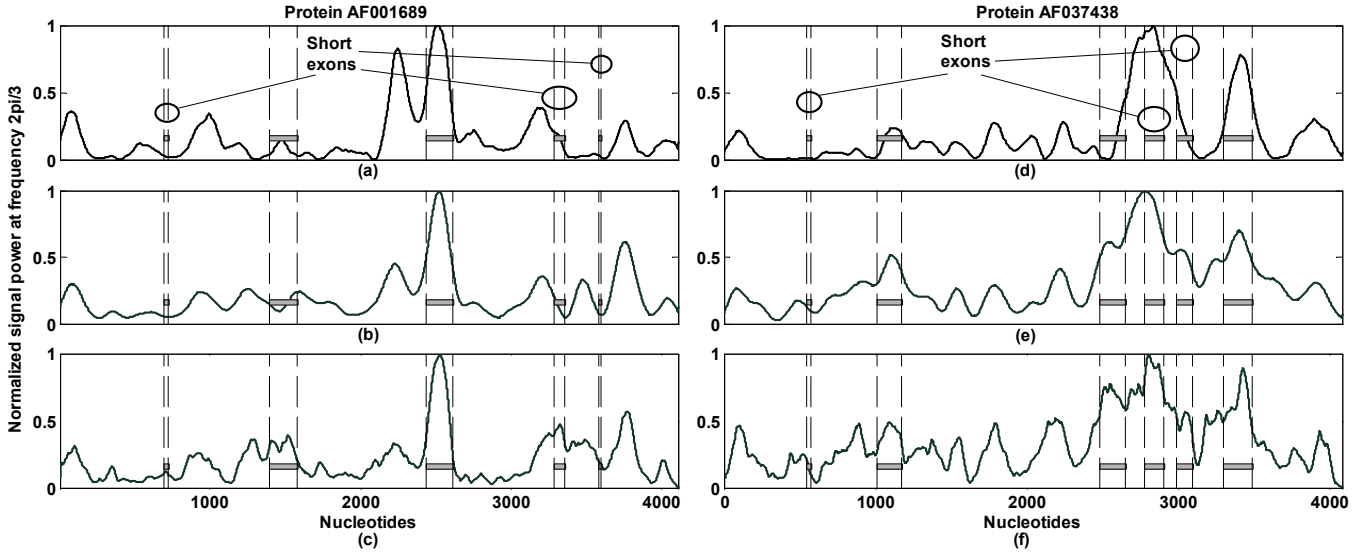

Fig. 12. Identification of short exons (less than 150 nucleotides) for genes AF001689 and AF037438: (a) and (d) correspond to ASF, (b) and (e) correspond to TIW, and (c) and (f) correspond to BPB.

[4] W. L. DeLano, "Unraveling hot spots in binding interfaces: Progress and challenges," *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 14–20, 2002.

[5] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol.*, vol. 268, pp. 78–94, 1997.

[6] A. Delcher, K. Bratke, E. Powers, and S. Salzberg, "Identifying bacterial genes and endosymbiont DNA with Glimmer," *Bioinformatics*, vol. 23, no. 6, pp. 673–679, 2007.

[7] E. Birney, M. Clamp, and R. Durbin, "Genewise and genomewise," *Genome Res.*, vol. 14, pp. 988–995, 2004.

[8] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein-protein complexes," *PNAS*, vol. 99, no. 22, pp. 14 116–14 121, Oct. 2002.

[9] Y. Gao, R. Wang, and L. Lai, "Structure-based method for analyzing protein-protein interfaces," *Journal of Molecular Modeling*, vol. 10, no. 1, pp. 44–54, Feb. 2004.

[10] S. Lise, C. Archambeau, M. Pontil, and D. T. Jones, "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods," *BMC Bioinformatics*, vol. 10, p. 365, 2009.

[11] S. Lise, D. Buchan, M. Pontil, and D. T. Jones, "Predictions of hot spot residues at protein-protein interfaces using support vector machines," *PLoS ONE*, vol. 6, no. 2, p. e16774, 2011.

[12] K. Cho, D. Kim, and D. Lee, "A feature-based approach to modeling protein-protein interaction hot spots," *Nucleic Acids Res.*, vol. 37, no. 8, pp. 2672–2687, 2009.

[13] J.-F. Xia, X.-M. Zhao, J. Song, and D.-S. Huang, "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *BMC Bioinf.*, vol. 11, p. 174, 2010.

[14] I. Meliciani, K. Klenin, T. Strunk, K. Schmitz, and W. Wenzel, "Probing hot spots on protein-protein interfaces with all-atom free-energy simulation," *J Chem Phys.*, vol. 131, no. 3:(034114), 2009.

[15] N. Tuncbag, A. Gursoy, and O. Keskin, "Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy," *Bioinformatics*, vol. 25, no. 12, pp. 1513–1520, 2009.

[16] S. J. Darnell, L. LeGault, and J. C. Mitchell, "KFC Server: Interactive forecasting of protein interaction hot spots," *Nucleic Acids Res.*, vol. 36, pp. W265–W269, 2008.

[17] N. Tuncbag, O. Keskin, and A. Gursoy, "HotPoint: hot spot prediction server for protein interfaces," *Nucleic Acids Res.*, vol. 38, pp. W402–W406, 2010.

[18] S. Huo, I. Massova, and P. A. Kollman, "Computational alanine scanning of the 1:1 human growth hormone-receptor complex," *J. Comput. Chem.*, vol. 23, no. 1, pp. 15–27, 2002.

[19] D. Rajamani, S. Thiel, S. Vajda, and C. J. Camacho, "Anchor residues in protein-protein interactions," *PNAS*, vol. 101, no. 31, pp. 11 287–11 292, 2004.

[20] Y. Ofran and B. Rost, "Protein-protein interaction hotspots carved into sequences," *PLoS Comput Biol.*, vol. 3, no. 7, p. e119, 2007.

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

[22] V. Veljković, I. Cosić, B. Dimitrijević, and D. Lalović, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?" *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 5, pp. 337–341, May 1985.

[23] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by fourier analysis of genomic sequences," *Comput. Appl. Biosci.*, vol. 13, no. 3, pp. 263–270, 1997.

[24] I. Cosic, "Macromolecular bioactivity: Is it resonant interaction between macromolecules?—Theory and applications," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 12, pp. 1101–1114, Dec. 1994.

[25] P. Ramachandran and A. Antoniou, "Identification of hot-spot locations in proteins using digital filters," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 378–389, Jun. 2008.

[26] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.

[27] P. P. Vaidyanathan and B.-J. Yoon, "Digital filters for gene prediction applications," in *Proc. Thirty-Sixth Asilomar Conf. on Sig., Syst., & Comp.*, Monterey, Nov. 2002, pp. 306–310.

[28] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions," *Genome Research*, vol. 13, no. 8, pp. 1930–1937, 2003.

[29] S. Datta, A. Asif, and H. Wang, "Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics," in *Proc. IEEE Sixth Int. Symp. on Multimedia Software Engineering*, Dec. 2004, pp. 160–163.

[30] 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.

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

[32] J. Tuqan and A. Rushdi, "A DSP perspective to the period-3 detection problem," in *Proc. IEEE Int. Workshop on Genomic Signal Processing and Statistics*, May 2006.

[33] ——, "A DSP approach for finding the codon bias in DNA sequences," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 343–356, Jun. 2008.

[34] P. Ramachandran, A. Antoniou, and P. P. Vaidyanathan, "Identification and location of hot spots in proteins using the short-time discrete Fourier transform," in *Proc. Thirty-Eighth Asilomar Conf. on Sig., Syst., & Comp.*, Pacific Groove, CA, Nov. 2004, pp. 1656–1660.

[35] K. Chen, J. T. Huzil, H. Freedman, P. Ramachandran, A. Antoniou, J. A. Tuszynski, and L. Kurgan, "Identification of tubulin drug binding sites and prediction of relative differences in binding affinities to tubulin isotypes using digital signal processing," *Journal of Molecular Graphics and Modelling*, vol. 27, no. 4, pp. 497–505, Nov. 2008.

[36] 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. Cir. Syst.*, Seattle, May 2008, pp. 2673–2676.
[37] ——, "Tuning technique for the location of hot spots in proteins using a bandpass notch digital filter," in *Proc. IEEE Int. Workshop on Genomic Signal Processing and Statistics*, Minneapolis, May 2009.
[38] A. Antoniou, *Digital Signal Processing: Signals, Systems, and Filters*. New York: McGraw-Hill, 2005.
[39] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. New York: Springer, 2007.
[40] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, pp. 353–367, 1996.
[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.
[42] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning*, vol. 30, pp. 195–215, 1998.
[43] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.
[44] P. E. Greenwood and M. S. Nikulin, *A guide to chi-squared testing*. Wiley, 1996.
[45] P. Ramachandran, "New techniques for the location of hot spots in proteins and exons in DNA using digital filters," Ph.D. dissertation, University of Victoria, 2010.
[46] P. Ramachandran and A. Antoniou, "Localization of hot spots in proteins using digital filters," in *Proc. IEEE Int. Symp. on Signal Processing and Information Technology*, Vancouver, Canada, Aug. 2006, pp. 926–931.
[47] P. Ramachandran, W.-S. Lu, and A. Antoniou, "Optimized numerical mapping scheme for filter-based exon location in DNA using a quasi-Newton algorithm," in *Proc. IEEE Int. Symp. Cir. Syst.*, Paris, May 2010.
[48] Protein Data Bank (PDB). Research Collaboratory for Structural Bioinformatics (RCSB). [Online]. Available: http://www.rcsb.org/pdb/
[49] K. S. Thorn and A. A. Bogan, "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions," *Bioinformatics*, vol. 17, no. 3, pp. 284–285, 2001.
[50] S. Rogic, A. K. Mackworth, and F. B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome Research*, vol. 11, no. 5, pp. 817–832, May 2001.
[51] Y. Zhou, Y. Liang, C. Hu, L. Wang, and X. Shi, "An artificial neural network method for combining gene prediction based on equitable weights," *Neurocomputing*, vol. 71, pp. 538–543, 2008.