Location of Exons in DNA Sequences Using Digital Filters

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*—A filtering technique for the location of hot spots in proteins proposed recently is applied for the location of exons in DNA sequences. The technique involves conversion of a DNA character sequence into a numerical sequence using the electron-ion interaction potential values and then filtering the numerical sequence using a narrowband bandpass digital filter whose passband is centered at the period-3 frequency, i.e., $2\pi/3$. The strength of the bandpass-filtered signal as a function of nucleotide location is then detected using a lowpass filter. A plot of the signal power versus location reveals the presence of exons as distinct peaks. Simulations have shown that the technique leads to more accurate exon locations than another computational technique based on the short-time discrete Fourier transform. Furthermore, the amount of computation required is reduced by as much as 97 percent thereby rendering the technique suitable for the processing of long DNA sequences, even complete genomes.**

***Index Terms*—DNA, exons, period-3 property, resonant recognition model (RRM), electron-ion interaction potential (EIIP), narrowband bandpass digital filters.**

I. Introduction

The complete set of instructions to build and maintain a living organism is encoded in its *genome*. The genome is made of DNA which is a biomolecule composed of smaller components called nucleotides [1]. A nucleotide can be one 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 as a string of characters. Regions in a genome that code for proteins are known as *genes*. In organisms with a distinct cellular nucleus, known as *eucaryotes*, the coding regions of genes are usually divided into several disconnected fragments known as *exons*. The noncoding regions occurring between the exons are known as *introns*. Before manufacturing proteins, the introns are removed by the cellular mechanism and the exons are joined together to form a corresponding uninterrupted gene. By such an interleaved arrangement of introns and exons, a type of data compression has evolved over the years that enables a single gene to produce different types of proteins by combining its exons in different ways. Fig. 1 illustrates the organization of genes in eucaryotes.

Accurate location of exons in genomes is very important for understanding life processes. The exon locations help determine protein codes thereby leading to an understanding of protein function. Consequently, knowledge pertaining to exon locations may result in the design of customized drugs and new cures for diseases. A popular strategy of locating exons has been to exploit the *period-3 property*. It turns out that the power spectra of DNA segments that correspond to exons tend to exhibit a relatively strong component at the *period-3 frequency*, i.e., $2\pi/3$, whereas segments that correspond to introns exhibit a relatively weak component at the period-3 frequency. Thus exons can be located by mapping the DNA characters into numbers in some way and then tracking the strength of the period-3 component along the length of the DNA sequence of interest. In a popular character-to-numeric mapping scheme proposed by Voss [2], a DNA sequence is represented by four binary *indicator* sequences, one for each of the four types of nucleotides whereby digits '1' and '0' are used to represent the presence or absence of the nucleotide of interest, e.g., the indicator sequence for nucleotide A in DNA sequence 'ATCCGCTTAGC' would be '10000000100'. Indicator sequences have been employed in [3], [4] for locating exons using the short-time discrete Fourier transform (STDFT). In another computational technique described in [5], a second-order digital filter has been used to process the indicator sequences.

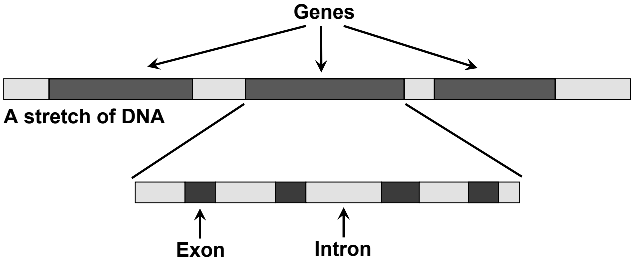


Fig. 1. Organization of genes in eucaryotes.

Electron-ion interaction potential (EIIP) values have been used in [6], [7] for the location of hot spots in proteins. In this technique, a narrowband bandpass digital filter is used to select the characteristic frequency of interest. A plot of the signal power versus location reveals the presence of hot spots as distinct peaks.

In this paper, we apply the filtering technique reported in [6], [7] for the location of exons in DNA sequences. Through a set of examples, we show that the technique yields more accurate exon locations and a much better computational efficiency when compared to a technique based on the STDFT. The paper

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

is organized as follows. Section II defines the EIIP and justifies its use for exon location. Section III presents the details of the technique. Section IV presents the results obtained on applying the technique to a set of example DNA sequences.

II. DNA REPRESENTATION USING EIIP VALUES

As an alternative to the set of indicator sequences proposed by Voss, the nucleotides of a DNA sequence can be mapped onto their EIIP values to obtain a single numerical sequence corresponding to the character sequence. The EIIP of a nucleotide is a physical quantity denoting the average energy of valence electrons in the nucleotide [8]. The EIIP values for the four nucleotides are listed in Table I. The EIIP sequence can be interpreted as a weighted sum of the four indicator sequences with the weights being the EIIP values. It can be represented by

$$\mathbf{x}_{EIIP} = w_A \mathbf{x}_A + w_T \mathbf{x}_T + w_G \mathbf{x}_G + w_C \mathbf{x}_C \qquad (1)$$

where w_A, w_T, w_G, w_C are the EIIP values and $\mathbf{x}_A$, $\mathbf{x}_T$, $\mathbf{x}_G$, $\mathbf{x}_C$ are the corresponding indicator sequences for the four nucleotides. The EIIP sequence is a better choice for numerically representing DNA when compared to indicator sequences for the following reasons. First, it involves only a single sequence instead of four in the case of indicator sequences and, secondly, it is biologically more meaningful as it represents a physical property when compared to the indicator values which represent just the presence or absence of a nucleotide.

In the past, EIIP values have been successfully applied for analyzing proteins in numerous studies such as those in [6], [7], [9]. Among over 200 different types of numerical mapping schemes, EIIP values have been shown to provide the most suitable mapping for spectral analysis of protein sequences [10]. In view of these works and also the well-known interrelations between the functioning of DNA and proteins, there is enough evidence to believe that EIIP values can be effectively applied to analyze DNA sequences.

A computational technique for the location of exons based on the use of EIIP values has been reported in [11]. In this technique, exons are identified by tracking the period-3 frequency using the STDFT.

III. EXON LOCATION USING DIGITAL FILTERS

From a DSP perspective, the protein hot-spot location problem using the resonant recognition model (RRM) and the exon location problem using the period-3 property are similar in nature, as both involve tracking the strength of a single frequency component along the length of a protein or a DNA sequence.

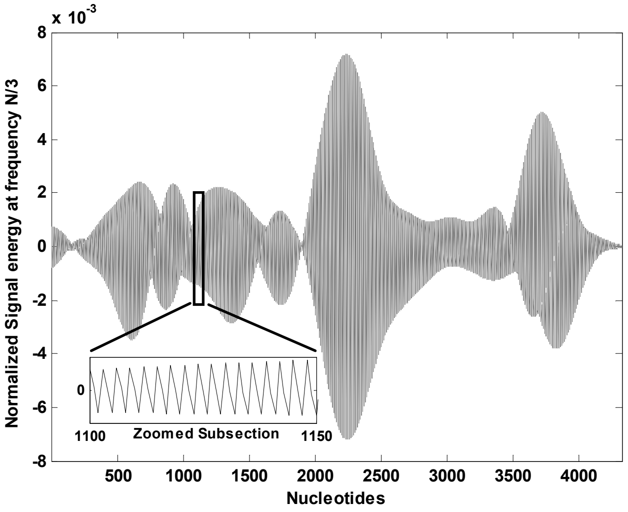


Fig. 2. Plot of $y[n]$ for gene AF039307.

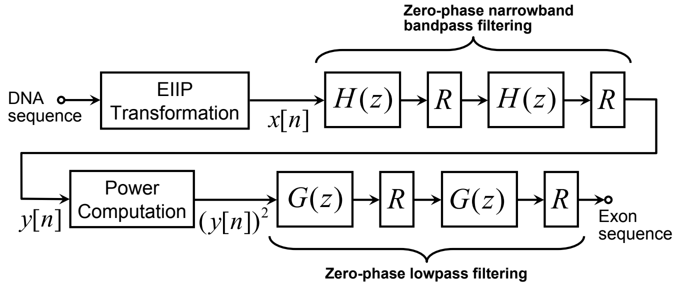
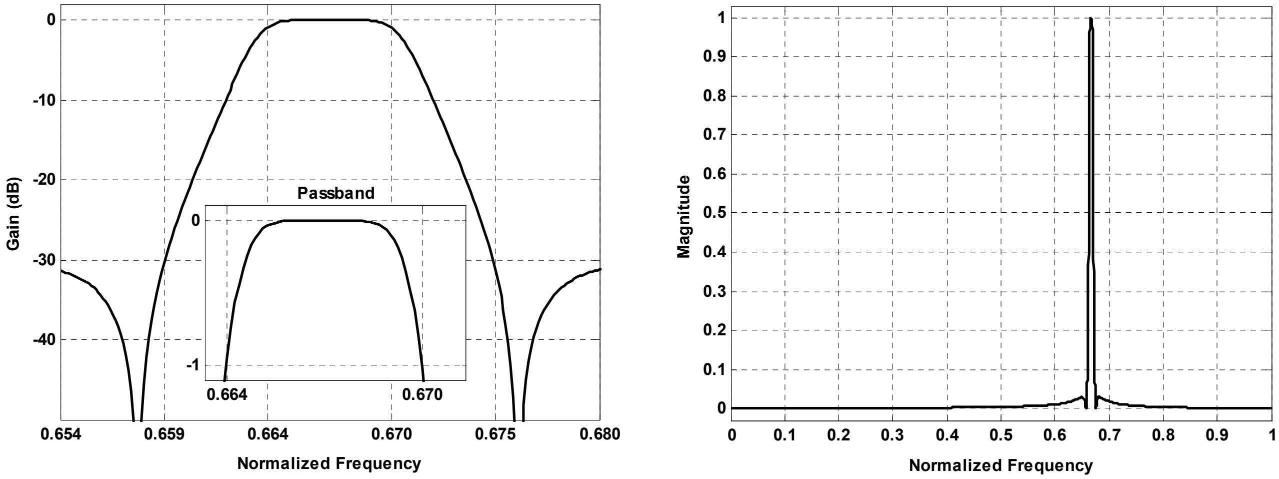


Fig. 3. The complete exon location system.

In our technique, for the location of exons in a DNA sequence, a series of DNA nucleotides are assigned EIIP values $eiip_1, eiip_2, \ldots, eiip_n, \ldots$ and a discrete signal is constructed as

$$x_{EIIP}[n] = eiip_n \qquad (2)$$

Then, a narrowband bandpass digital filter with its passband centered at the period-3 frequency of $2\pi/3$ is used to filter the DNA sequence. In order to eliminate the need to compute the phase response of the filter and to eliminate phase distortion, *zero-phase filtering* is employed [12]. The filtered output signal, $y[n]$, represents the period-3 component present in the DNA sequence, which oscillates about zero, and its amplitude at any one location would depend on whether an exon is present or not. A sample plot of $y[n]$ is shown in Fig. 2. In effect, this is an amplitude-modulated signal and as such its strength as a function of DNA location can be detected by filtering the power of the signal, $(y[n])^2$, using a lowpass filter as illustrated in Fig. 3. The locations of exons are identified as well defined peaks in the waveform of $(y[n])^2$. Since the original component to be identified has a fixed frequency, namely, $2\pi/3$, the bandpass and lowpass filters need to be designed only once per computer session.



(a)

(b)

Fig. 4. Amplitude response of the narrowband bandpass inverse-Chebyshev filter: (a) decibel (dB) scale, (b) linear scale.

TABLE II
AVERAGE CPU TIMES

Gene identifier	Sequence length	Average CPU time (milliseconds)	
		Filter-based technique	**STDFT-based technique**
AB009589	12414	16.9	553.7
AF039307	4322	6.2	194.7
AF042784	2234	3.5	101.1
AF009614	5195	7.4	236.0
AB003306	5006	7.1	224.9

IV. RESULTS

To evaluate the performance of the proposed exon location technique, we applied it to a set of five genes whose sequences and true exon locations were downloaded from the well-known HMR195 dataset [13]. The identifiers (known as accession numbers) for the genes are AB009589, AF039307, AF042784, AF009614, and AB003306.

We used an inverse-Chebyshev bandpass filter as it offers good selectivity for a reasonably low filter order along with a monotonic passband amplitude response. The amplitude response of the filter is illustrated in Fig. 4. A maximum passband attenuation of 1 dB and a minimum stopband attenuation of approximately 30 dB were found to give good results. The lower and upper passband edges were set to 0.664 and 0.670, respectively, while the lower and upper stopband edges were set to 0.659 and 0.675, respectively. The minimum filter order satisfying these specifications was found to be 6. For the inverse-Chebyshev lowpass filter, the maximum passband attenuation and minimum stopband attenuation were set to 1 dB and 80 dB, respectively, and the passband and stopband edges were set to 0.4 and 0.5, respectively, as these specifications were found to give good results. The minimum filter order satisfying these specifications was found to be 14.

The exon locations were identified by our technique for all the five genes. For comparison, we also implemented the STDFT-based technique employed in [11]. Results from both the techniques for genes AF039307 and AF009614 are shown in Figs. 5, 6, 7, and 8. The shaded blocks represent true exon locations. From the figures, it can be seen that the filter-based technique located exons with better accuracy when compared to the STDFT-based technique. For gene AF039307, the filter-based technique exhibited a definite peak corresponding to the exon near location 4000 while the STDFT-based technique did not exhibit a definite peak. For gene AF009614, the filter-based technique exhibited a single well-defined peak corresponding to the exon near location 4000 while the STDFT-based technique exhibited two peaks leading to ambiguous results.

To evaluate the computational efficiency of the filter-based technique when compared to the STDFT-based technique, we computed the average CPU times over 1000 runs of the techniques for the five genes using the `tic` and `toc` commands in MATLAB. The results are listed in Table II. From the table, we can see that the filter-based technique consistently requires only about 3% of the computational load required by the STDFT-based technique. Considering that DNA sequences are typically much longer than the ones used for our examples, the computational savings achieved by our technique would be substantial.

We are currently investigating alternative types of filters such as notch bandpass filters to enhance the accuracy and reduce the computational complexity further.

V. CONCLUSION

A filtering technique for the location of hot spots in proteins reported in [6], [7] based on the use of EIIP values and digital filters was applied for the location of exons in DNA sequences. The technique was applied to a set of example DNA sequences from a well-known database and the results obtained were compared with those obtained using an STDFT-based technique. The comparisons revealed that the proposed

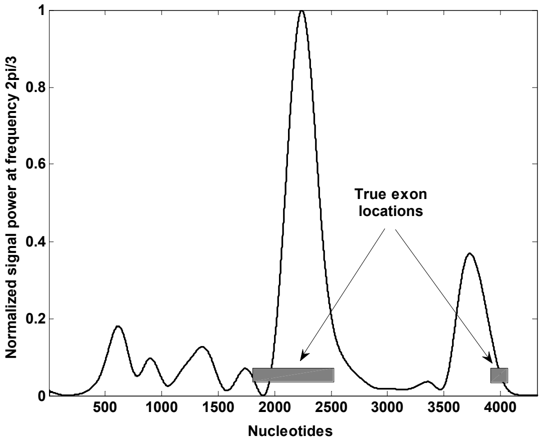


Fig. 5. Exon locations for gene AF039307 predicted by the filter-based technique.

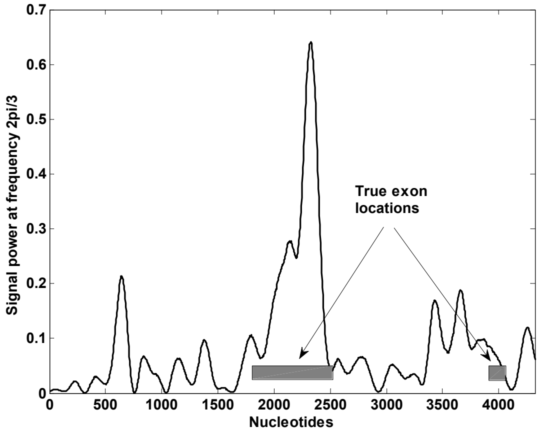


Fig. 6. Exon locations for gene AF039307 predicted by the STDFT-based technique.

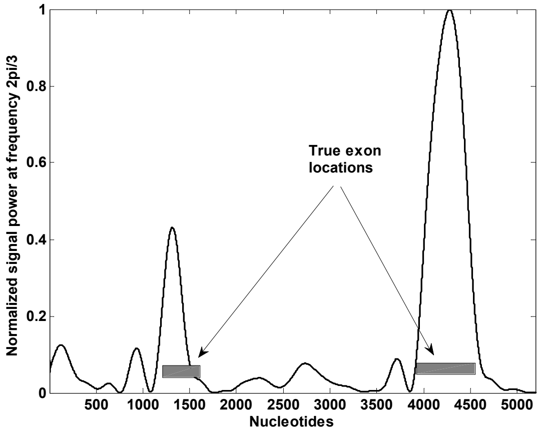


Fig. 7. Exon locations for gene AF009614 predicted by the filter-based technique.

technique is both more accurate and computationally much more efficient than another computational STDFT-based technique. Drawing from the successful application of EIIP values for protein analysis, the application of EIIP values for DNA

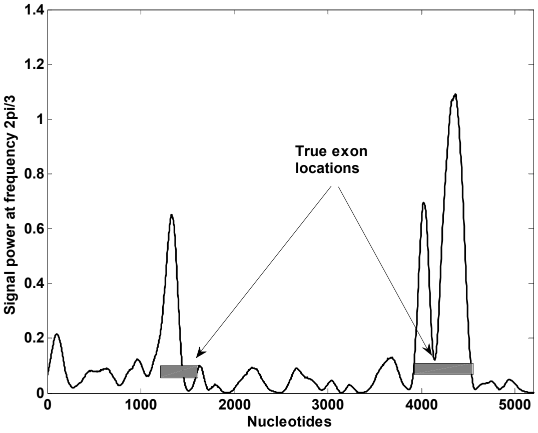


Fig. 8. Exon locations for gene AF009614 predicted by the STDFT-based technique.

analysis may lead to improved modeling of the interrelations between DNA and proteins.

Acknowledgment

The authors are grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) 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] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by fourier analysis of genomic sequences," *Computer Applications in the Biosciences (CABIOS)*, vol. 13, no. 3, pp. 263–270, 1997.

[4] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, pp. 8–20, Jul. 2001.

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

[6] P. Ramachandran, W.-S. Lu, and A. Antoniou, "Improved hot-spot location technique for proteins using a bandpass notch digital filter," in *IEEE International Symposium on Circuits and Systems*, Seattle, May 2008, pp. 2673–2676.

[7] 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.

[8] 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 Transactions on Biomedical Engineering*, vol. BME-32, no. 5, pp. 337–341, May 1985.

[9] 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.

[10] J. Lazović, "Selection of amino acid parameters for Fourier transform-based analysis of proteins," *Computer Applications in the Biosciences (CABIOS)*, vol. 12, no. 6, pp. 553–562, 1996.

[11] 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.

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

[13] 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.