Location of Exons in DNA Sequences Using Digital Filters

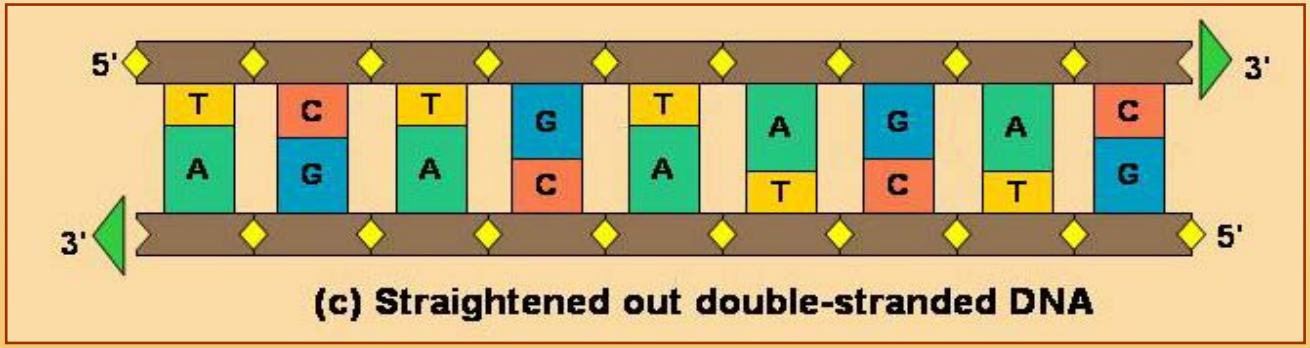
Parameswaran Ramachandran, Wu-Sheng Lu, and Andreas Antoniou

ISCAS, Taipei
May 27, 2009

Department of Electrical Engineering,
University of Victoria, BC, Canada.

DNA

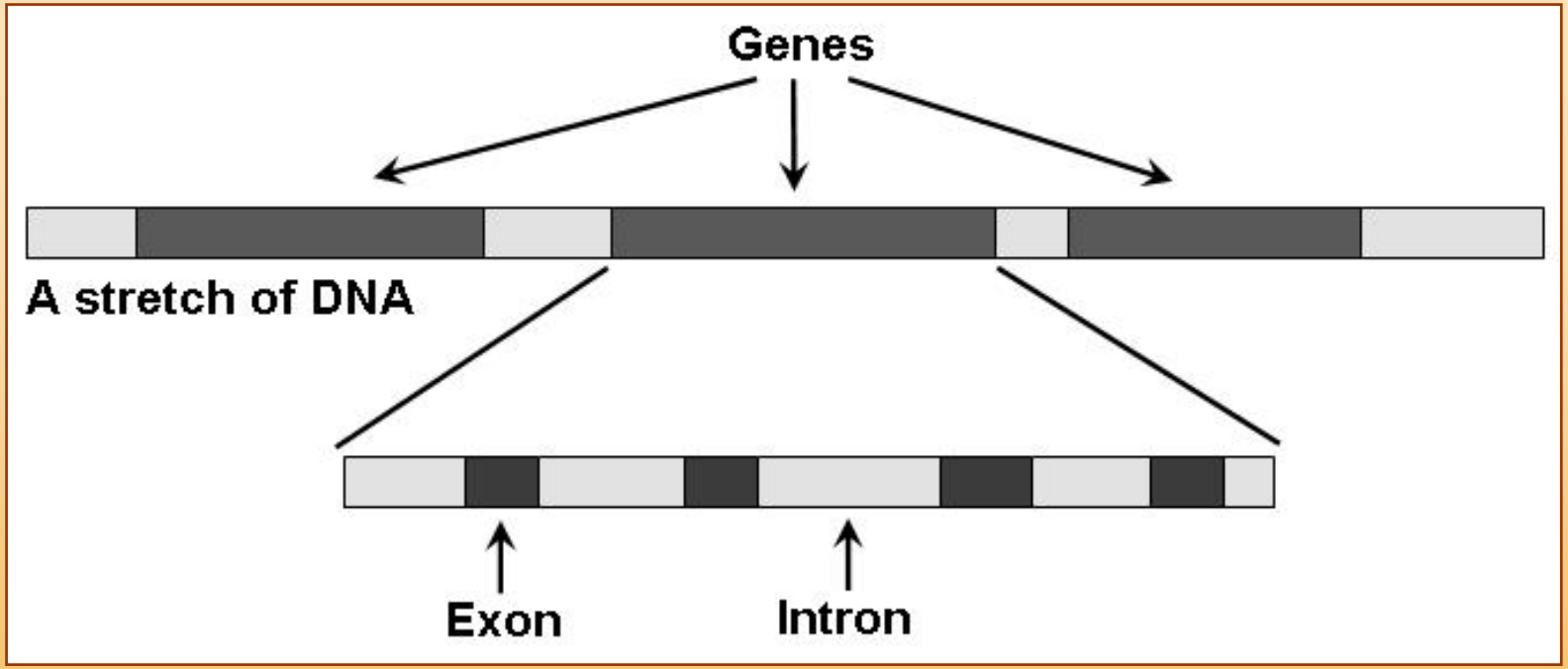
- The instructions to build and maintain a living organism are encoded in its *genome* which is made up of *DNA.*
- DNA is composed of smaller components called *nucleotides.* There are four types of nucleotides denoted by the letters A, T, G, and C.
- DNA comprises a pair of strands. Nucleotides pair up across the two strands. A always pairs with T and G always pairs with C; in effect, the two strands are *complementary.*



(c) Straightened out double-stranded DNA

Genes and Exons

- Regions in a genome that code for proteins are called genes. Genes are further split into coding regions called *exons* and noncoding regions called *introns*.
- Accurate location of exons in genomes is very important for understanding life processes.



Organization of genes.

Location of Exons

- The power spectra of DNA segments corresponding to exons exhibit a relatively strong component at $2\pi/3$. This is known as the period-3 property.
- Thus, exons can be located by mapping the DNA characters into numbers and then tracking the strength of the period-3 component along the length of the DNA sequence of interest.
- Electron-ion interaction potential (EIIP) values have been used by us earlier for the location of hot spots in proteins. A narrowband bandpass digital filter was used to select the characteristic frequency.
- Here, we apply the filtering technique for the location of exons in DNA sequences.

Filter-Based Exon Location Technique

1. The DNA character sequence of interest is mapped onto a numerical sequence using EIIP values.
   - The EIIP of a nucleotide is a physical quantity denoting the average energy of valence electrons in the nucleotide.
   - The EIIP sequence is a weighted sum of four indicator sequences and can be represented by

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

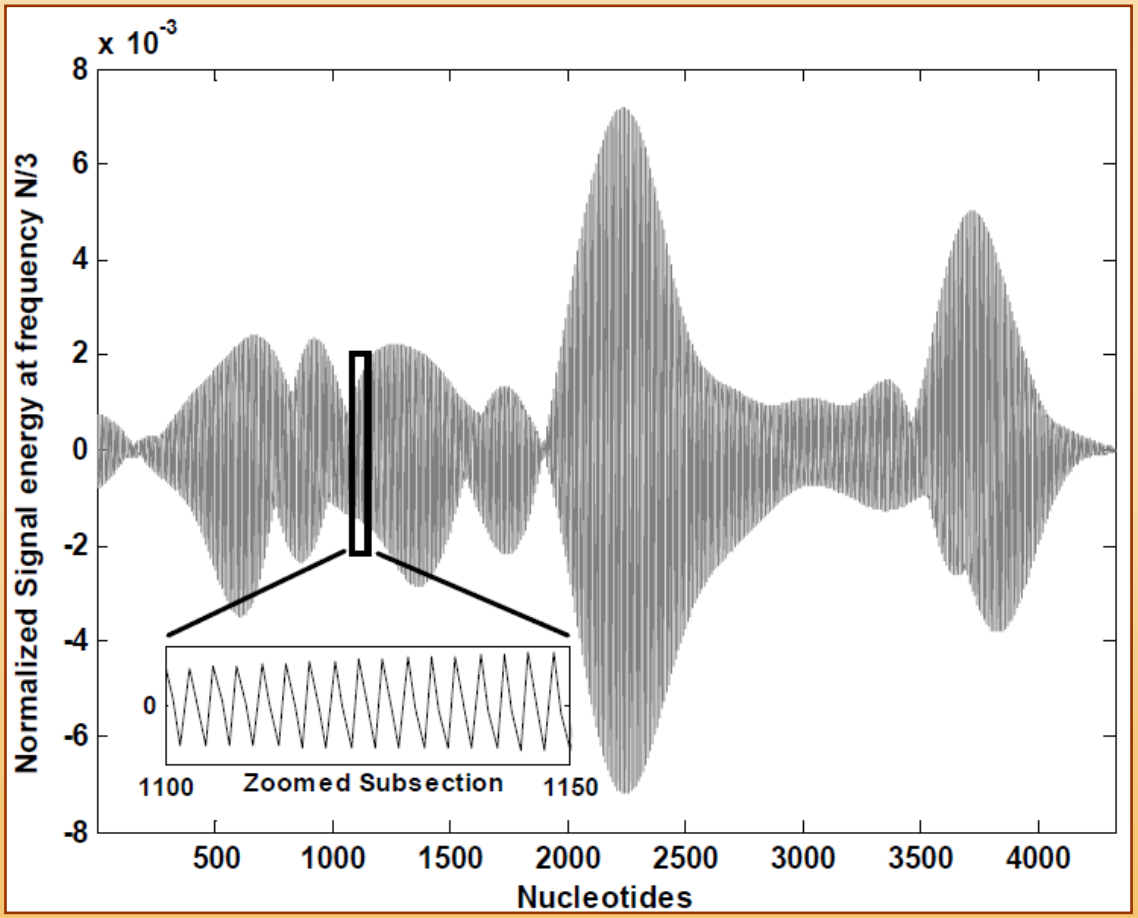
EIIP Values for the Nucleotides

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

Filter-Based Technique (cont'd)

2. A narrowband bandpass digital filter with its passband centered at the period-3 frequency is used to filter the DNA sequence.

3. The filtered output is an amplitude modulated signal as shown below.



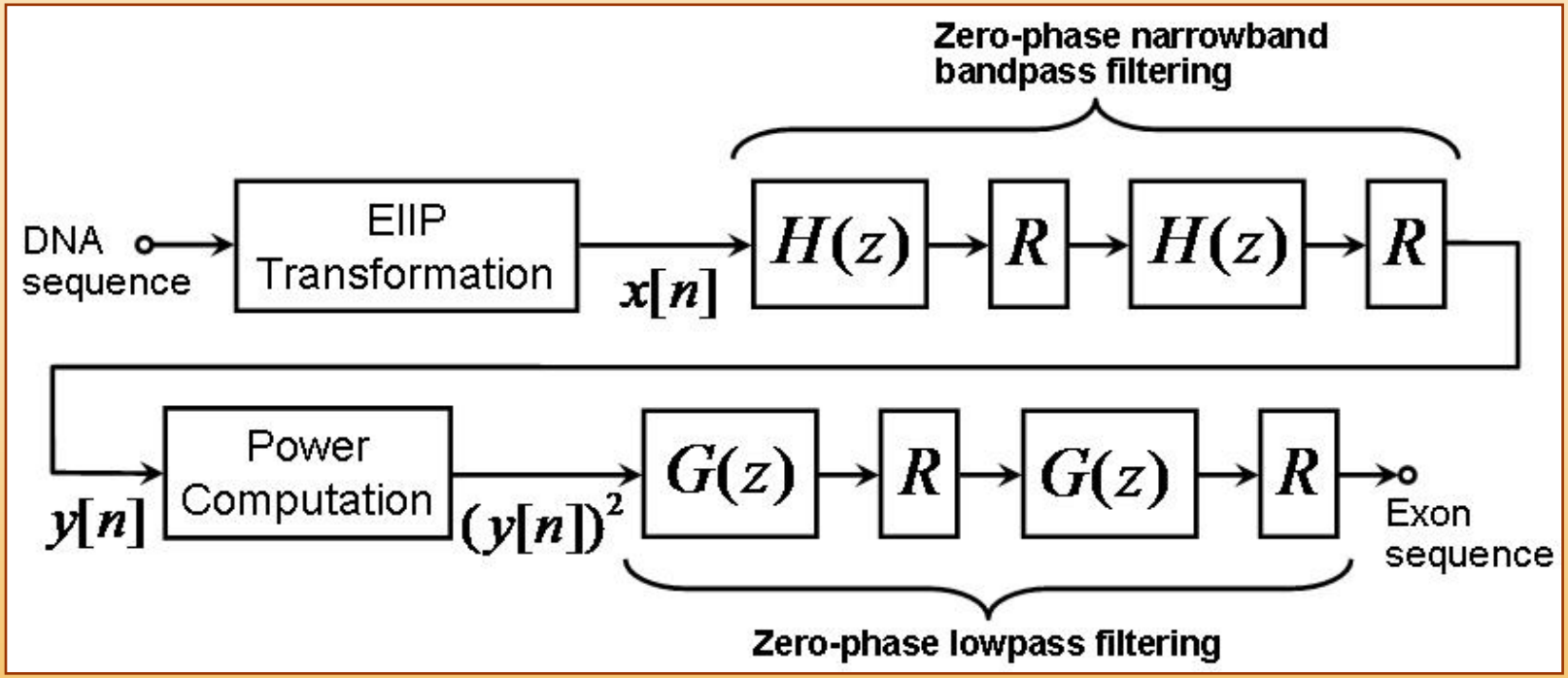
Plot of $y[n]$ for gene AF039307.

Filter-Based Technique (cont'd)

4. The power of the output signal, $(y[n])^2$, is filtered using a lowpass filter to identify the exon locations as distinct peaks.

   To eliminate phase distortion, zero-phase filtering is employed for both the filtering procedures.

   A schematic of the complete system is shown below.

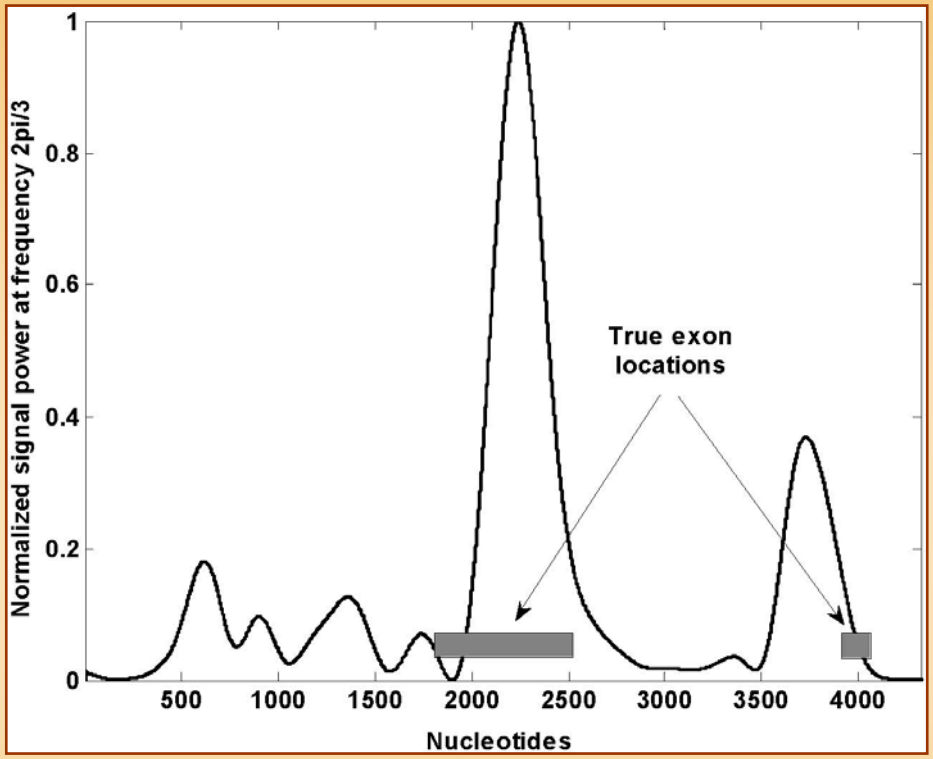


The complete exon location system.

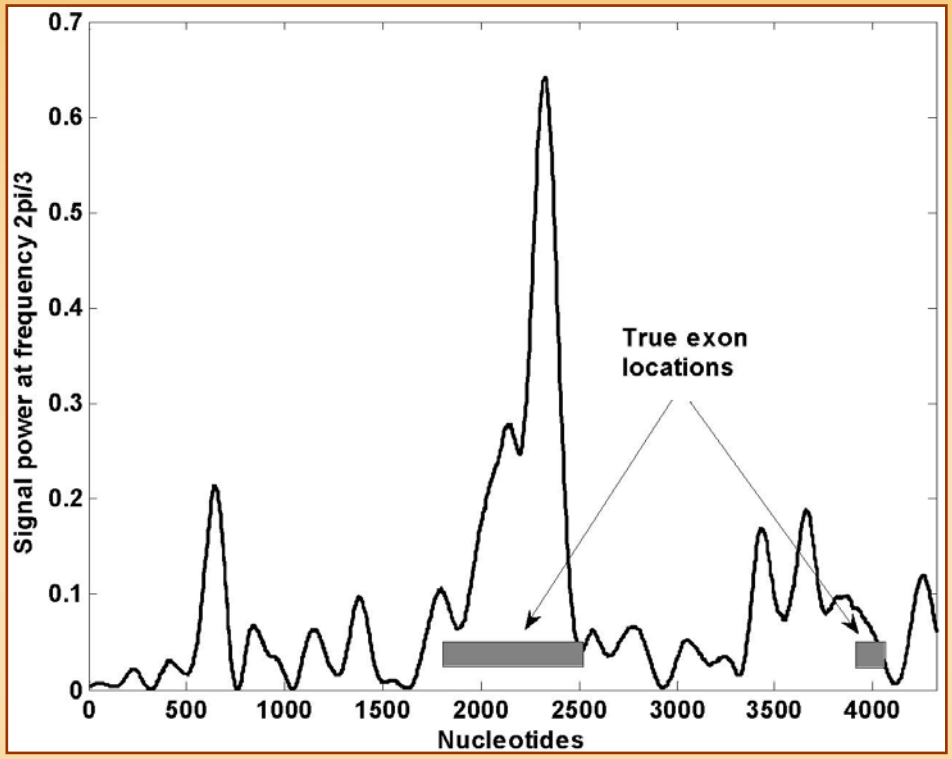
Results

- To evaluate the performance of the proposed technique, we applied it to a set of five genes whose sequences and true exon locations were downloaded from the well-known HMR195 dataset.
- We used an inverse-Chebyshev bandpass filter of order 6 followed by an inverse-Chebyshev lowpass filter of order 14.
- The exon locations were identified by our technique for all the five genes.
- For comparison, we also implemented the STDFT technique employed by Nair and Sreenadhan (2006).

Results (cont'd)



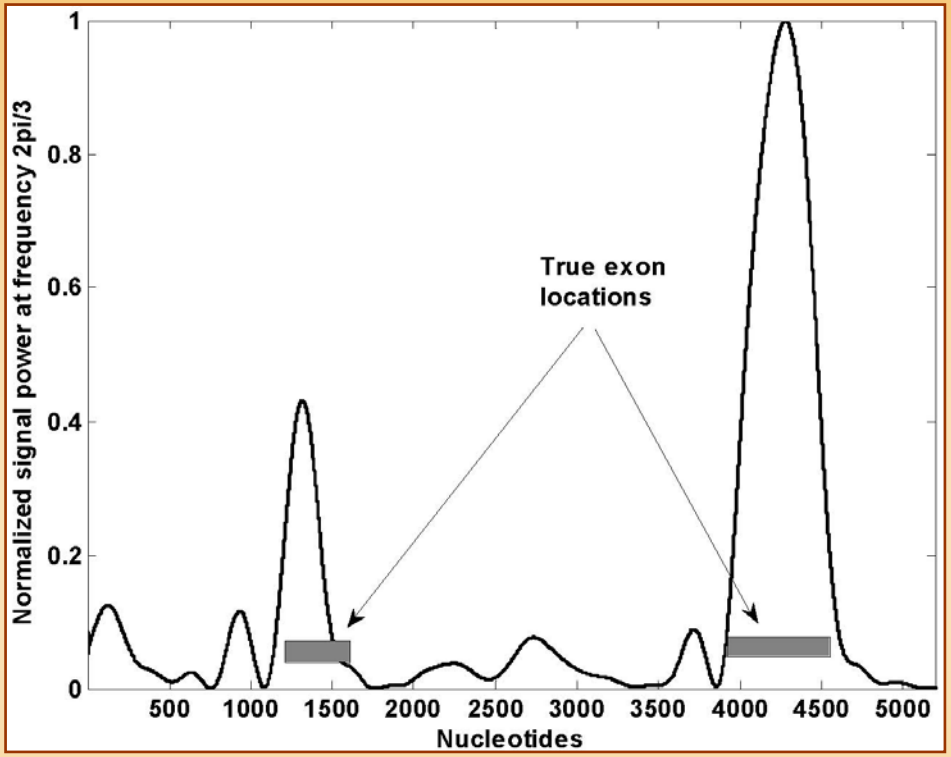
Exon locations for gene AF039307 predicted by the filter-based technique.



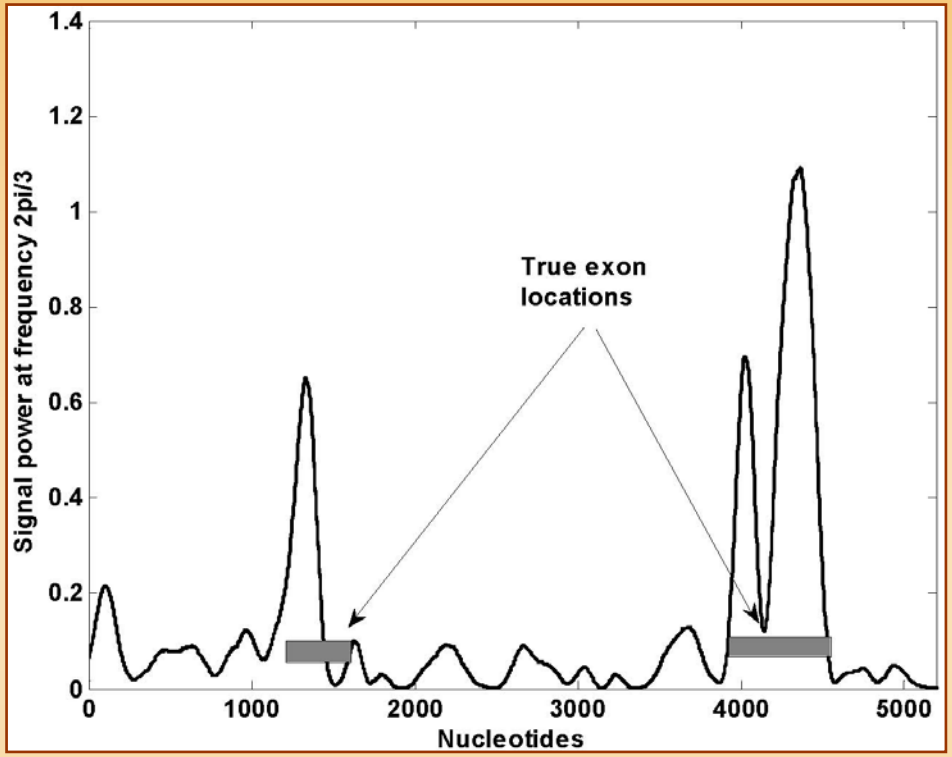
Exon locations for gene AF039307 predicted by the STDFT-based technique.

- Note that for the exon near location 4000, the filter-based technique exhibits a well-defined peak, while the STDFT technique does not.

Results (cont'd)



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



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

- Note that for the exon near location 4000, the filter-based technique exhibits a single well-defined peak, while the STDFT technique exhibits two peaks leading to an ambiguity.

Results (cont'd)

- To compare the computational efficiencies of the two techniques, we computed the average CPU times over 1000 runs of the techniques for the five genes.
- Results show that the filter-based technique requires only about 3% of the computational load required by the STDFT-based technique.

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

Conclusions

- A filtering technique for the location of hot spots in proteins reported by us earlier was applied for the location of exons in DNA sequences.
- Results show that the proposed 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 analysis may lead to improved modeling of the interrelations between DNA and proteins.

References

1. 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.
2. D. Anastassiou, “Genomic signal processing,” IEEE Signal Processing Magazine, pp. 8–20, Jul. 2001.
3. 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.
4. S. Rogic, A. K. Mackworth, and F. B. F. Ouellette, “Evaluation of genefinding programs on mammalian sequences,” Genome Research, vol. 11, no. 5, pp. 817–832, May 2001.
5. 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.