EURASIP Journal on Applied Signal Processing 2004:1, 108–114

A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression

Trevor W. Fox

Research and Development Department, Intelligent Engines Corporation, 903 42 St. SW, Calgary, Alberta, Canada T3C-1Y9
Email: tfox@bm.net

Alex Carreira

Department of Electrical and Computer Engineering, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, Canada T2N 1N4
Email: aycarrei@shaw.ca

Received 1 March 2003; Revised 15 September 2003

It has been observed that the protein-coding regions of DNA sequences exhibit period-three behaviour, which can be exploited to predict the location of coding regions within genes. Previously, discrete Fourier transform (DFT) and digital filter-based methods have been used for the identification of coding regions. However, these methods do not significantly suppress the noncoding regions in the DNA spectrum at $2\pi/3$. Consequently, a noncoding region may inadvertently be identified as a coding region. This paper introduces a new technique (a single digital filter operation followed by a quadratic window operation) that suppresses nearly all of the noncoding regions. The proposed method therefore improves the likelihood of correctly identifying coding regions in such genes.

Keywords and phrases: gene prediction, digital filter, DNA.

1. INTRODUCTION

Finding coding regions (exons) in a DNA strand involves searching amongst the many nucleotides that comprise a DNA strand. Typically a DNA molecule contains millions to hundreds of millions of elements [1]. The problem of finding exons in a DNA sequence is well suited to computers because DNA sequences can be represented by data that is easily processed by a computer. DNA strands can be represented by sequences of letters from a four-character alphabet. Convention dictates the use of the letters A, T, C, and G in each element to represent each of the four distinct nucleotides [1]. A nucleotide has two distinct ends: a 3′ end and a 5′ end. A covalent chemical bond links the 5′ end of one nucleotide to the 3′ end of another nucleotide. A DNA strand is comprised of many nucleotides linked in this fashion [1]. The DNA sequence representing a DNA strand consists of the letters A, T, C, and G listed in a left-to-right fashion corresponding to the nucleotides that make up the strand arranged left to right from their 5′ to 3′ ends [1].

A DNA strand can be divided into genes and intergenic spaces. Genes are responsible for protein synthesis. A gene can be further subdivided into exons and introns for cells with a nucleus (eukaryotes) [2]. Cells without a nucleus are called prokaryotes and do not contain introns [2]. The exons, coding regions within genes, are denoted by start and stop codons. Codons are a subsequence of three letters within the DNA sequence. Because codons are comprised of three letters from the four-letter alphabet that makes up a DNA sequence, there are 64 possible codons [1]. Of the 64 possible codons, there are one start codon and three stop codons, and the remainder of the codons correspond to one of the twenty possible amino acids of a protein [1]. The relationship between DNA sequences, genes, intergenic spaces, exons, introns, and codons is illustrated in Figure 1.

Some exons within the protein-coding regions of DNA sequences of eukaryotes tend to exhibit a period-three pattern [2, 3, 4, 5]. The period-three pattern of the exons can be exploited to predict gene locations and even predict specific exons within the genes of eukaryotic cells [2, 3, 4, 5].

Previous digital signal processing (DSP) methods for the identification of coding regions (exons) in DNA sequences include the application of the discrete Fourier transform (DFT) on overlapping windows [1, 3, 4] and the application of bandpass digital filters that are centered at $2\pi/3$ [2, 6]. The output of a bandpass digital filter centered at $2\pi/3$ can be thought of as one measure of the DNA spectral content at frequency $2\pi/3$. Digital filter methods are of interest because

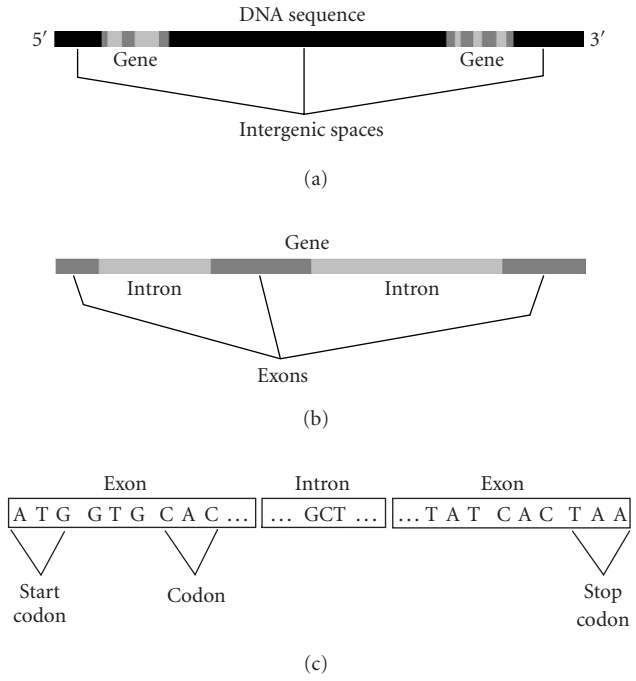

Figure 1: (a) An abstraction to illustrate the genes and intergenic spaces which comprise a DNA sequence. (b) An abstraction of a gene to illustrate the subdivision of a gene into exons and introns. (c) Various subsequences that comprise exons and introns in a gene (each three-letter grouping is a codon). The start codon is always ATG. However, one of the three possible stop codons is illustrated as (TAA).

they are significantly faster than the DFT method and they can be used to suppress more of the DNA background noise than it is possible by using the DFT method [2, 6].

DSP methods that only exploit period-three behaviour have many shortcomings. These methods are unable to reliably locate coding regions that do not have strong period-three characteristics. Methods based on hidden Markov models [7, 8, 9] provide superior results in these circumstances. The models used in these methods are also sufficiently accurate to account for exon and intron length distributions [10]. Alternatively, computational methods that exploit the heterogeneous statistical properties of DNA sequences to recursively segment homogeneous subsequences from their heterogeneous supersequences can be used for the identification of the borders between coding and noncoding regions [11, 12, 13]. The accuracy of these segmentation methods for coding region identification in DNA sequences surpasses the method presented in this paper and other DSP methods when applied to DNA sequences that do not have coding regions exhibiting a periodicity of three.

The method presented in this paper is an extension of DSP methods that exploit period-three behaviour. Previous DSP methods that exploit period-three behaviour do not entirely suppress the noncoding regions in the DNA spectrum at $2\pi/3$. As a result, a noncoding region may be incorrectly identified as a coding region. Also the methods presented in [2, 6] require four digital filter operations. In contrast, this paper presents a method that requires only one digital filter operation followed by a quadratic windowing operation. The quadratic window produces a signal that has almost zero energy in the noncoding regions. The proposed method can therefore improve the likelihood of correctly identifying coding regions over previous digital filtering methods. However, the accuracy of the proposed method suffers when dealing with coding regions that do not exhibit strong period-three behaviour. Also the methods presented in [7, 8, 9] are able to accurately model structures in genes, whereas the proposed method cannot. Despite these limitations, the method proposed in this paper can be used to generate one of the signals of a more complex gene finding method.

This paper is organized as follows. Section 2 reviews previous DSP methods for the identification of coding regions in DNA sequences. In particular, the DFT and digital filter methods are discussed. Section 3 presents a new computationally efficient one-step digital filter method for the identification of coding regions. Section 4 presents a new quadratic window operation that improves the suppression of noncoding regions from the DNA spectrum at frequency $2\pi/3$. In the example presented, noise suppression is improved by almost three orders of magnitude. Section 5 presents the conclusions of this research.

2. PREVIOUS DIGITAL SIGNAL PROCESSING METHODS FOR IDENTIFYING CODING REGIONS

Strands of DNA consist of four nucleotides (or bases), which are designated by the characters A, T, C, and G [1]. A character string composed of these four bases can be mapped to four signals [1]. The signal $u_A(n)$ takes the value of either 1 if A is present in the DNA sequence at index n, or 0 if A is absent at index n. For example, $u_A(n)$ for the DNA segment ATGCTGAA is 10000011. The signals $u_T(n)$, $u_C(n)$, and $u_G(n)$ can be obtained in a similar fashion.

The DFT of $u_A(n)$ over N samples is defined [14] as 3pt

$$U_A(k) = \sum_{n=0}^{N-1} u_A(n) e^{-j2\pi kn/N}, \quad 0 \le k \le N-1. \tag{1}$$

In a similar fashion, the DFT of $u_T(n)$, $u_C(n)$, and $u_G(n)$ can be obtained. For many genes, period-three behaviour has been observed and is useful for identifying coding regions [2, 3, 4, 5]. Specifically, the $(k = N/3)$-DFT coefficient magnitude is often significantly larger than the surrounding DFT coefficient magnitudes and corresponds to a coding region within the gene [1, 3, 4]. This effect varies and can be quite pronounced or quite weak, depending upon the gene [2].

A figure that can be used to measure the total spectral content of a DNA character string at frequency k is defined as [1, 4, 15]

$$S_{A+C+T+G}(k) = (U_A(k))^2 + (U_T(k))^2 + (U_C(k))^2 + (U_G(k))^2. \tag{2}$$

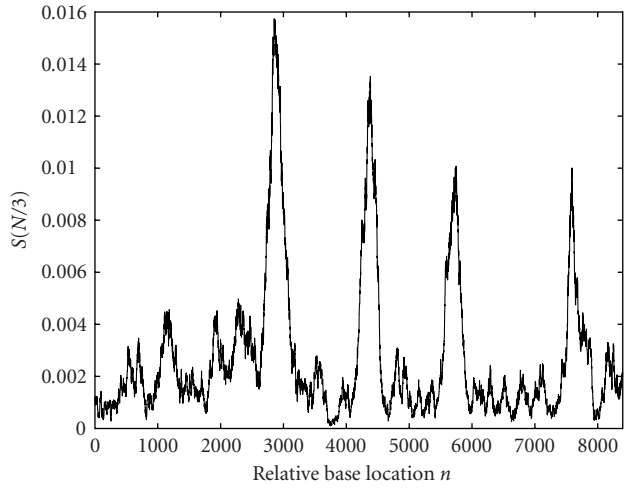

FIGURE 2: The signal $S_{A+C+T+G}(N/3)$ for gene F56F11.4 in the *C-elegans* chromosome III ($N = 351$).

The subscript of $S_{A+C+T+G}(k)$ indicates that all four nucleotide signals are considered. Corresponding to the previously described period-three behaviour, the value of $S_{A+C+T+G}(k)$ is large at $k = N/3$ when a coding region is present. The progression of $S_{A+C+T+G}(N/3)$ can be plotted by evaluating $S_{A+C+T+G}(N/3)$ over a window of N samples, sliding the window by one or more sample, and recalculating $S_{A+C+T+G}(N/3)$ [1]. This process can be carried out over the entire DNA sequence. As an example, consider the gene F56F11.4 in the *C-elegans* chromosome III. The value of $S_{A+C+T+G}(N/3)$ using $N = 351$ is plotted over the base numbers 7021 to 15080 in Figure 2.

The four dominant peaks in Figure 2 clearly indicate coding regions. However, a fifth coding region is present from 929 to 1135 but its small peak is obscured by $1/f$ DNA background noise. (The work presented in [15, 16, 17] observes the presence of $1/f$ background noise in DNA sequences.)

The DFT method for the identification of coding regions can be interpreted as a bandpass digital filter operation followed by a decimation operation [2]. The bandpass digital filter associated with the DFT method is centered at frequency $2\pi/3$ and has a minimum stopband attenuation of only 13 dB. High frequency selective bandpass digital filters for the identification of coding regions can be used instead of the DFT and have been presented in [2, 6] by Vaidyanathan and Yoon. The digital filter presented in [6] is a second-order antinotch filter. The digital filter presented in [2] is an eleventh-order bandpass digital filter with a minimum stopband attenuation of 60 dB.

The digital filter method for the identification of coding regions does not require the use of a sliding window [2, 6]. Instead, the signals $u_A(n)$, $u_C(n)$, $u_T(n)$, and $u_G(n)$ are individually processed using the same digital filter to produce the signals $y_A(n)$, $y_C(n)$, $y_T(n)$, and $y_G(n)$. A pseudomeasure of the total spectral content of a DNA sequence at frequency $2\pi/3$, $y_{A+C+T+G}(n)$, is given by [2, 6]

$$y_{A+C+T+G}(n) = |y_A(n)|^2 + |y_C(n)|^2 + |y_T(n)|^2 + |y_G(n)|^2. \quad (3)$$

The signal $y_{A+C+T+G}(n)$ produces large values in coding regions that exhibit strong period-three behaviour [2, 6] and is therefore an indicator for coding regions.

The digital filter method is much faster than the DFT method. For example, processing gene F56F11.4 in the *C-elegans* chromosome III using the DFT method requires 264 seconds on a 400 MHz Pentium II computer. In contrast, the digital filter method presented in [2] requires only 0.36 seconds, which is 733 times faster than the DFT method.

3. GENE PREDICTION USING A SINGLE DIGITAL FILTER

The methods presented by Vaidyanathan and Yoon in [2, 6] require a digital filtering operation for each of the four $u_A(n)$, $u_C(n)$, $u_T(n)$, and $u_G(n)$ signals for a total of four separate filtering operations. We now introduce a method that only requires one application of a digital filtering operation by filtering a single signal composed of $u_T(n)$ and $u_G(n)$. This new approach also removes much more of the DNA background noise than it is possible by using the methods presented in [2, 6]. In the following two sections, the optimization problem for creating this new signal is described and solved for a specific example.

3.1. Optimized signal construction

The number of digital filter operations can be reduced from four to one with the creation of a new signal that encapsulates the entire DNA sequence

$$u_{A+C+T+G}(n) = au_A(n) + cu_C(n) + tu_T(n) + gu_G(n), \quad (4)$$

where a, c, t, and g are real-valued parameters. Strand symmetry [18, 19, 20] can be exploited to further reduce the complexity of (4) to the sum of two terms. A long DNA sequence can be approximated using a two-symbol representation, where one symbol is either A or T and the other symbol is either C or G. In this case, the signal becomes

$$u_{T+G}(n) = tu_T(n) + gu_G(n). \quad (5)$$

Strand symmetry may not hold for shorter DNA sequences (on the order of 100 bases) and therefore strand symmetry should be verified before using (5) on short sequences. Section 3.2 compares the use of (4) and (5) for a test DNA sequence.

An optimization-based approach can be used to select the values of t and g (or a, c, t, and g if the strand symmetry is not used). A digital filter for gene prediction is first obtained from either the literature or from a suitable filter design method (this paper uses the digital filter presented in [2]). This digital filter is used in the optimization process to produce $v_{T+G}(n)$ from $u_{T+G}(n)$. A DNA sequence is selected where all of the coding regions are known. A pseudomeasure

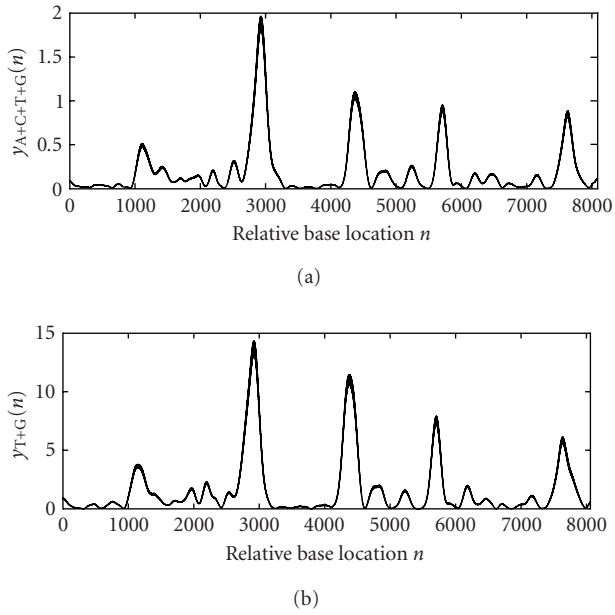

FIGURE 3: The signals $y_{T+G}(n)$ and $y_{A+C+T+G}(n)$ for gene F56F11.4 in the *C-elegans* chromosome III using the proposed single digital filter method.

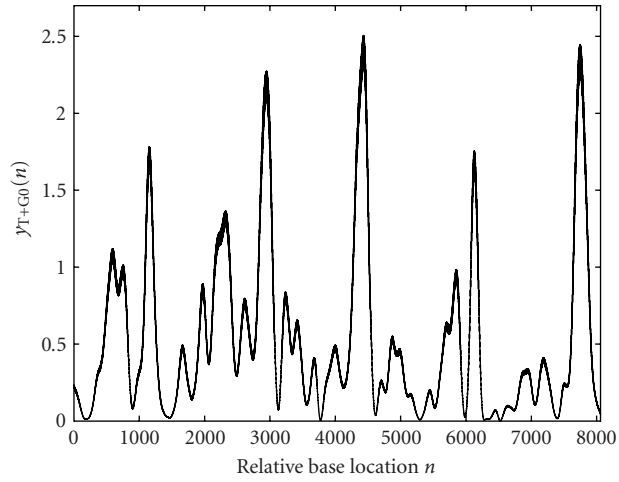

FIGURE 4: The signal $y_{T+G0}(n)$ for gene F56F11.4 in the *C-elegans* chromosome III.

of the total spectral content of a DNA sequence at $2\pi/3$ is given by

$$y_{T+G}(n) = v_{T+G}^2(n). \tag{6}$$

The ratio of $y_{T+G}^2(n)$ accumulated over all of the coding regions to $y_{T+G}^2(n)$ accumulated over all of the noncoding regions is maximized by choosing the t and g parameters:

$$\text{Maximize } \frac{\sum_{n_0 \in [\text{coding region}]} y_{T+G}^2(n_0)}{\sum_{n_1 \in [\text{noncoding region}]} y_{T+G}^2(n_1)}. \tag{7}$$

3.2. Applying the signal optimization

As an example, consider the use of the digital filter presented in [2] and the chromosome XVI of *S. cerevisiae* dataset. The quasi-Newton optimization method [21] is used to solve the above optimization problem for a two-symbol signal and for a four-symbol signal. The method proposed in this section is then used to process gene F56F11.4 in the *C-elegans* chromosome III over the base numbers 7021 to 15080 (see Figure 3). Figure 3 demonstrates that $y_{T+G}(n)$ and $y_{A+C+T+G}(n)$ are very similar due to the strand symmetry. The use of $y_{T+G}(n)$ is preferred because of its simplicity.

All five exons in Figure 3 are clearly visible in both $y_{T+G}(n)$ and $y_{A+C+T+G}(n)$. The remaining peaks do not have sufficient magnitude to obscure any of the coding regions. The total energy of $y_{T+G}(n)$ in the noncoding regions is defined as $\sum_{n \in [\text{noncoding region}]} {y_{T+G}}^2(n)$. This is a useful performance measure to gauge the effectiveness of a DSP gene prediction method for the suppression of the noncoding regions in $y_{T+G}(n)$. The total energy of $y_{T+G}(n)$ using the single digital filter method is 56.6. In contrast, the total energy of $y_{T+G}(n)$ in the noncoding regions using the multiple digital filter method as presented in [2] is 273.7, which is almost five times larger than the proposed single digital filter method. Clearly in this example, the proposed method improves the likelihood of correctly identifying the coding regions by reducing the total energy of $y_{T+G}(n)$ in the noncoding regions.

The initial coding region for gene F56F11.4 in the *C-elegans* chromosome III has a weak period-three characteristic, which is evident in Figures 2 and 3. In Figure 2, the initial coding region is obscured by noise. Optimizing the parameters t and g in $u_{T+G}(n)$ over a training sequence consisting of initial, internal, and terminal coding regions can be used to suppress a significant portion of this noise (see Figure 3). However, the relative height of the peak in $y_{T+G}(n)$ associated with the initial coding region is almost unchanged.

Our experiments indicate that the method proposed in this paper cannot be used to increase the relative height of the peaks in $y_{T+G}(n)$ associated with coding regions without also increasing the energy in the noncoding regions. We have attempted to optimize a new signal, $u_{T+G0}(n)$, that, when filtered, produces larger peaks for initial coding regions. A training dataset composed only of initial coding regions in XVI of *S. cerevisiae* was used to obtain t and g. Figure 4 shows $y_{T+G0}(n)$ for gene F56F11.4 in the *C-elegans* chromosome III. The relative height of the peak associated with the initial coding region shown in Figure 4 has increased but at the expense of a significant increase in the signal energy in the noncoding regions. Consequently, the use of $u_{T+G0}(n)$ has little practical benefit because the increased signal energy in the noncoding regions decrease the likelihood of correctly identifying the coding regions. Similar results can be obtained if t and g are optimized only for internal coding regions or only for terminal coding regions. In contrast, methods based on hidden Markov models [7, 8, 9] use sufficiently accurate models to predict the location of coding regions that do not have strong period-three characteristics.

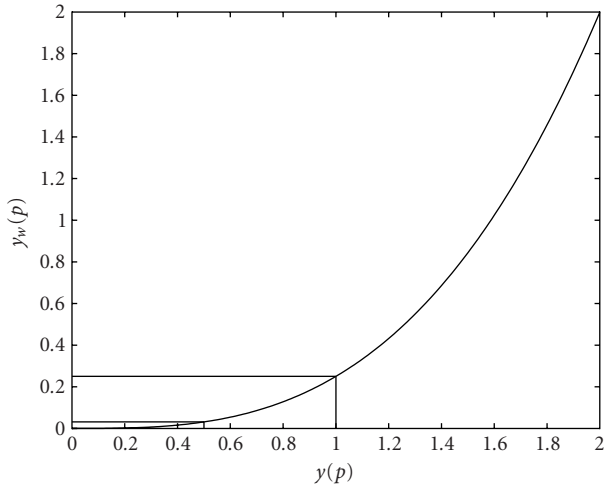

Figure 5: The quadratic window nonlinearity plotted for Maxvalue = 2.

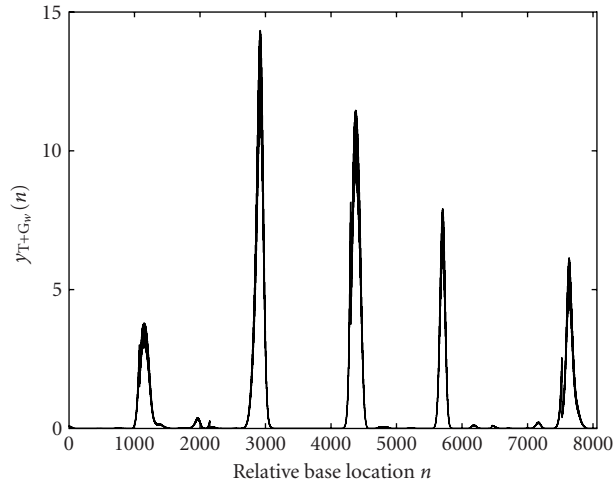

Figure 6: The signal $y_{T+G_w}(n)$ for gene F56F11.4 in the *C-elegans* chromosome III using the quadratic window (8).

4. A QUADRATIC WINDOW OPERATION TO SUPPRESS NONCODING REGIONS

The single digital filter method for the identification of coding regions does not always suppress all of the peaks found in the noncoding regions of $y_{T+G}(n)$ (see Figure 3). Consequently, the noncoding regions may obscure the coding regions in some datasets. To reduce uncertainty in the identification of coding regions, a new quadratic windowing operation is now introduced that can be used to effectively suppress the noncoding regions while preserving the coding regions. This quadratic windowing operation is performed after the single digital filter operation on $y_{T+G}(n)$.

The maximum value of $y_{T+G}(n)$ in a coding region is almost always greater than the maximum value of $y_{T+G}(n)$ in a noncoding region although the difference in magnitude between the two may be small. It is desirable to exaggerate the difference in magnitude between the coding and noncoding regions so that the coding regions can be more easily identified. To this end, a window of M samples is processed using the following operation:

$$y_w(p) = \left(\frac{y_{T+G}(p)}{\text{Maxvalue}}\right)^2 \cdot y_{T+G}(p), \quad 1 \le p \le M, \tag{8}$$

where p is the window sample index, M is the number of samples in the window, $y_w(p)$ is the pth windowed sample value, and Maxvalue is the largest value of $y_{T+G}(p)$ in the window.

The quadratic windowing operation defined in (8) multiplies $y_{T+G}(p)$ by a value that approaches zero in a quadratic fashion as $y_{T+G}(p)$ approaches zero. Noncoding regions in the window that have sample values less than Maxvalue are effectively suppressed. Consider a window of samples that has maximum sample value of 2. The quadratic window operation produces $y_w(p)$ values of 0.0313 and 0.25 for $y_{T+G}(p)$ values that equal 0.5 and 1, respectively, as shown in Figure 5.

To preserve the coding regions in $y_{T+G}(n)$, the size of the window should not contain more than one coding region. In this case, the sole coding region in the window is not suppressed because the value of the largest sample, which belongs to the coding region, is not changed when using (8). A DNA sequence, where all of the coding regions are known, can be used to select the window size. The window size is set to a value less than the minimum number of samples between adjacent coding regions and greater than the number of samples of the widest coding region.

After a window of M samples has been processed, the window is then moved M samples, which prevents the successive windowing operations from overlapping.

The quadratic windowing operation is now applied to the gene F56F11.4 in the *C-elegans* chromosome III over the base numbers 7021 to 15080. Figure 3 shows the original $y_{T+G}(n)$ signal obtained using the method discussed in Section 3.2. The quadratic window of (8) is used to obtain the signal $y_w(p)$, as shown in Figure 6. The window size is set to $M = 1100$ samples. The five coding regions (exons) dominate the signal $y_w(n)$. In the coding regions, the signal $y_w(n)$ has been suppressed to near-zero values, which improves the certainty of correctly identifying the coding regions.

Table 1 compares the suppression of the noncoding regions by comparing the total energy in these regions for the multiple digital filter gene prediction method presented in [2], the single digital filter method presented in Section 3, and the single digital filter method followed by the quadratic window operation presented in this section. This numerical experiment used gene F56F11.4 in the *C-elegans* chromosome III over the base numbers 7021 to 15080.

The multiple digital filter method does not effectively minimize the total energy in the noncoding regions. The total energy in the noncoding regions for the multiple digital filter method is 720 times greater than the total energy in noncoding regions for the method proposed in this section and almost five times greater than the method presented in Section 3. As a result, a noncoding region may inadvertently

Table 1: A comparison of the performance between competing gene prediction methods.

Gene prediction method	Total energy in the noncoding regions
Single digital filter method followed by the quadratic window operation	0.38
Single digital filter method	56.6
Multiple digital filter method [2]	273.7

Table 2: A comparison of SNR values between competing gene prediction methods.

Gene	SNR (single digital filter method followed by the quadratic window operation)	SNR (multiple digital filter method [2])
F56F11.4	107	4
ZK250.9	225	18
ZK250.10	848	22
F54D8.1	64	11

be identified as a coding region when using the multiple digital filter method. In contrast, all five coding regions can easily be identified using the methods presented in this section.

The quadratic windowing method (single digital filter method followed by a quadratic window operation) is now compared in more depth with Vaidyanathan and Yoon's multiple digital filter method [2]. Table 2 compares the signal-to-noise ratio (SNR), see (9), for the following test genes: F56F11.4 in the *C-elegans* chromosome III, ZK250.9 and ZK250.10 in the *C-elegans* chromosome II, and F54D8.1 in the *C-elegans* chromosome III.

The SNR performance measure considers both the energy in the coding and noncoding regions. High SNR signals have low energy levels in the noncoding regions and high energy levels in the coding regions. For high SNR signals, the task of identifying coding regions is greatly simplified because the coding regions dominate over the noncoding regions

$$\mathrm{SNR} = \frac{\sum_{n_0 \in [\text{coding region}]} y_{\mathrm{T+G}}^2(n_0)}{\sum_{n_1 \in [\text{noncoding region}]} y_{\mathrm{T+G}}^2(n_1)}. \tag{9}$$

Table 2 shows that the multiple digital filter method consistently generates significant lower SNR signals than does the method proposed in this paper. Consequently, the task of identifying coding regions in signals generated by the multiple digital filter method is more problematic.

5. CONCLUSION

Methods for the identification of coding regions that solely rely on digital filters [2, 6] are unable to significantly attenuate the noncoding regions in $y_{\mathrm{T+G}}(n)$. Consequently, a noncoding region may inadvertently be identified as a coding region. This paper introduced a new DSP technique (a single digital filter operation followed by a quadratic window operation) that can be used to suppress nearly all of the noncoding regions in $y_{\mathrm{T+G}}(n)$. This paper demonstrated that the total energy in the noncoding regions of $y_{\mathrm{T+G}}(n)$ can be reduced by a factor of 720 compared to the previous digital filter techniques for gene F56F11.4 in the *C-elegans* chromosome III. As a result, the proposed method can improve the likelihood of correctly identifying coding regions.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their comments and valuable suggestions which helped in improving this paper.

REFERENCES

[1] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[2] P. P. Vaidyanathan and B.-J. Yoon, "Digital filters for gene prediction applications," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, pp. 306–310, Pacific Grove, Calif, USA, November 2002.

[3] D. Anastassiou, "DSP in genomics," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1053–1056, Salt Lake City, Utah, USA, May 2001.

[4] 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.

[5] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Res.*, vol. 10, no. 17, pp. 5303–5318, 1982.

[6] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using allpass-based filters," in *Workshop on Genomic Signal Processing and Statistics*, Raleigh, NC, USA, October 2002.

[7] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in DNA with a hidden Markov model," *J. Comput Biol.*, vol. 4, no. 2, pp. 127–141, 1997.

[8] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, "A generalized hidden Markov model for the recognition of human genes in DNA," in *Proc. of the 4th International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, Calif, USA, 1996.

[9] A. Krogh, I. S. Mian, and D. Haussler, "A hidden Markov model that finds genes in E. coli DNA," *Nucleic Acids Res.*, vol. 22, no. 22, pp. 4768–4778, 1994.

[10] C. B. Burge and S. Karlin, "Finding the genes in genomic DNA," *Curr. Opin. Struct. Biol.*, vol. 8, no. 3, pp. 346–354, 1998.

[11] P. D. Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871–888, 2003.

[12] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse, "Applications of recursive segmentation to the analysis of DNA sequences," *Computers & Chemistry*, vol. 26, no. 5, pp. 491–510, 2002.

[13] W. Li, G. Stolovitzky, P. Bernaola-Galvan, and J. L. Oliver, "Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes," *Genome Research*, vol. 8, no. 9, pp. 916–928, 1998.

[14] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[15] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," *Phys. Rev. Lett.*, vol. 68, no. 25, pp. 3805–3808, 1992.
[16] W. Li, "The study of correlation structures of DNA sequences: a critical review," *Computers & Chemistry*, vol. 21, no. 4, pp. 257–271, 1997.
[17] W. Li and K. Kaneko, "Long-range correlation and partial $1/f^{\alpha}$ spectrum in a non-coding DNA sequence," *Europhys. Lett.*, vol. 17, no. 7, pp. 655–660, 1992.
[18] D. R. Forsdyke and J. R. Mortimer, "Chargaff's legacy," *Gene*, vol. 261, no. 1, pp. 127–137, 2000.
[19] W. Li, "The study of correlation structures of DNA sequences: a critical review," *Computers & Chemistry*, vol. 21, no. 4, pp. 257–272, 1997.
[20] J. W. Fickett, D. C. Torney, and D. R. Wolf, "Base compositional structure of genomes," *Genomics*, vol. 13, no. 4, pp. 1056–1064, 1992.
[21] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, Pa, USA, 1996.

Trevor W. Fox received his B.S. and Ph.D. degrees in electrical engineering from the University of Calgary in 1999 and 2002, respectively. Currently, he is working at the Intelligent Engines in Calgary, Canada. His main research interests include digital filter design, reconfigurable digital signal processing, and genomic signal processing.

Alex Carreira received his B.S. and M.S. degrees in electrical engineering from the University of Calgary, Canada, in 1999 and 2003, respectively. His main research interests are digital signal processing with programmable logic devices, configurable and reconfigurable computing, and rapid prototyping of systems for programmable logic devices.