A Signal Boosting Technique for Gene Prediction

Teddy Surya Gunawan[1,2], Eliathamby Ambikairajah[1], Julien Epps[1]

[1]School of Electrical Engineering and Telecommunications
The University of New South Wales
Sydney, NSW 2052, Australia

[2]Department of Electrical and Computer Engineering
International Islamic University Malaysia
Gombak, Selangor 53100, Malaysia

tsgunawan@ee.unsw.edu.au, ambi@ee.unsw.edu.au , j.epps@unsw.edu.au

***Abstract*—This paper presents the signal boosting technique for gene and exon identification of a DNA sequence. Newly proposed signal boosting technique is used to enhance the coding region and improve the likelihood of correctly identifying the coding region. This new method is compared with other genomic signal processing techniques, such as DFT, IIR anti-notch filter, and multistage filters. Results show that the ratio of coding to non-coding energy was enhanced by almost twice. Furthermore, the proposed method outperforms other methods in terms of sensitivity, specificity, and prediction rate, when testing with 78 sequences from HMR195 database.**

***Keywords*—genomic signal processing, gene prediction, digital filter, signal boosting, DNA**

I. INTRODUCTION

The gene identification problem, which identifies the protein-coding regions (exons) in DNA sequences through computational means, is of great importance nowadays. Figure 1 shows that a DNA sequence can be divided into genes and intergenic spaces. A gene can be divided into two sub-regions called coding regions (exons) and non-coding regions (introns). It is well known that protein-coding regions of DNA sequences tend to exhibit a period-3 pattern because of the codon structure involved in the translation of base sequences into amino acids [1-4]. Many researchers have regarded the period-3 property to be a good indicator of gene location.

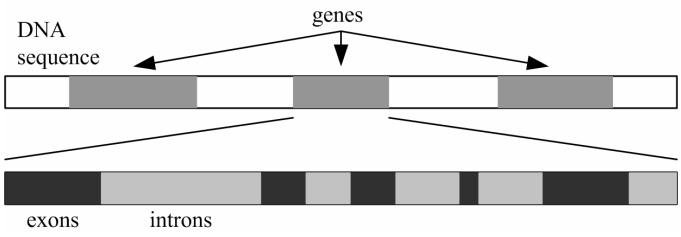

Figure 1. Various regions in a DNA molecule

Previous digital signal processing methods for the identification of exons in DNA sequences include the sliding window discrete Fourier transform (DFT) [1, 2], the anti-notch IIR filter, and the multi-stage bandpass filter centered at $2\pi/3$ [4]. The multi-stage bandpass filter is of particular interest because it can be used to suppress the $1/f$ noise exhibited by DNA sequences of many organisms in general [4, 5].

An extension of signal processing methods that exploit period-3 behavior is presented. Previous signal processing methods that utilize period-3 characteristics do not entirely suppress the non-coding regions in the DNA spectrum at $2\pi/3$. Consequently, a non-coding region may be incorrectly identified as a coding region. In this paper, we propose a signal boosting technique to enhance coding region detection.

A signal boosting technique has been successfully utilized for speech enhancement [6]. The same principle of signal boosting technique is applied here to the coding region for gene identification. The coding region is enhanced while the non-coding region is kept. The proposed method can therefore improve the likelihood of correctly identifying coding regions over previous gene identification methods, such as DFT techniques, IIR anti-notch filter, and multistage filters. In this paper, a multi layer neural network classifier with two hidden layers was used to evaluate the performance of the proposed algorithm.

The rest of the paper is organized as follows. Section II reviews previous signal processing methods for the identification of coding regions in DNA sequence. In particular, the DFT and digital filter methods are discussed. Section III presents a new technique to boost the coding region. Section IV provides the performance evaluation of various techniques on HMR195 database, while section V concludes the paper.

II. GENOMIC SIGNAL PROCESSING TECHNIQUES

It has been found that the base sequence in the coding region have a strong period-3 component [3]. This phenomenon could be due to nonuniform codon usage, i.e. even though there are several codons which could code a given amino acid, they are not used with uniform probability, and this creates a codon bias. In order to apply digital signal processing techniques, the character sequences of DNA should be first converted into a numerical representation, for example four binary indicator numeric sequences. The simplest and most popular mapping of a DNA sequence is known as the Voss representation [5]. For example, for a DNA sequence $x[n]$=CGATGACGAA, the binary indicator sequence for each base type, $x_\ell[n], \forall \ell \in \{A, C, G, T\}$, would be

$$x_A[n] = \{0,0,1,0,0,1,0,0,1,1\},\quad x_C[n] = \{1,0,0,0,0,0,1,0,0,0\}$$
$$x_G[n] = \{0,1,0,0,1,0,0,1,0,0\},\quad x_T[n] = \{0,0,0,1,0,0,0,0,0,0\}$$
$$x_A[n] + x_C[n] + x_G[n] + x_T[n] = 1$$

where n represents the base index. From a biological perspective, the Voss representation characterizes the

frequency of occurrence of each individual base in the DNA sequence. Other popular DNA representations for genomic signal processing can be found in [7]. We now review existing methods for the detection of gene regions:

A. DFT Technique

The DFT of a DNA sequence $x_\ell[n]$ over N samples is defined as

$$X_\ell[k] = \sum_{n=0}^{N-1} x_\ell[n] e^{-j\frac{2\pi k}{N}} \tag{1}$$

where $k = 0,\ldots,N-1$ and $\ell \in \{A,C,G,T\}$. The DFTs $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$ for the above indicator sequences can thus be calculated using Equation (1). The total spectral content of a DNA sequence can be calculated as follows [1, 2, 5]:

$$S[k] = |X_A[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 + |X_T[k]|^2 \tag{2}$$

The period-3 property of a DNA sequence implies that the spectral coefficients $S[N/3]$ are large. While this is generally true, the strength of the peak depends on the gene, and is thus sometimes is very pronounced and sometimes quite weak. The window length N should be sufficiently large [1], e.g. N = 351, so that the periodicity effect dominates the background $1/f$ spectrum which makes its strong presence in DNA sequences [5].

B. IIR Anti-Notch Filter

An IIR anti-notch filter $H(z)$ can be used for gene prediction [4]. The IIR filters can be obtained from a second order allpass filter with poles at $R \cdot e^{\pm j\theta}$ as follows,

$$A(z) = \frac{R^2 - 2R\cos\theta z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \tag{3}$$

where $R^2 < 1$, e.g. $R = 0.992$, for stability. The IIR anti-notch filter is then can be calculated as follows [4]:

$$H(z) = \frac{1 - A(z)}{2}. \tag{4}$$

C. Multistage filters

Improvements to the IIR anti-notch filter are proposed in [4], including using multistage filters with increased stopband attenuation compared with the IIR anti-notch filter, with only a slight increase in computation. Such filters are essential in order to suppress the background $1/f$ noise due to long-range correlation between based pairs. In [4], the multistage filters is designed as follows,

$$H(z) = H_1(z^3) H_2(z) \tag{5}$$

where $H_1(z)$ is a third order elliptic filter and $H_2(z) = (1 - z^{-1})^2$.

III. Signal Boosting Method

We propose a new method to enhance the coding region in DNA sequences which relatively suppresses the non-coding region, thus improving the likelihood of correct prediction of the coding region of a DNA sequence. To achieve this objective, the coding region is treated as the "signal" and non-coding region is treated as the "noise".

The DFT of a DNA sequence $x_\ell[n] \times w[n]$ over N samples is calculated as in Equation (1). The Bartlett window $w[n]$ was utilized, as it removes the extraneous peaks introduced by the abrupt edges of the rectangular window [2]. The total spectral content of a DNA sequence is then calculated as shown in Equation (2), and its value at $S[N/3]$ is stored, i.e. $\psi(m) = S_m[N/3]$ $\forall m$, $m = 1\ldots M$, while M is the length of the DNA sequence in base pairs.

The objective is now to find a gain function, $\Gamma(m)$, that weights the spectral content at $k = N/3$, $\psi(m)$, based on the coding (signal) to non-coding (noise) ratio. The SNR can be calculated by using the ratio of a short-term average signal energy, $P(m)$, and an estimate of the noise floor level, $Q(m)$. The short-term average signal energy is calculated as

$$P(m) = (1-\alpha)P(m-1) + \alpha \cdot \psi(m) \tag{6}$$

where α is a small positive constant, e.g. $\alpha = 0.2$, controlling the sensitivity of the algorithm to changes in signal energy and acting as a smoothing factor. The slowly varying noise floor estimate, $Q(m)$, is calculated as

$$Q(m) = \begin{cases} (1+\beta)Q(m-1), & Q(m-1) \le P(m) \\ P(m), & Q(m-1) > P(m) \end{cases} \tag{7}$$

where β is a small positive constant, e.g. $\beta = 0.0002$, controlling how fast the noise floor level estimate adapts to changes.

The boosted signal (coding region) $\hat{\psi}(m)$ is then calculated as follows:

$$\hat{\psi}(m) = \Gamma(m) \cdot \psi(m) = \frac{P(m)}{Q(m)} \psi(m) \tag{8}$$

To avoid over-amplification, we limit the gain $\Gamma(m)$ to be no greater than 10 dB.

Figure 3 shows the exon prediction results for gene F56F11.4 in the *C-elegans* chromosome III over the base numbers 7021 to 15080 which has five exons. Figure 3(a)-(c) shows the original $\psi(m)$ signal obtained using the DFT, anti-notch filter, and multistage filter techniques respectively. The

signal boosting technique is used to obtain the signal $\hat{\psi}(m)$ as shown in Figure 3(d). The five exons dominate the signal $\hat{\psi}(m)$ compared with the other methods. In the coding regions, the signal $\hat{\psi}(m)$ has been enhanced so that it will improve the certainty of correctly identifying the coding regions. For identifying the first exon in particular, which has the smallest amplitude compared to other exons, we can see clearly that the proposed method outperforms other methods. If a simple threshold is to be used to classify the DNA sequence, our proposed method will better separate the coding and non-coding regions. Table I compares the ratio of energy between coding and non-coding regions for various methods.

TABLE I. ENERGY RATIO OF VARIOUS METHODS

Gene prediction method	Energy ratio of coding and non-coding regions
DFT technique	0.744
IIR Anti-notch filter	0.767
Multistage filters	0.804
Signal boosting technique	*1.771*

IV. PERFORMANCE EVALUATION

In this section, the performance of various methods for gene prediction is evaluated. First, the performance metrics are explained. Then, the HMR195 dataset for training and testing and neural network classifier is described. Finally, the performance evaluation of the gene identification methods is discussed.

A. Evaluation Measures

To evaluate the performance of gene identification, we used prediction accuracy measures similar to [8], as shown in Figure 2. True positive (*TP*) is the number of coding nucleotides correctly predicted as coding. False negative (*FN*) is the number of coding nucleotides predicted as non-coding. True negative (*TN*) is the number of non-coding nucleotides correctly predicted as non-coding. False positive (*FP*) is the number of non-coding nucleotides predicted as coding. The sensitivity (*Sn*) provides a measure of the proportion of coding nucleotides that have been correctly predicted as coding. The specificity (*Sp*) provides the proportion of predicted coding nucleotides that are actually from the coding region. Finally, the precision (*P*) shows the recognition rate of the classifier.

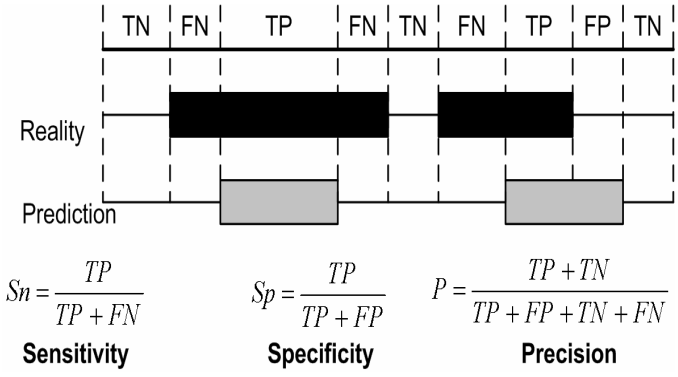

Figure 2. Nucleotide-level measures of prediction accuracy

B. Data sets

The HMR195 dataset [9] contains 195 mammalian sequences with exactly one complete either single-exon or multi-exon gene. The dataset was developed to evaluate different gene-finding programs. All sequences contain exactly one gene, which starts with the 'ATG' initial codon and ends with one of the stop codons, i.e. 'TAA', 'TAG', or 'TGA'. There are no in-frame stop codons in coding genes, and introns of multi-exons genes start with dinucleotides 'GT' and end with dinucleotide 'AG'. Sequences longer than 200,000 bp are not included in the set. In this dataset, the ratio of human:mouse:rat sequences is 103:82:10, with a mean length of 7096 bp. The set contains 43 single-exon genes, and 152 multi-exon genes. The proportion of coding regions in the sequences is 14% and the mean exon length is 208 bp.

In this paper, the HMR195 dataset was divided into 117 training set sequences (60%) and 78 testing set sequences (40%). The single and multi-exon sequences and human/mouse/rat sequences were evenly and randomly distributed into the training and testing sets. The training set had length 786338 bp, while the testing set had length 603400 bp.

C. Neural Network Classifier

The input to the classifier is the extracted features $\psi(m)$ using various methods, including DFT technique, IIR anti-notch filter, multistage filters, and the proposed signal boosting technique. The output of the classifier is coding or non-coding nucleotides. In this paper, a multilayer neural network trained using back-propagation algorithm was used as the classifier for all the gene prediction methods. The network was configured with two hidden layers and network size of 1-4-4-1 as this configuration provided good classification and efficient network training. All neurons in both hidden layers have tan-sigmoid transfer functions. The output neuron has a purely linear transfer function. The networks were trained using the 177 HMR195 training set sequences, as defined in previous section. Training was determined to be complete when the mean square error of the network fell below 0.001 of the training data. The resilient back-propagation algorithm was utilized to train the networks.

D. Results

The neural network classifier was evaluated using the 78 HMR195 testing set sequences. Table II shows the gene identification performance in terms of specificity, sensitivity, and precision. From Table II, we can see that the proposed signal boosting technique outperforms other methods. In particular, the specificity performance indicator for correctly identifying coding regions was higher than for any other method. Furthermore, the task of identifying coding regions in is simple relative to other methods, such as the multistage filters which proved more problematic than other methods.

TABLE II. PERFORMANCE COMPARISON OF VARIOUS METHODS

Gene prediction method	*Sp*	*Sn*	*P*
DFT technique	0.721	0.394	0.897
IIR Anti-notch filter	0.703	0.351	0.894
Multistage filters	0.673	0.266	0.885
Signal boosting technique	*0.725*	*0.471*	*0.911*

V. CONCLUSIONS

Signal processing methods for the identification of coding regions that solely rely on DFT techniques or digital filters are unable to significantly enhance the coding regions. Therefore, a non-coding region may inadvertently be identified as a coding region. This paper has introduced a new signal boosting technique that can be used to boost the coding region and can improve the likelihood of correctly identifying coding regions. Evaluation of various gene identification methods using a multi layer neural network classifier on the HMR195 database revealed that the proposed method outperforms other methods in terms of specificity, sensitivity, and prediction accuracy.

ACKNOWLEDGMENT

This research is fully supported by the University of New South Wales, Australia, Faculty Research Grant, 2007 for genomic signal processing research.

REFERENCES

[1] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 8-20, 2001.

[2] S. Datta, A. Asif, and H. Wang, "Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics," in IEEE 6th International Symposium on Multimedia Software Engineering, pp. 160-163, 2004.

[3] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Physica A: Statistical and Theoretical Physics*, vol. 249, pp. 511-516, 1998.

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

[5] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, pp. 3805, 1992.

[6] T. S. Gunawan, *Audio compression and speech enhancement using temporal masking models*, PhD Thesis, Sydney, The University of New South Wales, 2007.

[7] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representation for period-3 based exon prediction," in 5th International Workshop on Genomic Signal Processing and Statistics, 2007.

[8] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, pp. 353-367, 1996.

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

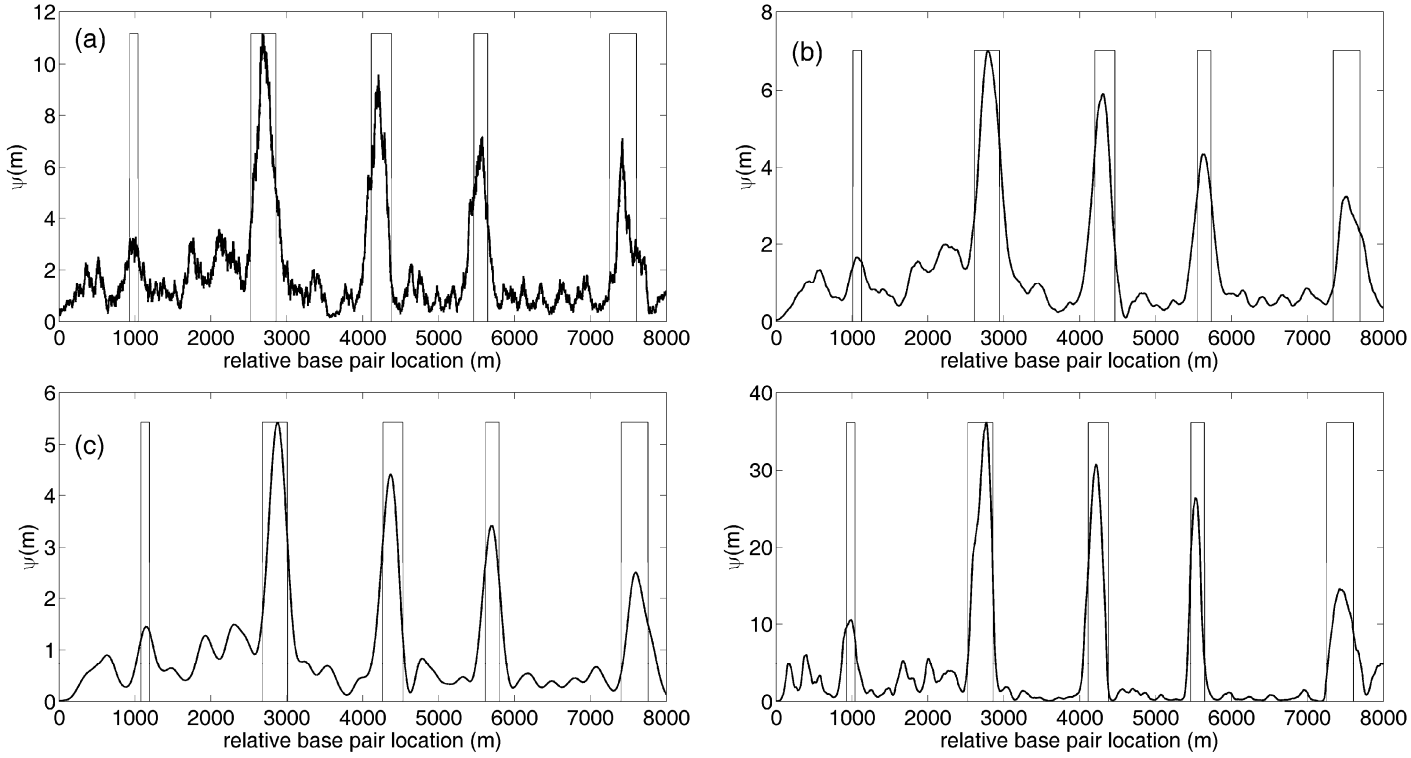

Figure 3. Plot of the total spectrum at $k = N/3$ $\psi[m]$ for various methods, (a) DFT technique, (b) IIR anti-notch filter, (c) Multistage filters, and (d) Proposed method (signal boosting technique $\hat{\psi}(m)$)
Note that, the faint binary signal in the background indicates the true locations of coding regions (high) and non-coding regions (low)