# Digital Signal Processing Techniques for Gene Finding in Eukaryotes

Mahmood Akhtar[1, 2], Eliathamby Ambikairajah[2], and Julien Epps[2]

[1] Centre for Health Informatics, University of New South Wales,
45 Beach St, Coogee NSW 2034, Australia
[2] School of EE&T, University of New South Wales,
Sydney 2052, Australia
mahmood@unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

**Abstract.** In this paper, we investigate the effects of window shape and length on a DFT-based method for gene and exon prediction in eukaryotes. We then propose a new gene finding method which combines the selected time-domain and frequency-domain methods, by employing the most effective DNA symbolic-to-numeric representation examined to date in conjunction with suitable window shape and length parameters and a signal boosting technique. It is shown herein that the new method outperforms major existing approaches. By comparison with the existing methods, the proposed method reveals relative improvements of 15.1% to 55.9% over different methods in terms of prediction accuracy of exonic nucleotides at a 5% false positive rate using the GENSCAN test set.

**Keywords:** DNA, periodicity, discrete Fourier transforms, signal boosting.

## 1 Introduction

It is well-known that deoxyribonucleic acid (DNA) is the material of heredity in most living organisms, and consists of genic and intergenic regions. Eukaryotes differ from prokaryotes in that their genes are further divided into relatively small protein coding segments known as *exons*, interrupted by non-coding spacers known as *introns*. Eukaryotic gene finding is a significant open problem in the field of DNA sequence analysis. The problem is difficult mainly due to the noncontiguous and non-continuous nature of genes. Furthermore, often the intergenic and intronic regions make up most of the genome. For example, in human genome the exonic fraction is as low as 2%.

The conversion of DNA nucleotide symbols (i.e., A, C, G, and T) into discrete numerical values enables novel and useful DSP-based applications for the solution of different sequence analysis related problems such as gene finding [1]. In recent years, a number of schemes have been introduced to map DNA nucleotides into numeric values [2]. The binary or Voss representation [3] is a popular scheme, which maps the nucleotides A, C, G, and T into the four binary indicator sequences $x_A[n]$, $x_C[n]$, $x_G[n]$, and $x_G[n]$ showing the presence (e.g. 1) or absence (e.g. 0) of the respective nucleotides. For example, for a DNA sequence $x[n]$ = AGTTCTACCGAGC…, the binary

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 144 – 152, 2008.

indicator sequences for each base type would resemble: $x_A[n]$ = {1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, ...}, $x_C[n]$ = {0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, ...}, $x_G[n]$ = {0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, ...}, $x_T[n]$ = {0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, …}, where $x_A[n] + x_C[n] + x_G[n] + x_T[n] = 1$, and $n$ represents the base index. In exons, the occurrence of identical nucleotides in identical codon (e.g. triplet encoding protein) positions is the basis for a periodicity of three interpretation in these regions [4]. The period-3 behaviour of exons has been widely used to identify these regions using DSP-based methods such as the discrete Fourier transform (DFT) [1], [5]–[7], time-domain algorithms [8], etc.

The discrete Fourier transform (DFT), the most commonly used method for spectrum analysis of a finite-length numerical sequence $x[n]$ of length $N$, is defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi nk}{N}}, \quad 0 \le k \le N-1 \tag{1}$$

Equation (1) can be used to calculate DFTs for four binary indicator sequences (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$). The periodicity of 3 in exon regions of a DNA sequence suggests that the DFT coefficient corresponding to $k = N/3$ (where $N$ is chosen to be a multiple of 3) in each DFT sequence should be large [1]. Various DFT based spectral measures exploiting the period-3 behaviour of exons for the identification of these regions have been proposed. The spectral content (SC) measure [5] combines the individual DFTs (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$) to obtain a total Fourier magnitude spectrum of the DNA sequence, as follows:

$$SC[k] = \sum_{m} |X[k]|^2, \quad m \in \{A, C, G, T\} \tag{2}$$

The spectral rotation (SR) measure [6] rotates four DFT vectors $X_A[k]$, $X_C[k]$, $X_G[k]$ and $X_T[k]$ clockwise, each by an angle equivalent to the average phase angle value in coding regions, to make all of them 'point' in the same direction. The SR measure also divides each term by the corresponding phase angle deviations to give more weight to exonic distributions. The feature

$$SR[k] = \left| \sum_{m} \frac{e^{-j\mu_m}}{\sigma_m} . X_m[k] \right|^2, \quad m = \{A, C, G, T\} \tag{3}$$

has been used for the detection of exons. The SR measure has been shown [6] to give better performance than the SC (2) measure at a 10% false positive gene detection rate. The paired and weighted spectral rotation (PWSR) measure [7] incorporates a statistical property of eukaryotic sequences, according to which introns are rich in nucleotides '*A*' and '*T*' whereas exons are rich in nucleotides '*C*' and '*G*'. This information leads to an alternative to the well-known period-3 behavior of exons. In this method, the DNA sequences are first converted into two binary indicators (i.e., $x_{A\text{-}T}[n]$ and $x_{C\text{-}G}[n]$). Using training data from DNA sequences of the same organism, the means $\mu_m$ and standard deviations $\sigma_m$ of the distributions of DFT phase angle averaged over coding regions (i.e., one phase angle value for one coding region) are calculated. Weights $w_m$ based on the frequency of occurrence of nucleotides '*A* or *T*' and '*C* or *G*' in coding regions of the training data are also calculated. The expression given in (4) can then be used as a feature, along one direction of the DNA sequence:

$$PWSR_{l}[k] = \left| \frac{e^{-j\mu_{A-T}}}{\sigma_{A-T}} . w_{A-T} . X_{A-T}[k] + \frac{e^{-j\mu_{C-G}}}{\sigma_{C-G}} . w_{C-G} . X_{C-G}[k] \right|^2 \quad (4)$$

where $l$ = forward ($F$) and reverse ($R$) directions of DNA sequence, and $X_m[k]$ ($m = A$-$T$, $C$-$G$) are the sliding DFT windows of two indicator sequences. The expression in (4) has been used in the reverse direction of the same DNA sequence (i.e., due to paired indicators, DFT in reverse direction of the same DNA strand is equivalent to DFT on its complementary strand). The PWSR measure is the sum of forward and reverse measures:

$$PWSR[k] = PWSR_{F}[k] + PWSR_{R}[k] \quad (5)$$

The PWSR measure has been shown [7] to give better performance than the SC (2) and SR (3) measures at 10%, 20%, and 30% false positive nucleotide detection rates.

Time-domain algorithms [8] use prefiltering of the four binary indicator sequences (i.e., pass them through a second order resonant filter with centre frequency of $2\pi/3$) to remove spectral components at $2\pi k/3$, $k \in \Im$, $k \neq 1$, which arises from the application of correlation-based approaches to a binary indicator sequence at a base-domain lag of 3. The resultant non-binary numerical sequences, emphasizing the period-3 behavior of DNA sequences but de-emphasizing the other components, are input to the average magnitude difference function (AMDF) or time domain periodogram (TDP) algorithms. The AMDF for a discrete signal $x[n]$ as a function of the period $k$, is defined as:

$$AMDF[k] = \frac{1}{N} \sum_{n=1}^{N} |x[n] - x[n-k]| \quad (6)$$

where $N$ is the window length. Practically, the AMDF will produce a deep null if significant correlation exists at period $k = 3$.

Despite the existence of these approaches and also data-driven approaches, the accuracy of gene prediction still needs to be improved. We address this shortcoming herein, investigating the effects of window shape and length parameters and proposing a new DSP-based gene finding method. The remainder of this paper is organized as follows. In Section 2, effects of window shape and length on the DFT-based gene and exon prediction are investigated. A new digital signal processing-based gene finding method is then proposed in Section 3. Finally, the proposed and existing methods are compared in Section 4.

## 2 Effects of Window Shape and Length

Most existing signal processing-based gene prediction methods use sliding window attributes to maintain a reasonable base-domain resolution. The DFT-based SC measure [5], SR measure [6], are well-known examples. All these authors use a conventional rectangular shaped window of length of 351 bp, with the argument that the data window should be reasonably long or few hundred base pairs long [5]-[6]. Datta and Asif [9] claim an improved DFT based period-3 detection using the Bartlett data window compared with the rectangular window of same size (i.e., $N = 351$) . How does window shape and/or window size affects the performance of exon prediction methods? What could be an optimal window size for a given set of sequences? Herein, we

address these questions through an investigation of the suitability of different window shapes and length for period-3 exon detection.

### 2.1 Database and Evaluation Metrics

The DNA sequences were first converted into four binary indicator sequences, using the Voss representation [3], as discussed in Section 1. The DFT-based SC measure (2) was then used for the period-3 detection, using different window shapes (with fixed length of 351) such as rectangular, Bartlett, Hanning, Hamming, Blackman, Kaiser and Gaussian (e.g. see Fig. 1). In each case, the entire HMR195 dataset [10] containing 195 mammalian gene sequences, was used. These sequences have maximum length 200,000 bp, and the ratio of human:mouse:rat sequences is 103:82:10, with the mean exon length of 208 bp. The receiver operating characteristic (ROC) curve measure was used to compare the performance of SC measure using different window shapes. An ROC curve explores the effects on true positive (*TP*) and false positive (*FP*) as the position of an arbitrary decision threshold is varied, and plots the *TP* as a function of *FP* of exonic and intronic nucleotide separation method for varying decision threshold values, where *TP* is the number of coding nucleotides correctly predicted as coding and *FP* is the number of non-coding nucleotides predicted as coding. One way of characterizing this result as a single number is to calculate the area under the ROC curve (AUC), with larger areas indicating more accurate detection. Empirically, we found shape parameter values $\alpha = 1.5$ and $\beta = 2$ respectively for the Gaussian and Kaiser windows, more suitable for the DFT-based gene prediction. The Kaiser window was further investigated for different lengths.

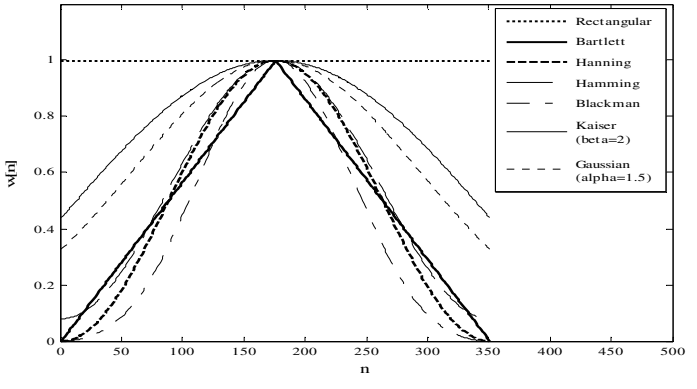


**Fig. 1.** Time-domain representation of different window shapes

### 2.2 Results and Discussion

Table 1 summarizes the AUC results for different window shapes (for a constant window length of 351). It is clear that the window type does affect the performance of the DFT based period-3 detection method. Results show that the Kaiser and Gaussian windows are better than other types, whereas the Blackman is poorer. The performance of rectangular, Bartlett, and Hamming windows seems comparable. It is very

interesting to see from the time-domain representation of different window types (in Fig. 1) that the extremely suppressed edges in different shapes (e.g. Blackman, and Hanning) result in a loss of useful DNA sequence information, and are presumably the main reason for poor gene prediction results using these types. On the other hand, abrupt truncation of sequences as in rectangular type, causes leakage of the DFT power into adjacent frequencies. This work suggests that a window shape with moderately suppressed edges (e.g. Gaussian $\alpha = 1.5$, and Kaiser $\beta = 2$ in our case) would be the more suitable choice for the DFT-based gene prediction.

**Table 1.** AUC results using different window types

| Window Type | Area under ROC curve (AUC) | Window Type | Area under ROC curve (AUC) | Window Type | Area under ROC curve (AUC) |
|---|---|---|---|---|---|
| Rectangular | 0.8008 | Hamming | 0.7979 | Kaiser ($\beta = 2$) | 0.8059 |
| Bartlett | 0.7985 | Blackman | 0.7816 | Gaussian ($\alpha = 1.5$) | 0.8056 |
| Hanning | 0.7923 | | | | |

The AUC as function of beta ($\beta$) for different length Kaiser windows is shown in Fig. 2. Regardless of the window length, the optimum performance of the Kaiser window lies around a value of $\beta = 2$. A similar experiment (results not shown) suggests that the Gaussian window give the optimum performance for a value of $\alpha = 1.5$ (approximate). Clearly, the effect of window length on DFT-based exon prediction is stronger than that due to the window shape parameter. It can be observed that due to an average exon length of 208 bp in HMR195 dataset, the Kaiser window with a length of 201 points performs better than those with lengths 351, 501, and 651, acknowledging results in [11], using a different window type on a completely different

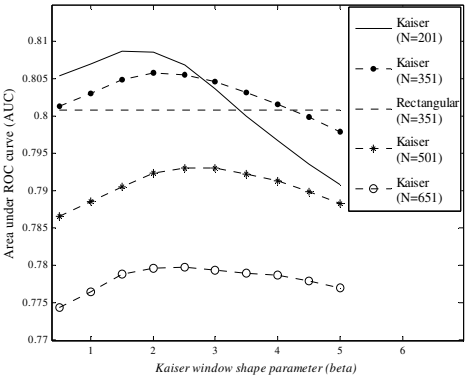


**Fig. 2.** *AUC* vs. *beta* ($\beta$) plot for different lengths of Kaiser window, using the HMR195 dataset

dataset. This too suggests that an unnecessary larger window size not only requires longer computation time, but it also gives poor performance for different window types, compromising the base-domain resolution.

## 3 Proposed Method for Gene Prediction

The paired numeric method has recently been shown [2] to be the most effective DNA symbolic-to-numeric mapping scheme for the DFT-based gene prediction. Results and discussion presented in Section 3, suggest that a Kaiser window ($\beta$ = 2) of length approximately equal to the average exon length of the given dataset is the more suitable choice of window shape and length parameters. Furthermore, the signal boosting technique has been shown [12] a successful post-processing of output signals to enhance genomic protein coding regions and suppress the non-coding regions. Herein, we first modify the frequency-domain PWSR measure (5) by employing the best DNA representation, a suitable window shape and length, and the signal boosting technique. The modified PWSR measure is then combined with the time-domain AMDF method (6) to improve gene prediction accuracy of existing methods. The block diagram of the proposed optimized time-frequency hybrid setup is shown in Fig. 3.

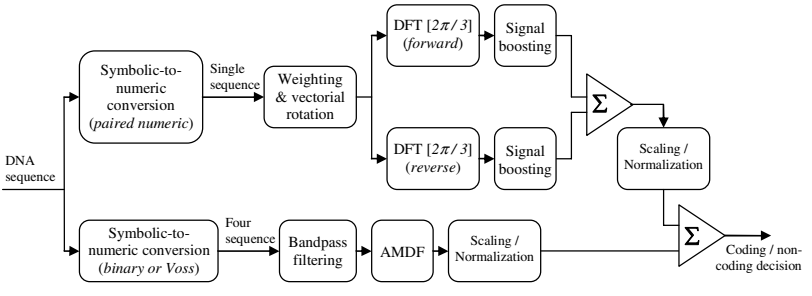


**Fig. 3.** Block diagram for the proposed optimized time-frequency hybrid method

Before applying the PWSR method, the DNA sequence is first converted into numeric values using the paired numeric representation [2], previously shown empirically to be the best available mapping scheme for the gene and exon prediction problem [2]. Weights $w_m$ based on the frequency of occurrence of bases 'A or T' and 'C or G' in coding regions of the training data are then assigned. A reduction in DFT processing is achieved after symbolic-numeric conversion by applying the spectral rotation and weighting of the PWSR measure before rather than after the DFT processing, recognizing that the DFT is a linear transform. The shape and length of the DFT window is another important performance parameter, investigated in Section 2, and a Kaiser window (with $\beta$ = 2) length of 150 base pairs is used herein, assuming human genomic sequences. The recently proposed signal

boosting technique [12] is then applied to the forward and reverse DFT features, to enhance their values in protein coding and suppress them in non-coding regions. According to the signal boosting technique, the protein coding regions are treated as the 'signal', while non-coding regions are treated as the 'noise', and a gain factor $\Gamma(m)$ is calculated as the ratio of a short-term average signal energy $P(m)$ to the estimate of the noise floor level $Q(m)$ for $m = 1,2,\ldots M$, where $M$ is the length of DNA sequence. The boosted signal $\hat{X}(m)$ is then calculated as [12]:

$$\hat{X}(m) = \Gamma(m).X(m) = \frac{P(m)}{Q(m)}.X(m) \tag{7}$$

where $X(m)$ is the period-3 detection DFT feature. Finally, the forward and reverse signal boosted PWSR features are combined with an unweighted sum. The resultant features are then used as an optimized PWSR feature for discrimination of coding and non-coding nucleotides. Finally, we combine 'time' and 'frequency' domain methods, similar to [7]. A simple fusion approach is employed, in which the features from each method are normalized to the range [0, 1] and combined with an unweighted sum.

## 4 Evaluation

### 4.1 Database and Evaluation Metrics

Two datasets consisting of human genomic sequences (e.g. GENSCAN learning and test sets [13]) were employed for the training and testing of different methods. A constant window size of 351 was used for the existing DFT-based SC, SR, and PWSR measures, as suggested in their original descriptions [5]-[7]. A frame size of 117 was used for the AMDF method, similar to [8]. In implementations of the SR, PWSR and the proposed method, prior information (frequency of nucleotide occurrence weights and angular mean and deviation values) was obtained from the GENSCAN learning set. The discrimination power of all methods was measured and compared at the nucleotide level, using evaluation measures such as ROC curves, AUC, and percentage of exonic nucleotides detected as false positives, similar to [7].

### 4.2 Gene Prediction Results

From the ROC curve and area under ROC curve results summarized in Fig. 4 and Table 2, we see that the proposed method outperforms the existing time-domain, frequency-domain, and combined time-frequency measures, giving consistently improved exonic nucleotide detection and the largest area under ROC curve. The proposed method reveals relative improvements of 55.9%, 49.3%, and 26.4% respectively over the SC, SR, and PWSR measures in the detection of exonic nucleotides at a 5% false positive rate. Furthermore, the proposed method gives relative improvements of 27.3% and 15.1% respectively over the AMDF and existing time-frequency hybrid (TFH) measure [7] in the detection of exonic nucleotides at a 5% false positive rate. Although the improvements over existing methods at a 20% or

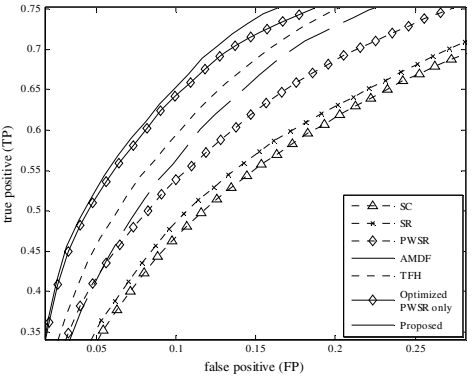


**Fig. 4.** ROC plot using GENSCAN test set

**Table 2.** Summary of results using GENSCAN test set

| Method | Area under ROC curve | % of exonic nucleotides detected as false positive | | | | |
|---|---|---|---|---|---|---|
| | | 5% | 10% | 15% | 20% | 30% |
| SC | 0.7778 | 33.8 | 46.7 | 55.2 | 61.6 | 71.0 |
| SR | 0.7800 | 35.3 | 48.6 | 57.0 | 62.9 | 72.4 |
| PWSR | 0.8123 | 41.7 | 53.8 | 62.5 | 68.7 | 77.3 |
| AMDF | 0.8338 | 41.4 | 56.2 | 66.5 | 72.9 | 81.7 |
| TFH | 0.8448 | 45.8 | 59.5 | 68.8 | 74.9 | 81.6 |
| Optimized PWSR only | 0.8532 | 51.6 | 64.3 | 71.8 | 76.3 | 82.6 |
| Proposed | **0.8650** | **52.7** | **65.4** | **73.8** | **78.3** | **84.8** |

larger false positive rate are more modest, results at low false positive rates are more significant, due to the high likelihood of false positives resulting from the low exonic fraction in eukaryotic genomes.

## 5 Conclusion

We have investigated the effects of window shape and length parameters on DFT-based gene and exon prediction. This revealed that the effect of window length is stronger than that of window shape, and the optimum window length for the given dataset is approximately equal to the average length of exons. We have also proposed a new gene prediction method which employs the most effective DNA representation examined to date in conjunction with suitable window shape and length parameters and a signal boosting technique. Using the GENSCAN test set of human gene sequences; the proposed method outperforms all existing methods in this comparison. Future work may combine this optimized signal processing method with data-driven methods to advance the state of the art in detection of exonic/intronic end-point signals (e.g. acceptor/donor splice sites, start/stop codons).

## References

1. Anastassiou, D.: Genomic Signal Processing. IEEE Signal Proc. Mag. 18(4), 8–20 (2001)
2. Akhtar, M., Epps, J., Ambikairajah, E.: On DNA Numerical Representations for Period-3 Based Exon Prediction. In: 5th IEEE Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland (2007)
3. Voss, R.F.: Evaluation of Long-range Fractal Correlations and 1/f Noise in DNA Base Sequences. Phy. Rev. Lett. 68(25), 3805–3808 (1992)
4. Fickett, J.W.: Recognition of Protein Coding Regions in DNA Sequences. Nucleic Acids Res. 10, 5303–5318 (1982)
5. Tiwari, S., Ramaswamy, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R.: Prediction of Probable genes by Fourier Analysis of Genomic Sequences. Comput. Appl. Biosci. 13, 263–270 (1997)
6. Kotlar, D., Lavner, Y.: Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein Coding Regions. Genome Res. 18, 1930–1937 (2003)
7. Akhtar, M., Epps, J., Ambikairajah, E.: Time and Frequency Domain Methods for Gene and Exon Prediction in Eukaryotes. In: IEEE ICASSP, pp. 573–576 (2007)
8. Ambikairajah, E., Epps, J., Akhtar, M.: Gene and Exon Prediction using Time-Domain Algorithms. In: 8th IEEE Int. Symp. on Sig. Proc. and its Appl., pp. 199–202 (2005)
9. Datta, S., Asif, A.: A Fast DFT Based Gene Prediction Algorithm for Identification of Protein Coding Regions. In: IEEE ICASSP, pp. 653–656 (2005)
10. Rogic, S., Mackworth, A.K., Ouellette, B.F.: Evaluation of Gene-Finding Programs on Mammalian Sequences. Genome Res. 11(5), 817–832 (2001)
11. Akhtar, M., Ambikairajah, E., Epps, J.: Optimizing Period-3 Methods for Eukaryotic Gene Prediction. In: IEEE ICASSP, pp. 621–624 (2008)
12. Gunawan, T.S., Ambikairajah, E., Epps, J.: A Boosting Approach to Exon Detection in DNA Sequences. IEEE Electronic Letters 44(4), 323–324 (2008)
13. Burge, C.: Identification of Genes in Human Genomic DNA. PhD Thesis Stanford University, Stanford, CA, USA (1997)