TIME AND FREQUENCY DOMAIN METHODS FOR GENE AND EXON PREDICTION IN EUKARYOTES

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

[1]The University of New South Wales, Sydney 2052, Australia
[2]UNSW Asia, 1 Kay Siang Road, Singapore 248922

ABSTRACT

The detection of period-3 components in exons of eukaryotic gene sequences enables signal processing based time-domain and frequency-domain methods to predict these regions. In this paper, we improve the prediction accuracy of frequency-domain methods by proposing a new algorithm known as the paired and weighted spectral rotation (PWSR) measure, which exploits both period-3 behaviour and another useful statistical property of genomic sequences. By comparison with existing frequency-domain approaches, the proposed PWSR method reveals relative improvements of 15.2% and 10.7% respectively over spectral content and spectral rotation measures in terms of prediction accuracy of exonic nucleotides at a 10% false positive rate using the GENSCAN test set. Finally, we combine the proposed PWSR with an existing time-domain method to demonstrate further signal processing-based improvements in gene and exon prediction accuracy.

Index Terms— DNA, Signal processing, Correlation, Discrete Fourier transforms, Time-frequency analysis

1. INTRODUCTION

Deoxyribonucleic acid (DNA) consists of genic and inter-genic regions. In eukaryotes, genes are further divided into relatively small protein coding segments known as exons, interrupted by non-coding spacers known as introns. The DNA codons (i.e., triplets of available four types of DNA nucleotides A, C, G, and T) in exon regions encode 20 amino acids and 3 terminator signals. In exons, occurrences of identical nucleotides in identical codon positions is the basis for a periodicity of three interpretation in these regions [1]. The periodicity of three behavior of exons in genomic sequences has been widely used to identify these regions using techniques such as the autocorrelation function (ACF) [1], discrete Fourier transforms (DFT) [2, 3, 4, 5], time-domain algorithms [6], singular value decomposition [7], etc. Accurate exon prediction requires detection of all of the nucleotides in the exon. Despite the existence of many digital signal processing (DSP) applications in this area, the accuracy of exon detection is still limited. The problem is difficult mainly due to noncontiguous and non-continuous nature of genes. Furthermore, often the intergenic and intronic regions make up most of the genome. For example, in the human genome the exonic fraction is as low as 2%. Existing DSP applications in this area can be divided into time-domain and frequency-domain methods. Previous work [8] has shown time-domain techniques perform better than frequency-domain methods for the detection of short coding regions. Despite the existence of these approaches and also data-driven approaches, the accuracy of exon prediction still needs to be improved. Furthermore, in terms of signal processing approaches exploiting period-3 behavior, there has been little comparative work examining exon prediction at the nucleotide level on large databases. In this paper, we review and compare existing approaches, and propose a new algorithm which exploits an alternative statistical property of sequences, computing DFT magnitude and phase angle on both DNA strands.

2. EXISTING METHODS FOR EXON PREDICTION

In order to apply digital signal processing techniques herein, the genomic sequences are first converted into four binary indicator sequences $x_A[n]$, $x_C[n]$, $x_G[n]$, and $x_T[n]$ similar to [7], showing the presence or absence of the respective nucleotides. Out of many available period-3 detection methods, here we consider only three (two 'frequency-domain' and one 'time-domain' approach).

2.1. Spectral Content (SC) Measure

In this fundamental frequency-domain method, sliding window DFTs of four indicator sequences are employed. The periodicity of three suggests a DFT peak at $k = N/3$ in exons, where N is window size, so that calculation of DFT at that point is sufficient. The window can then be moved by sliding one or more points. The plot of

$$SC[k] = \sum_{l} |X_l[k]|^2 \quad (l = A, C, G, and \quad T) \qquad (1)$$

has been used as a SC measure [2] , where $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$ represent DFTs of indicator sequences.

2.2. Spectral Rotation (SR) Measure

Kotlar and Lavner [4] recently proposed a modification to the DFT-based SC measure. They found that the distributions of the DFT phase angle at frequency $2\pi/3$ for

coding regions (i.e., exons) are narrower around a center value than those of non-coding regions, which are almost uniform, within genomic sequences of one particular organism. They proposed the SR measure, which 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 μ_l, to make all of them 'point' in the same direction. The SR measure also divides each term by the corresponding phase angle standard deviations σ_l to give more weight to narrower distributions. The feature

$$SR[k] = \left| \sum_l \frac{e^{-j\mu_l}}{\sigma_l} X_l[k] \right|^2 \quad (l = A, C, G \text{ and } T) \qquad (2)$$

has been used for the detection of exons [4].

2.3. Average Magnitude Difference Function (AMDF)

This well-known time-domain method can be adapted for a numeric DNA sequence $x[n]$ as a function of period $k = 3$:

$$AMDF\ [k] = \frac{1}{N} \sum_{n=1}^{N} |x[n] - x[n-k]| \qquad (3)$$

where N is the window size. The deep null produced by the AMDF at k=3 can be used for exon prediction. We have found in practice that this is enhanced by pre-processing with a second-order resonant filter tuned at $2\pi/3$, to emphasize the period-3 component. A linear combination of the AMDF outputs for each of the four indicator sequences gives the final feature values for exon prediction [6].

3. PWSR AND COMBINED TIME–FREQUENCY EXON PREDICTION

3.1. Nucleotide properties in exonic regions

In exon regions, it has been observed that the frequency of occurrence of DNA nucleotides 'C' and 'G' is higher than 'A' and 'T'. In particular, introns are rich in nucleotides 'A' and 'T' whereas exons are rich in nucleotides 'C' and 'G' [9]. Furthermore, if we calculate DFT phase angle histogram distributions for coding and non-coding regions from the GENSCAN learning set [10], we observe smaller and equivalent angular means for distributions of nucleotides 'C' and 'G' than those of 'A' and 'T'. To fully exploit this property, we pair these nucleotides and define two indicator sequences (i.e., $x_{A\text{-}T}[n]$ and $x_{C\text{-}G}[n]$), which also reduces the cost of DFT processing. A similar approach was used by Datta and Asif [5], however no motivation for the 'A-T' and 'C-G' pairing was given. Figure 1(a) shows DFT phase angle histogram distributions for all coding and non-coding nucleotides from the GENSCAN learning set. There is little difference in coding and non-coding distributions, contrary to claims in [4]. The distributions shown in Figure 1(b) were then obtained by averaging the phase angle values over each coding and non-coding regions for individual nucleotides (i.e., one phase angle value for one complete coding or non-coding region, per nucleotide). A roughly bell-shaped distribution for coding regions can be observed in this case, with distinct differences from that of the non-coding regions. These differences can be exploited by weighting the contributions of $x_{A\text{-}T}[n]$ and $x_{C\text{-}G}[n]$ differently, based on the means and variances of each sequence, as proposed in the following section.

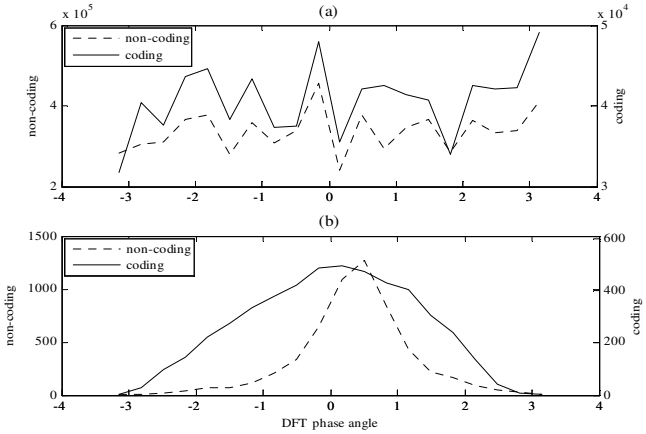

Figure 1. DFT phase angle distributions for coding and non-coding regions of four indicators using GENSCAN learning set, (a) calculated at nucleotide level, and (b) averaged over each entire coding and non-coding region.

3.2. Paired and Weighted Spectral Rotation (PWSR) Measure

As explained in section 3.1, we convert DNA sequences 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, we calculated means and standard deviations of the distributions of DFT phase angle averaged over coding regions (i.e., one phase angle value for one coding region). We also calculated weights based on the frequency of occurrence of nucleotides 'A or T' and 'C or G' in coding regions of the training data. The expression given in (4) is thus proposed as a measure of output feature value in one direction of the DNA sequence:

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

where l = forward (F) and reverse (R) directions of DNA sequence, μ_m and σ_m (m = A-T, C-G) are mean and standard deviation values obtained from distributions of the DFT phase angle averaged over coding regions using the GENSCAN learning set (here, $\mu_{A\text{-}T} = 0.1522$, $\mu_{C\text{-}G} = 0.0675$, $\sigma_{A\text{-}T} = 0.3984$, and $\sigma_{C\text{-}G} = 0.4187$), w_m are frequency of occurrence weights using the GENSCAN learning set (i.e., $w_{A\text{-}T} = 0.4368$, and $w_{C\text{-}G} = 0.5632$), and $X_m[k]$ are the sliding DFT windows of two indicator sequences. In a sliding window DFT, we normally calculate the DFT at a single point (i.e., $k = N/3$ where N is the window size) which suggests that different DFT results will be obtained at a

particular location of the DNA sequence when moving the window in different directions (i.e., the 5′ to 3′ and 3′ to 5′ directions of the same sequence). The expression in (4) is used in the reverse direction of the same DNA sequence (note that due to the paired indicator signals $x_{A\text{-}T}[n]$, $x_{C\text{-}G}[n]$, a DFT in the reverse direction of the same DNA strand is equivalent to DFT on its complementary strand). The proposed PSWR measure is then the combination of forward and reverse measures:

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

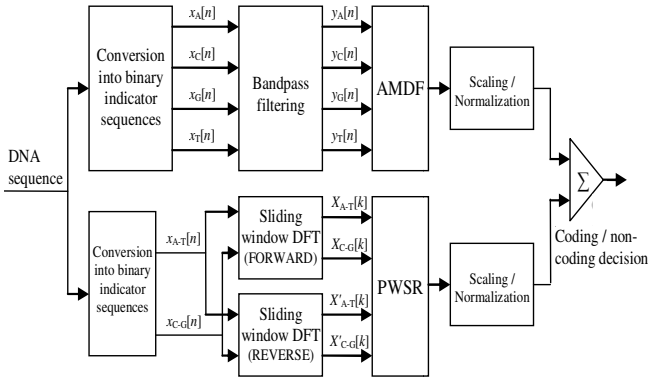

Figure 2. Block diagram for the proposed time-frequency hybrid measure

3.3 Time-Frequency Hybrid (TFH) Measure

We then combine 'time' and 'frequency' domain methods, since the DFT phase angle produces entirely different information to the 'time'-domain magnitude-based AMDF. It has already been shown by Kotlar and Lavner [4] that additionally considering the DFT phase angle is more informative than the magnitude alone. As seen in Figure 2, the proposed PWSR measure is combined with the AMDF, which has previously been found to produce better accuracy than other magnitude-based approaches [8]. 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. The resultant features are then used to discriminate coding and non-coding nucleotides. The length of the analysis window is also an important performance parameter. A shorter window may miss detection of larger exons and vice versa. In the proposed hybrid measure, different analysis window lengths were empirically determined for the AMDF and PWSR methods to ensure the optimal detection of both short and long exons.

4. EVALUATION

4.1. Databases

Two datasets consisting of human genomic sequences were employed for training and testing of methods: the GENSCAN learning set (188 multi-exon sequences), and the GENSCAN test set (64 available multi-exon gene sequences), as listed in [10].

4.2. System Configurations

A constant window size of 351 was used for all DFT-based methods in this work (i.e., SC, SR, and PWSR), consistent with previous work [2, 3, 4, 5]. Empirically, we found a frame size of 117 more suitable for the AMDF method [8]. In the SR implementation, mean and deviation values were obtained from average phase angle distributions for coding regions from the GENSCAN learning set.

4.3. Evaluation Criteria

In order to measure and compare the discriminatory power of all methods, we compared their prediction results at the nucleotide level, contrary to the existing comparisons at exon level or gene level, e.g. [4]. In exon-level detection, the feature value for one point (i.e., nucleotide) in an exon being greater than a decision threshold is sufficient for the detection of that particular exon. Here, we consider the feature values of all nucleotides in an exon, giving more insight into the robustness of a particular detection method. We calculate the percentage number of false positives and percentage specificity at different levels of percentage sensitivity. A threshold output feature value '*Th*' at a particular level of percentage sensitivity '*s*' is the minimum value for which '*s%*' of the exonic nucleotides have feature value greater than '*Th*' [10]. The specificity can be defined as *TP/(TP+FP)*, where *TP* = number of true positive nucleotides, and *FP* = number of false positive nucleotides. We also plot ROC curves for all methods. An ROC curve explores the effects on *TP* and *FP* as the position of an arbitrary decision threshold is varied, and plots the *TP* as a function of *FP* of exonic and intronic nucleotides. One way of characterizing this result as a single number is to calculate the area under the ROC curve, with larger areas indicating more accurate detection methods. We also show the percentage of exonic nucleotides detected as false positives by each method for 10%, 20% and 30%, since false positives inevitably occur in computational methods due to the fact that intronic and intergenic nucleotides make up more than 95% of the eukaryotic genomes.

4.4. Results and Discussion

From the specificity vs. sensitivity, ROC curve and area under ROC curve results summarized in Figures 3 and 4, and Table 1, we see that:

(i) the proposed DFT based PWSR measure outperforms two existing frequency-domain methods, giving consistently higher levels of specificity at each sensitivity level and improved exonic nucleotide detection,

(ii) by comparison with existing frequency-domain approaches, the PWSR method reveals relative

improvements of 15.2% and 10.7% respectively over the SC and SR measures in the detection of exonic nucleotides at a 10% false positive rate,

(iii) by comparison with the exiting time-domain method AMDF, the PWSR method gives consistently higher levels of specificity at each sensitivity level up to 40%, promising a relatively improved gene and exon prediction,

(iv) the time-domain AMDF method is more effective at extracting period-3 based features compared with frequency-domain methods for this application, and

(v) the proposed hybrid measure (TFH) is superior to AMDF alone and all other methods in this comparison.

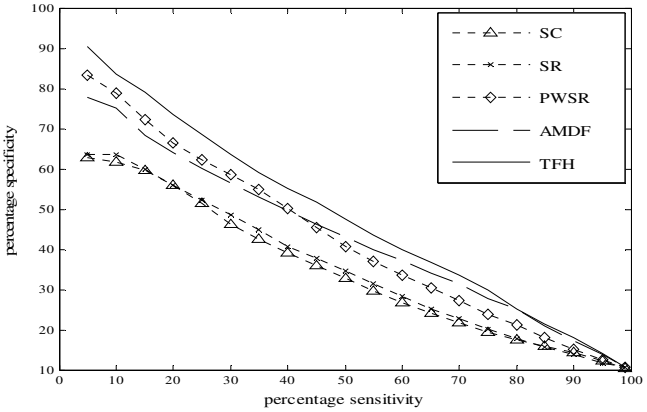

Figure 3. Specificity vs. sensitivity plot using GENSCAN test set

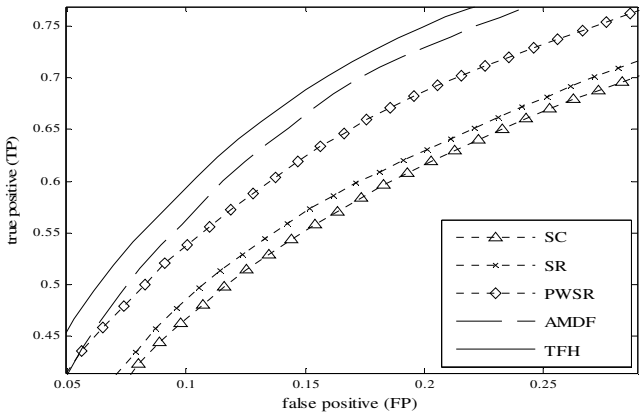

Figure 4. ROC plot using GENSCAN test set

Table 1. Summary of results

Method	Area under ROC curve	% impr. over SC	% of exonic nucleotides detected as false positive		
			10%	20%	30%
SC	0.77778	-	46.7	61.6	71.0
SR	0.78002	0.29	48.6	62.9	72.4
PWSR	0.81232	4.44	53.8	68.7	77.3
AMDF	0.83375	7.20	56.2	72.9	81.7
TFH	0.84484	8.62	59.5	74.9	81.6

5. CONCLUSION

This paper has reviewed three selected existing signal processing methods for gene and exon prediction in eukaryotes. In addition to period-3 behaviour, we have exploited another very useful statistical property of DNA sequences, proposing the paired and weighted spectral rotation (PWSR) measure for gene and exon prediction. A time-frequency hybrid measure has also been introduced that successfully combines the PWSR with the time domain AMDF technique. Using the GENSCAN test set of human genomic sequences, the new PWSR measure outperforms well-known frequency-domain spectral content and spectral rotation measures, while the proposed hybrid measure provides improved prediction accuracy relative to all existing methods. Future work will apply these new signal processing techniques to the detection of other biological signals (e.g. acceptor/donor splice sites, start/stop codons) in order to make a full comparison with state of the art data-driven gene and exon prediction packages such as GENSCAN.

6. REFERENCES

[1] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Res.*, vol. 10, pp. 5303–5318, 1982.

[2] 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, pp. 263–270, 1997.

[3] D. Anastassiou, "Genomic signal processing," *IEEE Signal Proc. Mag.*, vol. 18, no. 4, pp. 8–20, 2001.

[4] D. Kotlar, and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Res.*, vol. 18, pp. 1930–1937, 2003.

[5] S. Datta, and A. Asif, "A fast DFT based gene prediction algorithm for identification of protein coding regions," *ICASSP*, vol. 5, pp. 653–656, 2005.

[6] E. Ambikairajah, J. Epps, and M. Akhtar, "Gene and exon prediction using time-domain algorithms," *IEEE 8th Int. Symp. on Sig. Proc. and its Appl.*, pp. 199–202, 2005.

[7] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," *IEEE Int. Conf. on Emerging Technologies*, pp. 13–17, 2005.

[8] M. Akhtar, "Comparison of gene and exon prediction techniques for detection of short coding regions," *Int. J. of Inf. Tech., Special Issue on Bioinformatics and Biomedical Systems*, vol. 11, no. 8, pp. 26–35, 2005.

[9] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279–303, 2002.

[10] C. Burge, "Identification of genes in human genomic DNA," *PhD thesis Stanford University*, Stanford, CA, 1997. (Datasets are available at: ftp://ftp.cse.ucsc.edu/pub/dna/genes, and http://www.fruitfly.org/sequence/human-datasets.html)