OPTIMIZING PERIOD-3 METHODS FOR EUKARYOTIC GENE PREDICTION

Mahmood Akhtar, Eliathamby Ambikairajah, and Julien Epps

School of Electrical Engineering and Telecommunications
The 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 firstly investigate the effect of window lengths on selected signal processing-based gene and exon prediction methods. We then optimize these methods to improve their prediction accuracy by employing the best DNA representation, a suitable window length, and boosting the output signals to enhance protein coding and suppress the non-coding regions. It is shown herein that the proposed method outperforms major existing time-domain, frequency-domain, and combined time-frequency approaches. By comparison with the existing DFT-based methods, the proposed method not only requires 50% less processing but also exhibits relative improvements of 53.3%, 46.7%, and 24.2% respectively over spectral content, spectral rotation, and paired and weighted spectral rotation measures in terms of prediction accuracy of exonic nucleotides at a 5% false positive rate using the GENSCAN test set.

Index Terms— Signal processing, Discrete Fourier transform, DNA, sequence analysis, genomic signal processing

1. INTRODUCTION

In most eukaryotic genomic sequences, genes are divided into relatively small protein coding regions known as exons, and large non-coding regions known as introns. Different periodicities for these sequences have been reported in the literature. The periodicity of three (which appears mainly due to the occurrence of identical nucleotides in identical codon positions) behavior of exons has been widely used to identify these regions with the help of different DSP methods such as discrete Fourier transforms [1, 2, 3, 4], time-domain algorithms [5] and allpass-based filters [6]. Despite the existence of these and other applications in this area, the accuracy of exon detection is still limited. Apart from the difficulty of the problem itself, which is mainly due to the noncontiguous and non-continuous nature of genes and the low exonic fraction in eukaryotic genomes, the available methods are not well equipped to capture complementary properties of exonic / intronic regions and deal with the background noise in detection of exons at their nucleotide levels.

We show herein that existing methods can be optimized and their exon detection accuracy can be further increased in two respects, by investigating the effects of window lengths on gene prediction methods, and by enhancing the signal strength in coding regions using the recently proposed signal boosting technique [7].

2. EXISTING METHODS FOR GENE PREDICTION

This section gives an overview of four existing signal processing-based methods that exploit periodicity of three and other relevant properties to identify genomic protein coding regions. All frequency-domain methods reviewed in this section employ sliding window based discrete Fourier transform (DFT) to measure peaks at $k = N/3$ arising from the period-3 exon behaviour, where N is the window length.

The spectral content (SC) measure [1] is perhaps the most fundamental DFT-based method for period-3 detection. In this approach, the DNA is first converted into four binary indicator sequences, $x_A[n]$, $x_C[n]$, $x_G[n]$, and $x_T[n]$ showing the presence (i.e., '1') and absence (i.e., '0') of the respective base. The expression given in (1), which combines the magnitudes of individual DFTs (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$), is then used to obtain a total Fourier magnitude spectrum for a segment, or window, of the DNA sequence:

$$S[k] = |X_A[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 + |X_T[k]|^2 \quad (1)$$

The spectral rotation (SR) measure [3] modifies the SC method by rotating four DFT vectors clockwise, each by an angle equivalent to the average phase angle value in coding regions μ (to make all of them 'point' in the same direction). It also divides each term by the corresponding phase angle standard deviations σ to give more weight to narrower distributions of exons. The motivation for this came from observations by authors [3] that the distribution of DFT phase angle at $\theta = 2\pi/3$ is bell-shaped for protein coding regions and close to uniform for non-coding regions.

$$SR[k] = \left| \frac{e^{-j\mu_A}}{\sigma_A} X_A[k] + \frac{e^{-j\mu_C}}{\sigma_C} X_C[k] + \frac{e^{-j\mu_G}}{\sigma_G} X_G[k] + \frac{e^{-j\mu_T}}{\sigma_T} X_T[k] \right|^2 \quad (2)$$

The recently proposed paired and weighted spectral rotation (PWSR) measure [4] 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'), and computes the DFT magnitude and phase angle in the forward and reverse directions of the same DNA strand. The DNA is first converted into two binary indicators, $x_{A\text{-}T}[n]$, $x_{C\text{-}G}[n]$ containing the presence of either of the paired bases. The PWSR then rotates the two complex DFT values $X_{A\text{-}T}[k]$, $X_{C\text{-}G}[k]$ clockwise, each by an angle equivalent to the means of the distributions of the DFT phase angle averaged over coding regions of training data μ_m (one phase angle value per coding region is calculated) to align exonic vectors more effectively than the SR method. Weights w_m based on the frequency of occurrence of the bases 'A or T' and 'C or G' in coding regions of the training data are also assigned. The expression given in (3) can then be used as a feature:

$$PWSR_l[k] = \left| \sum_m \frac{e^{-j\mu_m}}{\sigma_m} . w_m . X_m[k] \right|^2, \quad m \in \{A-T, C-G\} \qquad (3)$$

The PWSR measure then combines (3) for l = forward (F) and reverse (R) directions of the same DNA sequence:

$$PWSR[k] = PWSR_F[k] + PWSR_R[k] \qquad (4)$$

The PWSR measure has been shown to improve on the SC and SR measures for a nucleotide level comparison [4].

The time-domain average magnitude difference function (AMDF), has been shown to be more effective than the period-3 detection measures discussed above [4].

3. EFFECT OF WINDOW LENGTHS

Most existing signal processing-based gene prediction methods have used a window size of 351 with the arbitrary argument that the data window should be 'reasonably long' or a few hundred base pairs long [1, 2, 3, 4]. Herein, we perform the first investigation of the suitability of different window sizes for period-3 exon detection.

3.1. Database and Evaluation Metrics

The combined Burset / Guigo 1996 [8] and HMR195 [9] data set containing 765 vertebrate and mammalian gene sequences, was divided into four subsets. The sequences were categorized according to the average exon length of individual sequences, and for each data subset, the number of gene sequences, number of exons, and average exon length were calculated, as shown in Table 1.

In this experiment, the DFT-based SC measure and AMDF methods for gene and exon prediction were used, representative of 'frequency domain' and 'time domain' methods respectively. Rectangular windows, varying in length between 27 and 378 in increments of 27 bp, were employed. In each case, the area under receiver operating characteristic (ROC) curve, AUC, was calculated for both methods, using the subsets of genomic data from Table 1. The ROC curve evaluation measure can be explained with the help of Figure 1, where 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, and false positive (*FP*) is the number of non-coding nucleotides predicted as coding. An ROC curve explores the effects on *TP* and *FP* as the position of an arbitrary decision threshold is varied.

Table 1. Categorization of combined Burset / Guigo 1996 and HMR195 sequences for window length investigation

Data Subset	Average Exon Length of Individual Sequences (bp)	Number of Gene Sequences	Number of Exons	Average Exon Length of Data Subset
Short (S)	≤ 100	120	603	83
Medium (M)	> 100 & ≤ 200	426	2372	143
Long (L)	> 200 & ≤ 300	73	303	235
Very Long (VL)	> 300	146	319	578

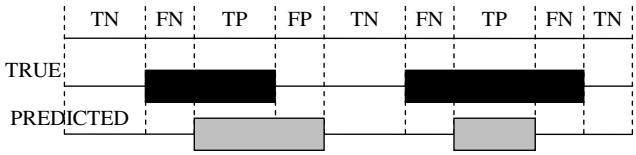

Figure 1. Nucleotide-level measures of prediction accuracy. The black / gray blocks are the actual / predicted exonic regions.

3.2. Window Length Results and Discussion

Figure 2 shows AUC as a function of window size for both methods, across different data subsets. The optimum window length for the DFT-based method depends to a large extent on the average length of exon regions in the dataset, whereas for the AMDF method, this point lies within a short range. A window size somewhere between 100 and 150 seems optimal for the AMDF method, whereas the DFT method requires larger window sizes, especially for very long sequences.

It can be observed that for the DFT method, there is not a large variation in optimal window sizes using short, medium, and long data sets. This suggests that for genomic sequences having exons shorter than 300 bp, the DFT method could give its best performance with a window size in the range 100 to 250 bp. However, to identify exons longer than 300 bp, a much larger window is to be preferred. Clearly the commonly used window size of 351 for DFT based methods is a crude assumption that may only be valid for long exonic regions. It can be further observed from Figure 2 that small exons are poorly detected by both methods, with AMDF always better than DFT. For the detection of larger exons, the computationally cheaper AMDF method still performs better, even with a much smaller window size than the DFT-based SC method.

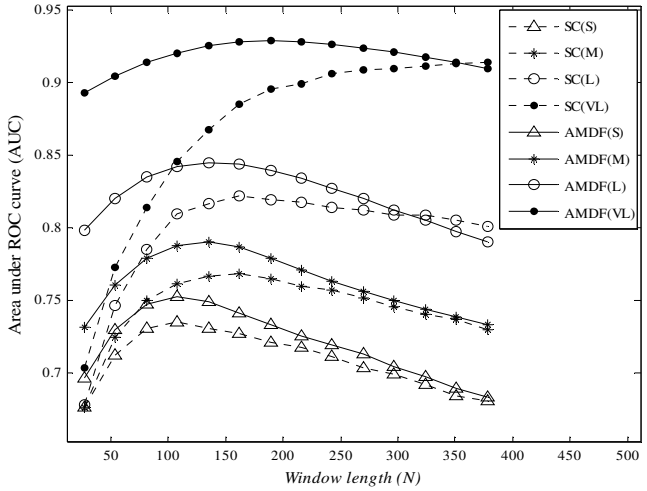

Figure 2. Variation of area under ROC curve with different window lengths for different average exon lengths, for the DFT-based SC and AMDF methods.

4. PROPOSED OPTIMIZED METHOD FOR GENE PREDICTION

We modify the frequency-domain PWSR measure to improve on existing methods [4], in terms of computational complexity and relative accuracy for gene prediction. The block diagram of the proposed setup is shown in Figure 3, where the DNA sequence is firstly converted into numeric values using the paired numeric representation, previously shown empirically to be the best available mapping scheme for the gene and exon prediction problem [10].

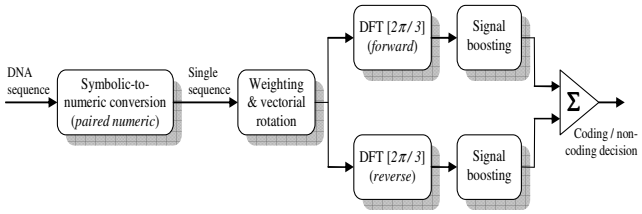

Figure 3. Block diagram for the proposed optimized method.

According to the paired numeric representation, values of +1 and –1 are assigned to A-T and C-G nucleotide pairs respectively. For example, the DNA sequence 'AGTTCTACCGA' has paired numeric representation:

$$x[n] = \{+1, -1, +1, +1, -1, +1, +1, -1, -1, -1, +1\}.$$

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 length of the DFT window is another important performance parameter, investigated in section 3, and here we use a window length of 150 base pairs, assuming human genomic sequences.

The recently proposed signal boosting technique [7] 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, used originally in enhancement, the protein coding regions are treated as the 'signal', while non-coding regions are treated as the 'noise'. The gain factor $\Gamma(m)$, which weights the period-3 detection feature $X(m)$, can be 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,… M, where M is the length of DNA sequence. The short-term signal energy is calculated as:

$$P(m) = (1-\alpha).P(m-1) + \alpha.X(m) \qquad (5)$$

where α is a small positive constant responsible for controlling the changes in signal energy and smoothing the signal. The noise floor estimate is a slowly varying factor and is calculated as:

$$Q(m) = \begin{cases} (1+\beta).Q(m-1), & Q(m-1) \le P(m) \\ P(m) & Q(m-1) > P(m) \end{cases} \qquad (6)$$

where β is a positive constant much smaller than α that controls the speed of adaptation of the noise floor level to the changes. The boosted signal $\hat{X}(m)$ is then calculated as:

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

Finally, a simple fusion approach is employed to combine the boosted spectral content based features, in which the forward and reverse boosted features are normalized to the range [0, 1] and combined with an unweighted sum. The resultant features are then used as a feature for discrimination of coding and non-coding nucleotides.

5. EVALUATION

5.1. Database and Evaluation Metrics

Two datasets consisting of human genomic sequences were employed for the training and testing: the GENSCAN learning set (188 multi-exon sequences), and the GENSCAN test set (64 available multi-exon gene sequences), as listed in [11].

A constant window size of 351 was used for the existing DFT-based SC, SR, and PWSR measures, as suggested in their original descriptions [1, 3, 4]. A frame size of 117 was used for the AMDF method, similar to [5]. 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. Empirically, we found $\alpha = 0.01$ and $\beta = 0.0005$ most suitable parameters, to enhance the signal strength in protein coding and suppress them in non-coding regions.

The discriminatory power of all methods was measured and compared at the nucleotide level, using evaluation measures such as AUC, and percentage of exonic nucleotides detected as false positives, similarly to [4].

5.2. Gene Prediction Results

Table 2 summarizes the comparative evaluation of the proposed gene prediction method with selected existing approaches. 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 53.3%, 46.7%, and 24.2% 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 25.1% and 13.1% respectively over the AMDF and time-frequency hybrid (TFH) measure [4] in the detection of exonic nucleotides at a 5% false positive rate. Although the improvements over existing methods at a 20% or 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. We conjecture here that a further small gain in accuracy over existing methods may be obtained by combining the proposed method with the AMDF in a way similar to TFH measure in [4].

Table 2. Comparison of period-3 exon detection methods evaluated on the 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
Proposed (window length selection only)	0.8501	50.8	63.2	70.9	76.2	**83.1**
Proposed	**0.8527**	**51.8**	**64.3**	**71.5**	**76.4**	82.3

6. CONCLUSION

We have investigated the effects of window lengths on two gene and exon prediction methods: AMDF and the DFT-based SC measure. This revealed the optimum window length for the AMDF method, of around 150 bp, to be relatively independent of the average exon lengths. For the DFT-based SC measure, a longer window is generally required, except for exons shorter than 300 bp. Results on the combined Burset / Guigo 1996 [8] and HMR195 [9] data set strongly suggest that *a priori* knowledge of the average exon length of an organism can help researchers decide the optimal window length for signal processing methods applied to the detection of unknown exons of same organism. For example, in the human genome, about 80% of the exons on each chromosome are smaller than 200 bp in length [12], so for the detection of most human exons, a window length around 150 bp could be expected to give good DFT-based performance.

We have also proposed an optimized method for eukaryotic gene prediction which employs the most effective DNA representation examined to date in conjunction with a suitable window length, the paired and weighted spectral rotation measure and a signal boosting technique. Using the GENSCAN test set of human genomic sequences, the proposed method outperforms all existing methods in this comparison. Future work will 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).

7. ACKNOWLEDGMENT

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

8. REFERENCES

[1] 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.

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

[3] 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.

[4] M. Akhtar, J. Epps, and E. Ambikairajah, "Time and frequency domain methods for gene and exon prediction in eukaryotes," in *Proc. IEEE ICASSP*, pp. 573–576, 2007.

[5] 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.

[6] P. P. Vaidyanathan, and B. -J. Yoon, "Gene and exon prediction using allpass-based filters," in *Proc. IEEE GENSIPS* (Raleigh, NC, USA), 2002.

[7] T. S. Gunawan, E. Ambikairajah, J. Epps, "A signal boosting technique for gene prediction," in *Proc. IEEE ICICS*, 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 B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome Research*, vol. 11, no. 5, pp. 817–832, 2001.

[10] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in *Proc. IEEE GENSIPS* (Tuusula, Finland), 2007.

[11] C. Burge, "Identification of genes in human genomic DNA," *PhD thesis Stanford University*, Stanford, CA, 1997.

[12] M. K. Sakharkar, V. T. Chow, and P. Kangueane, "Distributions of exons and introns in the human genome," *In Silico Biology*, vol. 4, no. 4, pp. 387–393, 2004.