Identification of Eukaryotic Genes with Improved Noise Suppression

D. K. Shakya[1], Rajiv Saxena[2], S. N. Sharma[3]

[1] *Department of Biomedical Engineering, Samrat Ashok Technological Institute, Vidisha, 464001, India*
Email:devendrashakya@rediffmail.com

[2] *Department of Electronics and Communication Engineering, Jaypee University of Engineering and Technology, Raghogarh, Guna, 473226, India*
Email:rsaxena2001@yahoo.com

[3] *Department of Electronics and Instrumentation Engineering, Samrat Ashok Technological Institute, Vidisha, 464001, India*
Email:sanjeev_n_sharma@rediffmail.com

Abstract

The proposed algorithm for gene prediction compares the N/3 spectral components of DNA signal with the corresponding spectrum of period-3 suppressed DNA signal. In the DNA N/3 spectrum, for the bases for which the difference between these two spectrums is within a predefined threshold level, the signal values are replaced by the difference signal of the two spectrums. This substitution suppresses the noise in the non-coding regions of N/3 spectrum, while the coding region signal values are not affected, resulting in an improved detection. Performance of this algorithm when evaluated on HMR 195 dataset, in terms of area under the ROC curve, exhibits a 5% improvement on the DFT-based spectral content measure.

Key Words*: Genomic Signal Processing, DNA, Protein coding regions, DFT, IIR Digital Filters.*

1. Introduction

Deoxyribonucleic Acid (DNA) sequences are symbolically represented by a character string consisting of four alphabets, A, T, C and G, representing four nucleotide bases, adenine, thymine, cytosine and guanine respectively. A DNA sequence is separated into genic and intergenic regions, and in eukaryotic cells (cells with nuclei) genes are further divided into alternating sub-regions called introns and exons. One of the present challenges of analyzing the DNA sequences is to determine the protein coding regions (exons) in eukaryotic gene structures [1, 2].

Base sequence in the protein-coding region has a strong period-3 component due to codon structure involved in the translation of the base sequence into amino acids [3]. Based on the period-3 property a number of algorithms have been developed to identify the protein coding regions. In the period-3 based methods like anti-notch and multistage filtering [4], quadratic windowing [5], and boosting method [6], emphasis has been on suppressing the signal in the non-coding regions and boosting the coding regions [9]. In this work a simple algorithm to achieve better discrimination between coding and non-coding regions is proposed. Noise present in the non-coding regions has been captured using a notch filter. This noise has been

used to reduce the noise level in non-coding regions without affecting the signal in the coding regions.

2. DNA Numerical Representation and Period-3 Spectrum

Processing of DNA sequences using DSP methods require their conversion from a character string comprising of four characters, A, T, C, G, into numerical sequences as a first step. The binary or Voss representation scheme for DNA mapping used in [6] is known for high computational requirements [1]. Therefore EIIP (electron-ion-interaction-potential) values associated with each nucleotide are used in this paper to map DNA character strings into numerical sequences for computational work [2]. In the EIIP method [2], the EIIP values associated with four nucleotides are used to map DNA character string into a single numerical sequence, x(n). The EIIP values for the nucleotides are as follows- A = 0.1260, G = 0.0806, T = 0.1335, and C = 0.1340.

After this mapping, Discrete Fourier Transform (DFT) is used for spectrum analysis of finite-length windowed DNA numerical sequences. The DFT of a length-N block of x(n) is defined as-

$$X[k] = \sum_{n=0}^{N-1} \mathrm{x(n)} \ . \mathrm{w(n)} . \mathrm{e}^{-\mathrm{j}2\pi \mathrm{kn/N}} \quad , \qquad 0 \le k \le N-1 \tag{1}$$

where, w(n) is a window function. Because of the period-3 property, the DFT coefficients corresponding to k=N/3 are large in coding regions. The N/3 coefficients are obtained using (1) and are then used to obtain the spectral content (SC) measure [3], as follows-

$$S[k] = \left| X[k] \right|^2 \tag{2}$$

The window is then slided by one or more bases and S[N/3] is recalculated. The plot between S[N/3] and nucleotide position gives a clear identification between coding and non-coding regions.

3. Gene prediction algorithm (GPA)

Proposed algorithm captures the noise present in the non-coding regions by comparing the N/3 spectral components of DNA signal with the corresponding spectrum of period-3 suppressed DNA signal. In regions having 3-base repeat magnitude of two spectrums differ significantly, whereas they follow each other closely in other regions. If the difference between the two N/3 spectrum magnitudes is within a predefined threshold level the codon is assumed in the non-coding region. The threshold value for optimum detection is determined by plotting Receiver Operating Characteristic Curves (ROC) [1], for different thresholds, for a small set of sequences selected from the dataset. In the DNA N/3 spectrum, for the bases for which the difference between these two spectrums is within this threshold level, the signal values are replaced by the difference signal of the two spectrums. This substitution suppresses the noise in the non-coding regions of N/3 spectrum, while the coding region signal values are not affected, resulting in an improved detection. Detection of protein coding regions by GPA comprises of following steps-

Step 1 – Calculate the period-3 components for DNA signal using (1) at all base locations.

Step 2 – Filter the DNA signal by passing it through a notch filter with filter notch centered at $2\pi/3$.

Step 3 – Calculate the period-3 components for the filtered signal obtained in Step-2 using (1) at all base locations.

Step 4 – Compute difference signal by subtracting signal obtained in Step-3 from the corresponding signal of Step1.

Step 5 – Identify the base locations for which the difference signal is less than the predefined threshold level. For these base locations modify the signal at Step-1 by substituting difference signal for their initial values.

Step6 – Calculate the power spectrum of the modified Step1 signal obtained in Step-using (2).

4. Experimental Results

The performance of this algorithm has been evaluated on HMR195 data set [7]. Bartlett window of length 351 has been used in (1), as it provides the optimal window shape for processing genomic sequences in the HMR195 data set [8]. For notch filtering a second order all pass IIR filter [4] has been used. The transfer function of the filter with poles at $Re^{\pm j\theta}$ is given by (3). The filter notch is centered at frequency $2\pi/3$ and the value of R is selected as 0.992.

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

Notch filtering process removes the period-3 component of genomic sequence. The filtered signal is subjected to a sliding window based DFT operation to obtain the period-3 spectrum. This spectral output approximates the noisy component present in the coding and non-coding regions. The threshold level used is 0.2 and is obtained using 24 single and multi-exon sequences selected arbitrarily from the HMR195 data set. This threshold level has been decided by plotting ROC curves for different threshold levels, using Steps 1 to 6 of GPA. These curves are shown in Fig.1, with curve corresponding to a threshold of 0.2 providing maximum area under the ROC curve (AUC). To illustrate the quality of the GPA results, F56F11.4 DNA sequence from 7021 to 15080 has been taken as a sample sequence. Noise capturing by GPA and its comparative gene detection performance with respect to DFT-based SC measure [6] for this sample sequence, is illustrated in Fig.2 and Fig.3 respectively.

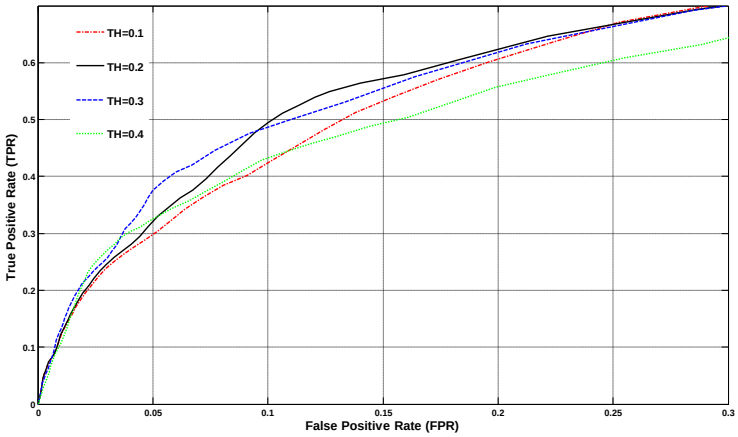

Figure 1. Comparative Threshold Levels

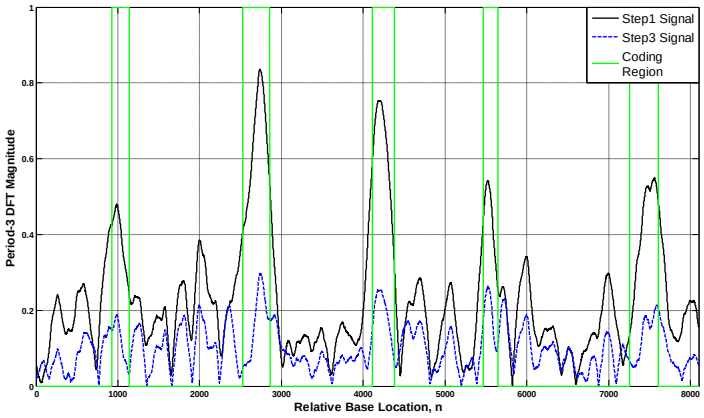

Figure 2. Capturing Noise for Gene F56F11.4

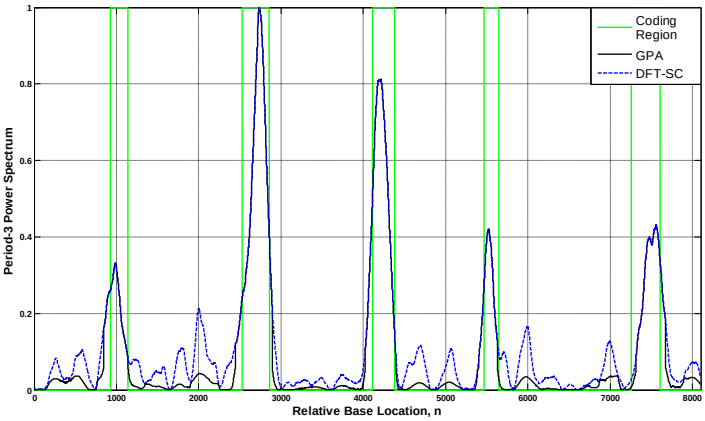

Figure 3. Comparative Results for F56F11.4

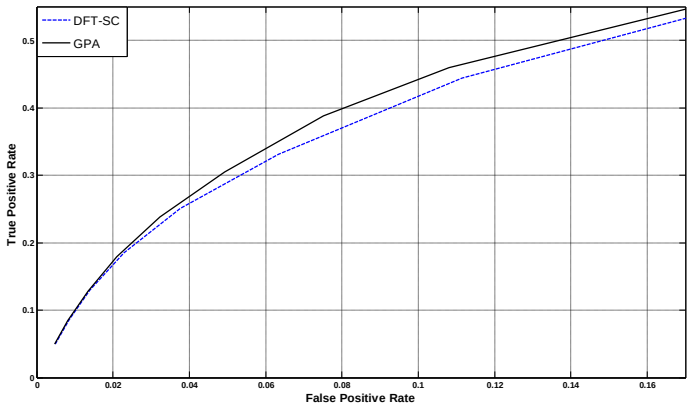

Figure 4. ROC Curves for Period-3 Detection

5. Comparative Performance Analysis

To compare the performance of GPA with DFT-based SC measure ROC curves for HMR 195 data set are plotted in Fig 4. False positives inevitably occur in computational methods due to the fact that intronic and intergenic nucleotides make up more than 95% of the eukaryotic genome. Owing to high likelihood of false positives resulting from the low exonic fraction in eukaryotic genomes, results at low false positive rates are more significant. The proposed GPA provides a larger true positive rate in this region. In Fig.4, AUC in this region, for DFT-based SC measure and GPA are 0.715 and 0.751 respectively. Thus GPA exhibits relatively more accurate gene and exon prediction improving on the well known DFT-based SC measure by 5%.

6. Conclusion

A simple algorithm based on direct capturing of noise in period-3 power spectrum has been developed for better detection of protein coding regions. Performance of the algorithm has been evaluated on HMR195 data set. It provides 5% improvement in terms of AUC over DFT-based SC measure. Existing approaches are heavily dependent on the compositional statistics of the training set sequences. Recently reported boosting technique [6] uses 60% of the HMR 195 dataset as training set sequences. The proposed algorithm does not require a large training set for the empirical determination of threshold. Also the threshold value is model independent and can be used for any other data which is not a part of HMR195 data set. The threshold is observed to be dependent only on the mapping scheme and window function. This algorithm can also be used for the detection of other periodicities present in the DNA sequences.

References

[1] M.Akhtar, J.Epps, and E.Ambikairajah, "Signal processing in sequence analysis: Advances in eukaryotic gene prediction," IEEE Journal of Selected Topics in Signal Processing, vol. 2, no. 3, pp. 310-321, June 2008.

[2] K.D. Rao and M.N.S. Swamy "Analysis of genomics and proteomics using DSP techniques," IEEE Transactions on Circuits and Systems-1, vol. 55, no. 1, pp. 370-378, February 2008.

[3] S.Tiwari, S.Ramachandran, A.Bhattacharya, S.Bhattacharya and R.Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," CABIOS, vol. 13, no. 3, pp. 263-270, 1997.

[4] P.P.Vadyanathan and B.J.Yoon, "Digital filters for gene prediction applications," in Proceedings 36th Asilomer Conference on Signals Systems and Computers, Monterey, CA, November 2002.

[5] T.W.Fox and A.Carreira, "A digital signal processing method for gene prediction with improved noise suppression," EURASIP Journal on Applied Signal Processing, vol. 2004, no. 1, pp. 108-114, 2004.

[6] T.S.Gunawan, J.Epps, and E.Ambikairajah, "Boosting approach to exon detection in DNA sequences," Electronics Letters, vol. 44, no. 4, pp. 323-324, 2008.

[7] S. Rogic, A.K. Mackworth and B.F. Ouellette, " Evaluation of gene finding program on mammalian sequence", Genomic Research, vol. 11, no. 5, pp. 817-832, 2001.

[8] T.S. Gunawan, "On the optimal window shape for genomic signal processing", Proceeding of the International Conference in Computer and Communication Engineering, pp. 252-255, May 13-15, 2008.

[9] D.K. Shakya, Rajiv Saxena, and S.N. Sharma, 'A Simple Algorithm for Gene Prediction with Improved Noise Suppression', Proceedings of the 10th IEEE International Conference on Signal Processing, Beijing, China, Oct. 2010, pp.1765-1768

Authors

D. K. Shakya received the B.E. degree in Electronics and Instrumentation Engineering from Barkatullah University, Bhopal, M.P., India, in 1999, and the M.E. in Digital Technique and Instrumentation from Rajiv Gandhi Proudyogiki Vishwavidyalaya Bhopal, M.P., India, in 2002. He is currently working as an assistant professor in the Department of Biomedical Engineering, Samrat Ashok Technological Institute, Vidisha, M.P., India, and pursuing Ph.D. degree. His teaching and research interests include genomic and proteomic signal processing, digital filter design, bio-signal and image processing.

S. N. Sharma received the B.E. degree in Electronics and Instrumentation Engineering from Barkatullah University, Bhopal, M.P., India, in 1991, M.E. degree in Measurements & Instrumentation Engineering from University of Roorkee (IIT Roorkee), India, in 1993, and the Ph.D. degree in Electronics and Communication Engineering with specialization in Signal Processing from Thapar University, Patiala. He is currently working as an associate professor in the Department of Electronics and Instrumentation Engineering, Samrat Ashok Technological Institute, Vidisha, M.P., India, His teaching and research interests include digital signal processing, genomic signal processing, fractional Fourier transform, digital filter design, bio-signal processing. He is having 10 publications in reputed journals and refereed conferences.

Rajiv Saxena received the B.E. degree in Electronics and Telecommunication Engineering in the year 1982 from Jabalpur University, M.E. degree in Digital Techniques & Data processing Engineering) from Jiwaji University, Gwalior in 1990, and the Ph.D. degree from University of Roorkee (IIT Roorkee), India, in year 1996 in Electronics & Computer Engineering. He is currently working as a professor and head and Dean (Acd. Affairs) in the Department of Electronics and Communication Engineering, Jaypee University of Engineering and Technology, Raghogarh, Guna, M.P., India. His teaching and research interests include digital signal processing, communication engineering and system, integral transform, digital image processing, mobile communication system. He has 25 year experience in teaching. He has guided 6 Ph.Ds. scholars and published more than 30 research papers in various international and national refereed journals and conferences. He received best paper award by IETE, New Delhi, in year the 2008. Dr. Saxena had an Industrial Experience of about two years in Automation of Textile Machinery with Reliance, GRASIM, CIMMCO and Orimpex Textile Industries.