



A DSP-Based Approach for Gene Prediction in Eukaryotic Genes

D. K. Shakya¹, Rajiv Saxena², and S. N. Sharma³

¹Department of Biomedical Engineering,
Samrat Ashok Technological Institute, Vidisha, 464001, India

²Department of Electronics and Communication Engineering,
Jaypee University of Engineering and Technology, Raghogarh, Guna, 473226, India

³Department of Electronics and Instrumentation Engineering,
Samrat Ashok Technological Institute, Vidisha, 464001, India

devendrashakya@rediffmail.com, rsaxena2001@yahoo.com, sanjeev_n_sharma@rediffmail.com

Abstract: A simple algorithm to improve the identification accuracy of protein coding regions (exons) in Deoxyribonucleic Acid (DNA) sequences exploiting period-3 property is proposed. Three base periodicity is quite pronounced in exons and is commonly used in Digital Signal Processing (DSP) based methods to locate the exonic regions. Improvement in the accuracy of the protein coding regions has been achieved by extracting the background noise that comes from long range correlation present in DNA sequences and then eliminating this noise from the period-3 power spectrum. Proposed algorithm is data independent as it does not require the empirical determination of any parameter for increasing the discrimination between coding and non-coding regions of a DNA sequence. Performance of the algorithm has been evaluated on F56F11 *C.elegans* chromosome-III nucleotide sequences. Performance of this algorithm has been compared with the spectral content method and an improvement in the correlation coefficient (CC), the performance metric used in this work, is observed.

KeyWords: Deoxyribonucleic Acid (DNA), Protein coding regions, Period-3 property, Discrete Fourier Transform (DFT), IIR Digital Filters, Genomic Signal Processing.

1. Introduction

DNA sequences are of fundamental importance in understanding living organisms, since all the information of the hereditary and species evolution is contained in these macromolecules. The DNA sequence comprises of four key chemicals, adenine (A), thymine (T), guanine (G), and cytosine (C). One of the present challenges of analyzing the DNA sequences is to determine the protein coding regions (exons) in eukaryotic gene structures [1, 2]. The difficulty of the problem is mainly due to the noncontiguous and non-continuous nature of genes (i.e., DNA consists of genic and intergenic regions, and eukaryotic genes are further divided into relatively small protein coding segments known as exons, interrupted by non-coding spacers known as introns). Furthermore, often the intergenic and intronic regions make up most of the genome. Figure 1 shows a DNA sequence.

In eukaryotes, exon regions are separated by introns, whereas in procaryotes these regions are continuous. Base sequences in the protein-coding regions have a strong period-3 component due to codon structure involved in the translation of the base sequences into amino acids [3]. Fourier analysis of DNA sequences is used to identify possible patterns in coding and non-coding regions. While intronic sequences show a rather random pattern, exonic sequences show periodicities of 3, 10.5, 200, and 400 [4]. Three base periodicity is quite pronounced and is commonly used in Digital Signal Processing (DSP) based methods to locate the exonic regions. Periodicity of three is present in the example periodic sequence: A-- A-- A-- A-- ..., where blanks can be filled randomly by A, T, C or G. This sequence shows a periodicity of three because of the repetition of the base A. Based on the period-3 property a number of algorithms have been developed to identify the protein coding

threshold are considered to be in coding regions otherwise they are recorded as non-coding regions codons.

3. Proposed Algorithm (PA)

The proposed algorithm (PA) is shown in Figure 2. The DFT magnitude values at $k = N/3$ for DNA signal are obtained using (1).

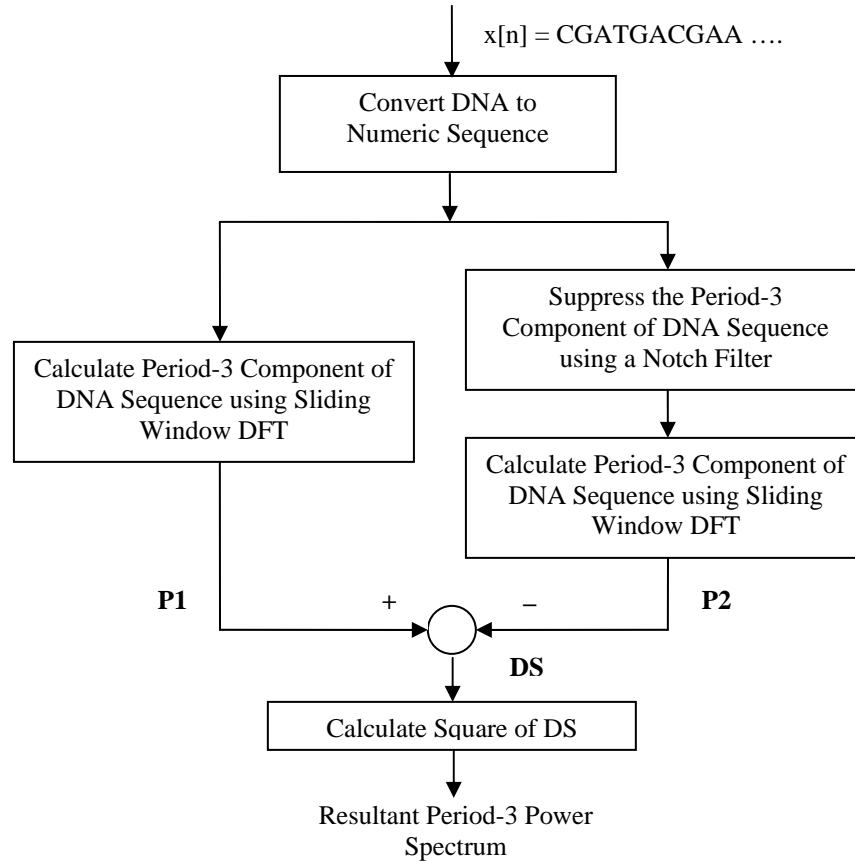


Figure 2. Proposed Algorithm for Gene Prediction

In (1) Bartlett window of length 351 has been used as it provides optimal window shape for processing genomic sequences in [11]. By sliding the window by one sample the process is carried out over the entire DNA sequence and the resultant signal obtained is shown in the algorithm by P1. In signal P1, that represents period-3 magnitude components of DNA data, non-coding region signals representing the noise are not suppressed effectively.

To capture this background noise which comes due to long-range correlation exhibited by DNA sequences both in the genic regions and intergenic regions, and eliminate it from P1, the numeric DNA sequence is first passed through a second order all pass Infinity Impulse Response (IIR) notch filter [7]. IIR filters require less computation and memory than FIR filters and can be very efficient here. Such filters can be built from second order allpass filters. The transfer function of the filter with pole at $\text{Re}^{\pm j\theta}$ is given by (3).

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

4. Comparative Performance Evaluation

The performance of proposed algorithm is compared with the DFT spectral content method [3]. Comparative performance is illustrated in Figure 4 to Figure 6. These figures illustrate the suppression of noise present in non-coding regions. For evaluation of gene structure prediction programs different measures of prediction have been discussed in [12, 13] and can be explained with the aid of Figure 7. 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. Sensitivity (S_n) is the probability of a nucleotide being predicted as coding given that it is actually coding and specificity (S_p) is the probability of a nucleotide being actually coding given that it has been predicted as coding. Both S_n and S_p can be viewed as conditional probabilities. Neither S_p nor S_n alone constitutes good measures of global accuracy.

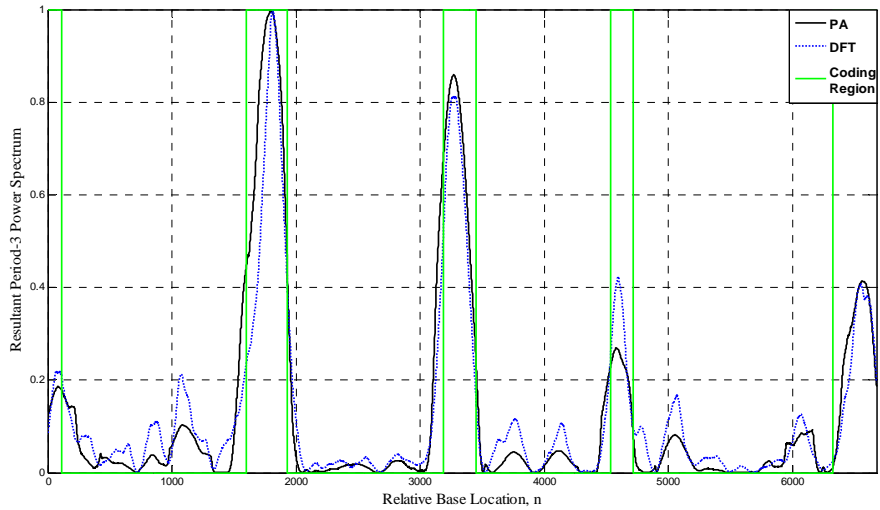


Figure 5. Comparative Results for F56F11.4a

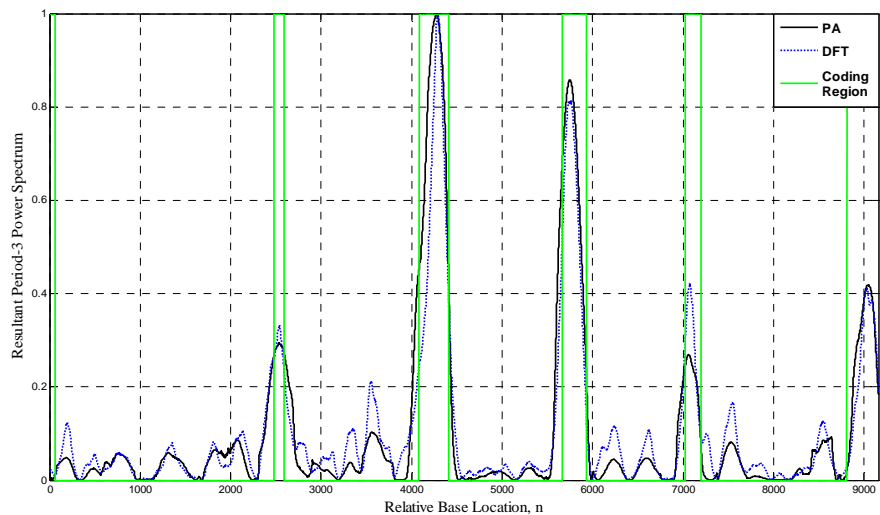


Figure 6. Comparative Results for F56F11.4b

future work by avoiding the suppression of coding region signals. Also the proposed algorithm will be generalized to improve identification accuracy using other transforms like wavelet.

References

- [1] J. Tuqnan and A. Rushdi, "A DSP Approach for finding the codon bias in DNA sequence," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 343-356, 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] 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.
- [5] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8-20, July 2001.
- [6] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes," *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 29-42, 2004.
- [7] 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.
- [8] 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.
- [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, 2010, pp.1765-1768
- [10] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proc. IEEE*, vol. 66, pp. 51-83, 1978.
- [11] 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.
- [12] C. Burge, "Identification of genes in human genomic DNA", Ph.D. dissertation, Stanford University, Stanford, CA, 1997.
- [13] M. Burset and R. Guigo, "Evaluation of gene structure prediction program", *Genomic*, vol. 34, pp. 353-367, 1996.

