Computational Study and Performance Evaluation of Different Genomic Signal Processing Methods for Identification of Protein Coding Regions (Exon Regions) of DNA Sequence

M.N.Vamsi Thalatam
Associate Professor
Department of Computer Applications
GVP College for Degree & PG Courses,
Visakhapatnam, AP, India-530045,

Allam Appa Rao, PhD
Director, CR Rao Advanced Institution for
Mathematics, Statistics &Computer Science
University of Hyderabad Campus,
Hyderabad, AP, India.

ABSTRACT
Recent developments in the area of genomic signal processing (GSP) reveal that this approach has important role in the analysis of genomic sequence, structure and function as well as the gene regulation of different organisms. In this paper we analyze different genomic signal processing methods used for identification of exon coding regions in DNA sequence. The gene sequences of interest are mapped to electron ion interaction potential (EIIP) values of nucleotides and these transformed numerical gene sequences are processed through different signal processing techniques like discrete Fourier transform (DFT), auto regressive (AR) and adaptive auto regressive (AAR) methods. The performance evaluation in terms of computational time is estimated and analyzed. By applying the EIIP mapped sequence to these DFT, AR and AAR methods, the effective computational time is abruptly reduced in AAR method compared to the DFT and AR methods. We tested five sequences of c-elegans) [AF099922], [FO080874.2], [FO081434]), fruitfly [NM_170135] & homosapien (BDNF [NG_011794]).

Keywords
Genomic Signal Processing (GSP), EIIP, DFT, AR and AAR

1. INTRODUCTION
Most signals and processes in nature are continuous. However, genomic information occurs in the form of discrete sequences. The meaning of genomic information is nothing but the information about DNA and protein data. Application of digital signal processing (DSP) techniques to the genomic data is termed as genomic signal processing (GSP). The fundamental functionality of any living cell of an organism is the production of proteins and these proteins are basically transformed from the DNA sequence. The DNA is composed of chromosomes that consist of genes. The important regions of DNA that code for proteins are called Exons and Introns. The challenging task of biologist is the identification of these protein coding regions with their traditional methods [1]. Different computational approaches are proposed by many people for identification of exon regions in the area of genomic signal processing [2]. Genomic signal processing involves the conversion of DNA alphabet sequence into numerical sequence by using mapping techniques like Binary Indicator sequence method [3], Electron Ion Interaction Potential (EIIP) method [4] and Complex indicator Sequence method [5].The "EIIP indicator sequence" is generated by replacing the electron ion interactions potentials of the nucleotides A, T, G, Cs in the given DNA sequence. Using this sequence we can reduce the time complexity by nearly 75% as well as the spectrum obtained from this sequence reveals the better period three property than the binary indicator sequence [6].

For small DNA sequences the predicted protein coding regions are not effective when Fourier based methods are used also these methods are restricted in terms of frequency resolution and spectral leakage issues [7]. To overcome these problems autoregressive (AR) model [8] and adaptive auto regressive model (AAR) [9] are proposed for detection of period-3 property in DNA sequence and to locate the protein coding regions. In this paper we apply the EIIP indicator sequence of five DNA sequences of three different species to the three methods DFT, AR and AAR models and estimated and analyzed the computational time complexity.

1.1 Numerical representation of nucleotide data
Most of the identified nucleotide data is available freely in different online databases over the web. Mainly all such data is retrieved through the Entrez search engine available at National Centre for Biotechnology Information (NCBI) [10].The DNA sequence is in form of sequence of four nucleotides. In order to process these sequences using digital signal processing methods the nucleotide character has to be represented in equivalent numerical form. For most suitable and reliable representation of these sequences into numerical sequence, each nucleotide is assigned to a physical characteristic of that nucleotide and that should be relevant to the biological activity of the molecule. The energy of delocalized electrons in nucleotides has been calculated as the Electro Ion Interaction potential (EIIP) [11]. The EIIP values of the respective nucleotides are shown in table (1). This EIIP is most suitable known nucleotide property that can be used in sequence analysis of protein coding regions. To perform the gene prediction based on period-3 property the total DNA sequence is first converted into EIIP indicator sequence. The DNA sequence D(n) is mapped into EIIP (n),which indicate the presence of nucleotides with their respective EIIP value given below at location n.

Table 1. EIIP Values of Nucleotides

Nucleotide	EIIP Value
A	**0.1260**
G	**0.0806**
C	**0.1340**
T	**0.1335**

For example, D(n)= {G A T A C G T C T T},then E(n)=D_{EIIP}(n)={ 0.0806 0.1260 0.1335 0.1260 0.1340 0.0806 0.1335 0.1340 0.1335 0.1335} is the equivalent sequence of the above example.

2. METHODS

2.1 Discrete Fourier Transform (DFT):

If E(n) is the given sequence and let D_{EIIP}(k) or E(k) is the corresponding discrete Fourier transform and is given by,

$$E(k) = \sum_{n=0}^{N-1} D(n)\ e^{(-\frac{j2\pi kn}{N})}$$

Where j= $\sqrt{-1}$ and D(n) is the discrete sequence over a period of N. And k = 0,1……N-1.And the corresponding absolute value of the power spectrum of E(k) is given by,

$$S_E(k) = | [E(k)]^2$$

When S_E(k) is plotted against the relative base location k, we get a peak at locations N/3for coding regions and no such intense peak is observed at other regions. This is called period-3 property [13] and by using this we can estimate the probable coding regions in the given DNA sequence. All such peaks are observed in specific window of the sequence whose standard and optimal length [12] (window size) has been taken as 351 and same is implemented in these methods.

2.2 Autoregressive (AR) model:

The most important kinds of digital filter that are used in many digital processing techniques are known as finite-impulse-response (FIR) filters and autoregressive filters. The AR modeling can be utilized in the frequency domain as spectral matching problem. The auto regressive modeling approach is used in biomedical signal processing for the analysis of physiological signals like human electroencephalogram (EEG) and speech recognition [13].

The operation of the AR filter is shown in Fig 1. [14], can be described in terms of polynomial arithmetic division. Specifically when a finite sequence is filtered by an AR filter with zero initial condition, then the output sequence corresponds to the quotient polynomial under polynomial division, whose coefficients are the weights.

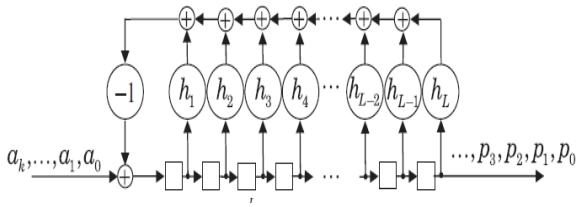

Fig 1: AR Model [14]

If the output E (n) is the function of both the 'L' most recent outputs say E(n-1),E(n-2), ….. E(n-L) and 'a_n' is the present input of the AR filter and by appropriate choice of the signs of the weights hi is given by

$$E(n) = -\sum_{i=1}^{L} h_i\, E(n-i) + a_n$$

The AR model described can be specified in the frequency domain by taking the Z-Transform of the original expression given above. Then the transfer function of the AR model is given by

$$H(z) = \frac{1}{1 + \sum_{i=1}^{L} h_i Z^{-k}}$$

The frequency response H(w) is determined by evaluating the H(z) along the unit circle in the Z-plane,

Where $Z = e^{jwT}$, for a sample period of T. And the spectrum of the output sequence S_E(k) is given by,

$$S_E(k) = \frac{\sigma^2}{|1+\sum_{i=1}^{L} h_i Z^{-k}|^2}$$

$$=\frac{\sigma^2}{|1+\sum_{i=1}^{L} h_i e^{j2\pi ik/N}|^2}$$

where σ^2 is the variance of the input signal. This above equation is used to predict the protein coding region of the given DNA sequence at frequency w = $\frac{2\pi}{3}$ or at N/3 point.

2.3 Adaptive AR Model:

An adaptive autoregressive (AAR) focuses the problem of non-stationary spectral analysis. Adaptively estimated autoregressive parameters are useful in many applications like online spectral analysis of heart rate variability [15] and ECG based brain computer interfacing [16]. For achieving on line prediction of gene and exon the computational time needs to be reduced. Further the fixed AR method requires all data to be available simultaneously which is not always feasible [9]. Hence considering these issues as limitations of AR model, an adaptive AR model based approach is suggested. The AR process can be viewed as an adaptive prediction error filter (all-zero filter) that adaptively adjusts its coefficients to flatten the spectrum of the signal to be observed. It is a fact that with a proper learning algorithm like LMS and RLS the weight vector of the adaptive prediction error filter converges to optimal AR coefficients.

2.3.1 The LMS prediction error filter

A signal E(n) is modeled as a L order AR process[10] can be expressed as

$$\overline{E}(n) = \sum_{i=1}^{L} h_i(n) E(n-i) + e(n)$$

where e(n) is the prediction error and h_1,h_2, · · ·h_L are AR coefficients. The LMS prediction error filter is used to adaptively estimate the optimal AR coefficients by minimizing the mean square value of prediction error.

The weight update equation of the filter is given by

$$h(n) = h(n-1) + \mu \overline{E}(n) e(n)$$

where $e(n) = E(n) - \widehat{E}(n)$ is the prediction error

$\widehat{E}(n) = E^T(n)\, H(n)$ is the prediction of E(n),

$\overline{E}(n)$= [E(n-1)E(n-2)E(n-3)…E(n-L)]

H(n) = [h_1(n)h_2(n) · · ·h_L(n)] and μ is the step size that determines the rate of converge and stability of weights which lies between 0 to 1.

The power spectra is estimated using the prediction error filter is given as

$$S_E(k) = \frac{\sigma_e^2}{|1 - \sum_{i=1}^{L} h(L)\, e^{j2\pi ik/N}|^2}$$

where σ_e^2 is the variance of the prediction error signal. Hence the power spectrum at frequency (N/3) is computed within a

window and the window is slides along the DNA sequence to identify the protein coding regions.

The flow of the system is shown in the block diagram given in the Fig 2.

3. RESULTS & DISCUSSIONS:

The authors have estimated the power spectrum of the protein coding regions of different species using the conventional DFT, AR model and AAR models by considering the all sequences as EIIP mapping sequences. All these three methods are implemented in MATLAB. Good peaks at relatively right locations (near N/3) for the optimal window size of 351 were observed for the three methods. The result of the three methods (DFT, AR model and AAR model) for the benchmark gene of C-elegance with accession numbers AF099922 shown in Fig 3, Fig 4, and Fig 5 respectively. The effective computational time (CPU time) is abruptly reduced in AAR model compared to the DFT and AR methods. Also relatively the peaks at N/3 point are predicted for the stipulated window size. The comparative result of all the sequences of C-elegance [FO080874.2, FO081434] and other genes of fruit fly [NM_170135] and homosapien (BDNF) [NG_011794] are shown in the Table 2.

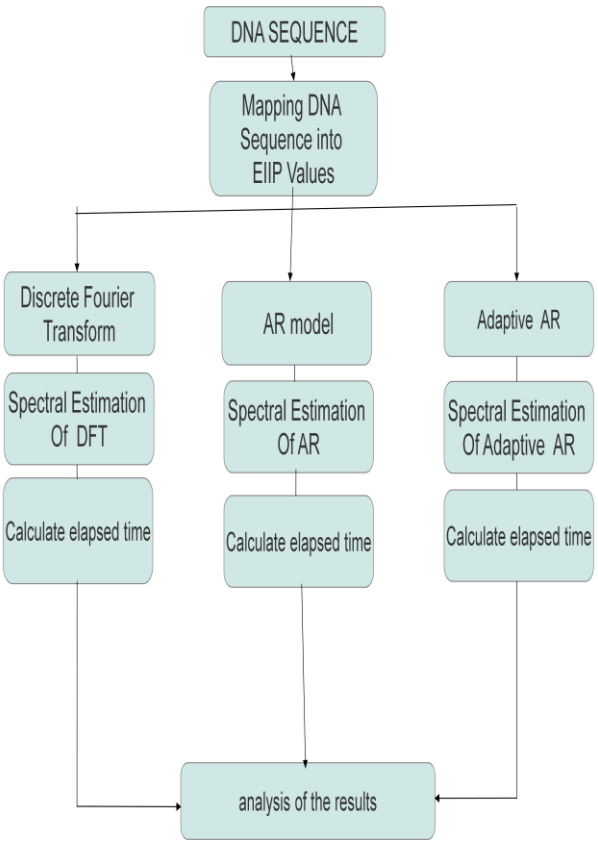

Fig 2: Block diagram

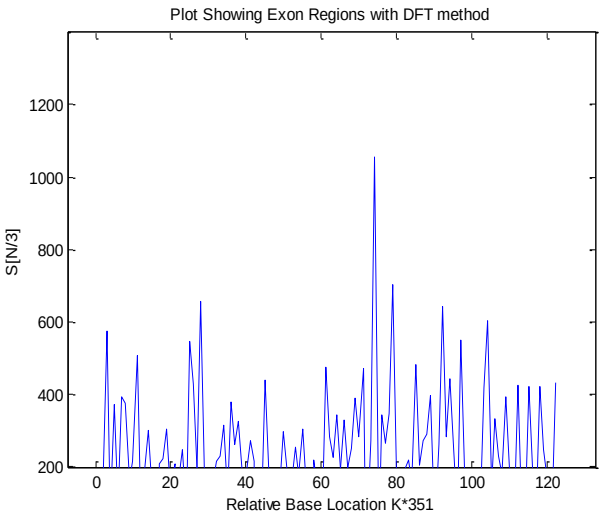

Fig 3: Exon regions with DFT method

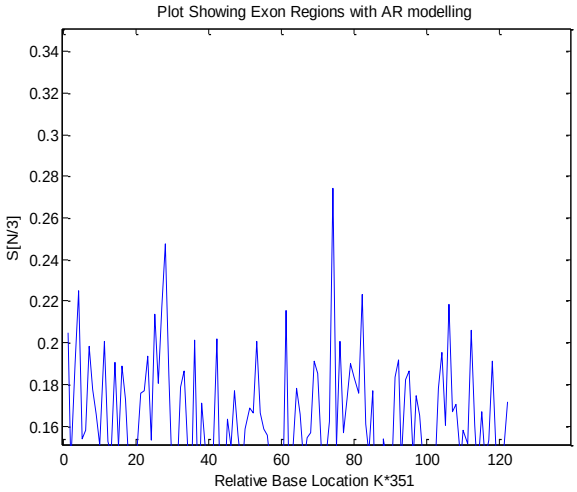

Fig 4: Exon regions with AR modeling method

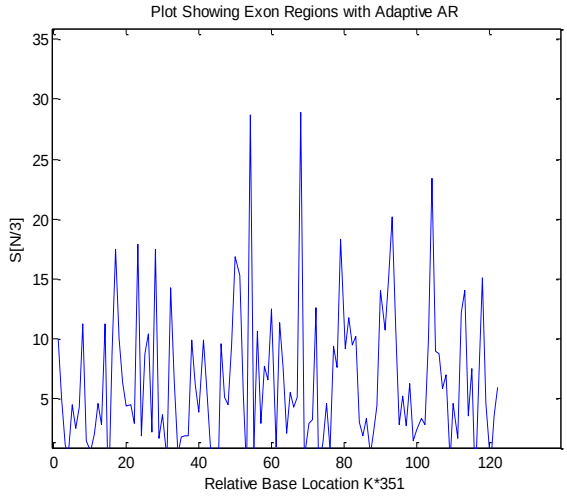

Fig 5: Exon regions with Adaptive AR method

4. CONCLUSION

With these results presented in this paper we conclude that, for the prediction of protein coding region or localization of exonic region by using Genomic Signal Processing methods are more suitable comparing to any other statistical models. The implementation of AR or AAR methods with the EIIP indicator sequence has given discrimination results with less computational Complexity than the binary indicator sequence method. Also the possibility of applying AAR approach with RLS, recursive AR and Kalman filtering may give good results. Finally we hope that with the amalgamation of EIIP sequence method with standard DSP and statistical approaches can give a thrust in this area of research and these EIIP values of nucleotides and amino acids will be used in many gnomic and proteomic applications like prediction of localization of Hotspots etc.

Table 2: Comparative results of the three methods

S No	Species	Gene Accession Number	No of bases	CPU Time(sec) in DFT	CPU Time(sec) in AR (order 25)	CPU Time(sec) in AAR (order 25 ,step size 0.002)
1	C-elegans	AF099922	42799	18.9174	5.6565	4.047
2	C-elegans	FO080874.2	37019	16.2497	3.9976	3.4442
3	C-elegans	FO081434	42981	18.7648	5.6447	3.969
4	Fruit fly	NM_170135	1776	2.2404	0.51507	0.60669
5	Homosapien	NG_011794[BDNF]	74164	32.267	25.4352	6.6015

5. REFERENCES

[1] Roy, M., S. Biswas and S. Barman, 2009. Identification and Analysis of Coding and Noncoding Regions of a DNA Sequence by Positional Frequency Distribution of Nucleotides (PFDN) Algorithm, 4th International Conference on Computers and Devices for Communication, pp: 1-4.

[2] Akhtar, M., E. Ambikairajah and J. Epps, 2008. Advances in Eukaryotic Gene Prediction, IEEE Journal of Signal Processing in Sequence Analysis, Selected Topics in Signal Processing, 2(3): 310-321.

[3] Voss RF. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequence. Physical Review Letters 68: 3805-3808.

[4] I. Cosic, IEEE Trans Biomed Eng., 41:12 (1994), [PMID: 7851912]

[5] Hota, M.K. and V.K. Srivastava, 2008. DSP technique for gene and exon prediction taking complex Indicator sequence, IEEE Region 10 Conference (TENCON) pp: 1-6.

[6] Achuthsankar S. Nair, Sivarama Pillai Sreenadhan A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), ISSN 0973-2063,2006, Bioinformation.

[7] Rao N, Shepherd SJ: Detection of 3-periodicity for small genomic sequence based on AR techniques. In Int. Conf. On comm., IRC. And Sys, Volume 2 2004:1032–1036.

[8] Akhatar M, Ambikairajah E, Epps J: Detection of Period-3 behavior in genomic sequence using singular value decomposition. In IEEE Int. Conf. On Emerging Technologies 2005:13–17.

[9] Sitanshu Sekhar Sahu, Ganapati Panda, An efficient signal processing approach in eukaryotic gene prediction. Vol.1-2010/Iss.2, pp. 75-79, IJSIP

[10]National Center for Biotechnology Information, website address available at: http://www.ncbi.nlm.nih.gov/.

[11] V.Veljkovic & I.Slavic, (1972) "Simple General-Model Pseudopotential", Physical ReviewLetters, Vol. 29, No. 2, pp 105-107.

[12] S. Tiwari et al., "Prediction of probable genes by Fourier analysis of genomic sequences," CABIOS, vol. 13, no. 3, 1997. [PMID: 9183531]

[13]J.Pardey,S.Roberts,L.Tarassenko, "A review of parametric modeling techniques for EEG analysis. University of Oxford, Oxford.

[14]Richard E.Blahut, "Fast Algorithms for Signal Processing",pp:10-17,Cambridge University Press.

[15] Bianchi A., Mainardi L., Meloni C., Chierchia S., Cerutti S. Continuous monitoring of the Sympatho-Vagal Balance through spectral analysis. IEEE Engineering in Medicine and Biology. 16(5): 64-73, 1997.

[16]Pfurtscheller G, Neuper C, Schlögl A, Lugger K. Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. IEEE Trans Rehabil Eng. 6(3):316-25, 1998.

[17] A. S. Nair & T. Mahalakshmi, In Silico Biology, 6: 0019 (2006) [PMID: 16922684]

[18] Haykin S. Adaptive Filter Theory. Prentice Hall, Englewood Cliffs, NJ, 1996.

[19] T. K. Attwood and D. J. Parry-Smith, An Introduction to Bioinformatics, Addison Wesley Longman

[20] David W. Mount, Introduction to Bioinformatics, Cold Spring Harbor Press.

[21] Vinay K.Ingle , John G. Proakis, Digital Signal Processing Using MATLAB.