ANNALS OF FACULTY ENGINEERING HUNEDOARA
- INTERNATIONAL JOURNAL OF ENGINEERING
Tome IX (Year 2011). Fascicule 3. (ISSN 1584 - 2673)

[1] M. ROY, [2] S. BARMAN (MANDAL)

SPECTRAL ANALYSIS OF CODING AND NON-CODING REGIONS OF A DNA SEQUENCE BY PARAMETRIC AND NONPARAMETRIC METHODS: A COMPARATIVE APPROACH

[1] THE CALCUTTA TECHNICAL SCHOOL, GOVT. OF W.B., KOLKATA-700013, INDIA
[2] INSTITUTE OF RADIO PHYSICS & ELECTRONICS, UNIVERSITY OF CALCUTTA, KOLKATA-700 009, INDIA

ABSTRACT: Identification and analysis of hidden features of coding and non-coding regions of DNA sequence is a challenging problem in the area of genomics. The objective of this paper is to estimate and compare spectral content of coding and non-coding segments of DNA sequence both by Parametric and Non-parametric methods. Consequently an attempt has been made so that some hidden internal properties of the DNA sequence can be brought into light in order to identify coding regions from non-coding ones. In this approach the DNA sequence from various Homo Sapien genes have been identified for sample test and assigned numerical values based on weak-strong hydrogen bonding (WSHB) before application of digital signal analysis techniques. The statistical methodology applied for computation of Spectral content are simple and the Spectrum plots obtained show satisfactory results.

KEYWORDS: Parametric, Non-parametric, Periodogram, Auto-regressive, Spectral content, codon

❖ INTRODUCTION

It has been observed that in most cases data sequence vary with time but in few situations data may vary with location points in space. Though in DNA sequences the variation is in position of nucleotide bases, it is treated as a time-series signal. From point of view of statistics such sequences are termed as Categorical time series [17]. Recently researchers from various cross-fields have concentrated in the field of DNA sequence analysis in order to extract the vast information content hidden in it [14],[1],[2]. Coding region of a DNA sequence being the actual information bearing part proper identification and study of this part is of prime importance as no full proof algorithm is yet available which is universally applicable to all data bases giving accurate results. Today Digital Signal Processing (DSP) plays an important role in this effort. In this paper Power Spectral Density of coding and non-coding regions of DNA sequences have been estimated by Parametric and Non-Parametric methods and an attempt has been made to compare and differentiate coding regions from non-coding ones. DNA (de-oxyribo-nucleic acid) is a huge data base available to us in Public Domain having hereditary traits hidden in it [19]. Genetic information is stored in the particular order of four kinds of nucleotide bases, Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) which comprises the DNA molecule along with Sugar-Phosphate backbone. There are two complementary DNA chains twisted around one another in a right handed double helix structure. A straightened form of a DNA helical structure has been shown in Figure1.

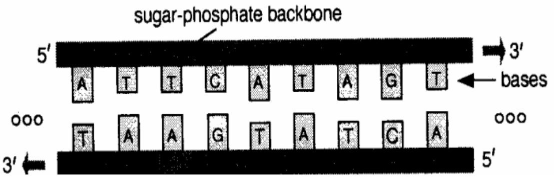


Figure 1. DNA double helix structure

Adenine (A) of one strand always pairs with Thymine (T) of opposite strand while Guanine (G) always pairs with Cytosine (C). DNA base sequence is always written from 5' end to 3' end which is called polarity of DNA chain.. Two types of nitrogen bases purine and pyrimidine are present in DNA. 'A' and 'T 'are purine bases where as 'C' and 'G' are pyrimidine bases. The DNA strands are held together mainly by hydrogen bonds between bases. There are two hydrogen bonds between 'A' and 'T' while three hydrogen bonds between 'C' and 'G'. Hence 'C' and 'G' bonds are stronger than 'A' and 'T' bonds as depicted in Figure 2 & Figure 3.

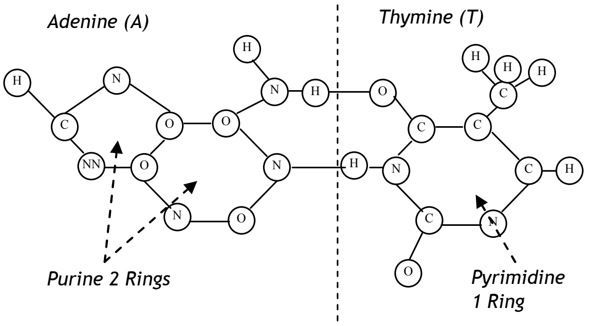


Figure 2. Double Hydrogen Bond A = T Signifies Weak bond

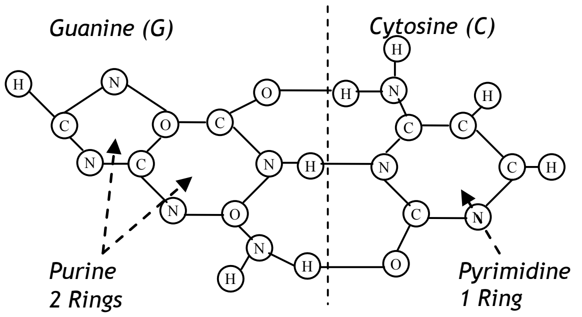


Figure 3. Triple Hydrogen Bond C Ξ G Signifies strong bond

The DNA sequence can be divided into genes and inter-genic spaces. The genes can again be subdivided into exons (coding region) and introns (non-coding region). Even though all the cells in an organism have identical genes only a selected subsets are activated in any family of cells. Exons of a DNA sequence are the most information bearing part because only the exons take part in protein coding while the introns are spliced off during protein synthesis.

Gene prediction refers to detecting locations of the protein coding regions of genes in a long DNA sequence. There has been great deal of work done in applying Digital Signal Processing and Statistics methods to DNA in the recent past, some of which are mentioned here in a nutshell. It has been established that base sequences in the exon regions of DNA molecules exhibit a period-3 property because of the codon structure involved in the translation of nucleotide bases into amino acids [4],[12],[16]. Investigation into long range correlation has also been the focus of attention for many researchers [15],[8],[7]. A coding measure scheme employing electron-ion-interaction pseudo-potential (EIIP) was presented as a revision for binary indicator sequences [9].Implementation of digital filters to extract period-3 components and effectively eliminate back ground 1/f noise present in DNA sequence has given good results [14],[13].Positional Frequency Distribution of nucleotides has also given interesting results [11]. In this article the authors have presented methods for identifying coding and non-coding regions of DNA sequence based on graphical representation of PSD plots using low order Auto Regressive Yule Walker Algorithm.

❖ DSP TECHNIQUES FOR SPECTRAL ESTIMATION OF DNA SEQUENCES

There is vast genomic data available in the NCBI Genbank and DSP can be used as an effective tool for analysis of this data. DSP technique is applicable only to numerical data but genomic data consists of four alphabets A,T,C and G. Hence a mapping technique is required to convert the 4-letter alphabet sequence into numerals before applying DSP techniques. Different researchers have adopted different mapping methods for this purpose. Here the authors have attempted applying a new mapping rule based on weak-strong hydrogen bonding for digitization. As nucleotides ‘A’ and ‘T’ have two hydrogen bonds in their molecular structure they have been treated as weak bond and assigned integer value ‘2’. Nucleotides ‘C’ and ‘G’ have three hydrogen bonds so they are treated as strong and have been assigned integer value ‘3’.

For example a DNA sequence of length N:

$$x[n] = [A\ T\ G\ C\ C\ T\ T\ A\ G\ G\ A\ T] \tag{1}$$

After mapping:

$$x_{sw}[n] = [2\ 2\ 3\ 3\ 3\ 2\ 2\ 2\ 3\ 3\ 2\ 2] \tag{2}$$

This method is employed to the data sequence for parametric and non-parametric analysis of DNA sequences.

❖ NON-PARAMETRIC ANALYSIS OF DNA SEQUENCE

Non-parametric technique of spectrum estimation is based on the idea of first estimating the auto-correlation of data sequence and then taking its Fourier Transform to obtain its Power Spectral Density (PSD).This method also known as Periodogram method was first introduced by Schuster in 1898 in his study of periodicities in sunspot numbers. Although periodogram is easy to compute it is limited in its ability to produce an accurate estimate of the power spectrum, particularly for short data records. For improvement of statistical property of periodogram method a variety of modifications have been proposed such as Barlett’s method, Welch’s method and the Blackman-Tukey method. In periodogram method PSD is estimated directly from signal itself [10]. The Fourier Transform of the estimated auto-correlation of data sequence is given by the following equation:

$$P_x(e^{jw}) = \sum_{k=-\infty}^{\infty} r_x(k) e^{-jwk} \tag{3}$$

The estimated auto-correlation function:

$$r_x(k)=\frac{1}{N}\sum_{n=0}^{N-1-k}x(n+k)x^*(n)\,,\qquad \text{where}\quad k=0,1,2\ldots\ldots\ldots N-1\ . \tag{4}$$

with $r_x(k)$ set equal to 0 for $|k| \geq N$.

With values of $r_x(k)$, for k<0 defined using conjugate symmetry as: $r_x(-k)= r_x^*(k)$

Taking DTFT of $r_x(k)$ leads to the estimate of the power spectrum known as Periodogram.

$$P_{per}(e^{jw})=\sum_{k=-N+1}^{N-1}r_x(k)e^{-jwk} \tag{5}$$

❖ Parametric Analysis of DNA sequence

The Parametric method uses a different approach to Spectral estimation. Instead of estimating PSD from data directly as is done in non-parametric method, it models the data as output of a linear system driven by white noise and attempts to estimate parameters of this linear system. The most frequently used linear system model is the all pole model, a filter with all of its zeroes at the origin on the z-plane. The output of such a filter for white noise input is an AR process, known as AR method of spectral estimation. There are different types of AR methods such as Burg method, Covariance and Modified Covariance method, Yule-Walker (auto-correlation) method etc. The advantage of Yule-Walker Autoregressive method is that it always produces a stable model. Parametric methods can yield higher resolution than non-parametric methods when the signal length is short [18],[3],[6].

As already stated the signal spectrum estimated in parametric method is based on the PSD of a linear system driven by white noise. The output of such a system with white noise input referred to as Autoregressive (AR) process has been implemented here (Figure 4).

The pth order power spectrum of Auto Regressive process is given by:

$$P_{AR}(e^{jw})=\frac{|b(0)|^2}{|1+a_p(k)e^{-jwk}|^2} \tag{6}$$

where b(0) and $a_p(k)$ are estimated from given data.

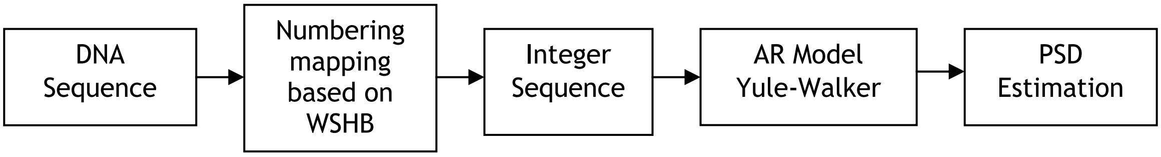


Figure 4. Block diagram realization of an AR model PSD estimation system

Here Yule-Walker Autoregressive method has been implemented efficiently for Parametric Analysis of DNA sequence. AR models are popular because with them an accurate estimation of PSD can be obtained by solving linear equations. Since in above equations $|b(0)|^2$ is constant, the only value that are needed for calculating the shape of PSD are the coefficients $a_p(k)$. Though there are various methods to find these coefficients, the Yule-Walker (auto-correlation) method has been used here for its simplicity.

❖ Results and Discussion

In this paper the authors have presented a comparative study of Spectral content of DNA sequences both by Periodogram and Yule-Walker Autoregressive methods. The nucleotide bases from coding and non-coding segments of various Homo Sapien genes have been used separately as raw data for spectrum analysis. It has been observed that Yule-Walker PSD spectrum plots are smooth, distinct and devoid of any spurious noise component compared to the plots obtained by non-parametric method. It is also evident from the plots that in case of Yule-Walker Autoregressive Power Spectrum estimation method the frequency resolution is independent of number of databases. Where as Classical Periodogram methods show that the frequency resolution deteriorates as the number of nucleotide databases increase Figure 5.

The other advantage of parametric method over periodogram method is that the spectrum can be interpreted easily from its graphical plots and it is free from spectral leakage. The limitation on part of Periodogram method is that the accuracy of PSD estimation depends mainly on accuracy of estimated auto-correlation. The power spectrum resolution is better than parametric method only if $r_x(k)\approx 0$ for $k| > 0$. Several genes from Homo Sapiens have been taken into consideration for this comparative study. Some of the values have been tabulated in Table-1 and plotted.

In this paper the authors have succeeded in bringing out special spectral characteristics in PSD plot of exon region as compared to intron region when low order Autoregressive PSD estimation method is applied to nucleotide databases. The two sided symmetric plots of PSD vs normalized frequency show that exon regions show a valley where as introns exhibits distinct peak at the centre of the frequency axis. This spectral signature is retained in all the databases from Homo Sapien genes that have been analyzed Figures. 6 & 7.

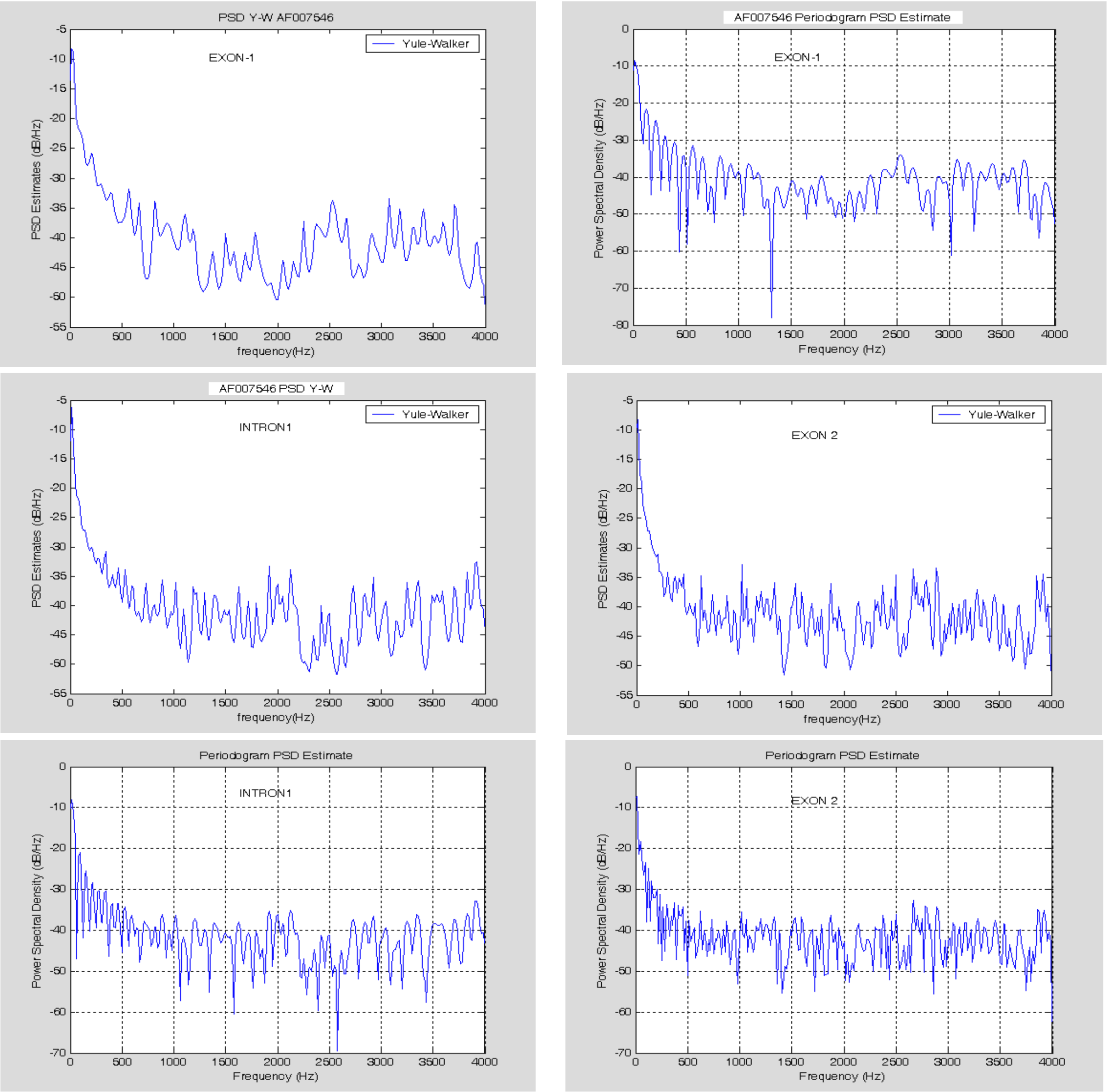


Figure 5. Accession No. 007546 Homo Sapiens Beta Globin Gene Parametric (Yule-Walker Plots) and Non-Parametric (Periodogram) Power Spectral Density Plot for Exons & Introns

Table 1. Average Std. Deviation/ Variance/ Normalized Variance values for various Accession numbers of Homo Sapien genes by parametric and non-parametric methods.

S.N.				Non-Parametric Periodogram			Parametric Yule-Walker AR		
	Accession No.	EXON INTRON	Bp length (k)	Avg.Std. Dev^n. ∂_x	Variance ∂_x^2	Var{m_x} =1/k ∂_x^2	Avg.Std. Dev^n. ∂_x	Varianc e ∂_x^2	Var{m_x} =1/k ∂_x^2
1	L26462.1	E	444	7.187	51.653	.1163	6.648	44.196	0.0995
		I	980	6.7615	45.7179	.0466	6.293	39.60	0.0404
2	AF083883	E	344	7.496	56.19	.1633	6.67	44.489	0.129
		I	980	6.793	46.145	.047	6.296	39.639	0.0404
3	AF007546	E	748	7.187	51.653	.069	6.369	40.564	0.054
		I	980	6.573	43.204	.044	6.286	39.514	0.040
4	D13156.1	E	354	7.439	55.339	.156	6.90	47.61	0.1344
		I	859	6.74	45.427	.053	6.278	39.413	0.459
5	M62420.1	E	834	7.234	52.33	.0627	6.983	48.762	0.0584
		I	1933	6.78	47.97	.0237	6.252	39.087	0.0202
6	AF015224	E	282	8.211	67.42	.239	6.85	46.922	0.166
		I	2477	6.476	41.938	.0169	6.122	37.478	0.015
7	AF065986	E	789	7.830	61.309	.0777	7.06	49.84	0.0633
		I	891	7.39	54.61	.0612	6.584	43.349	0.0486

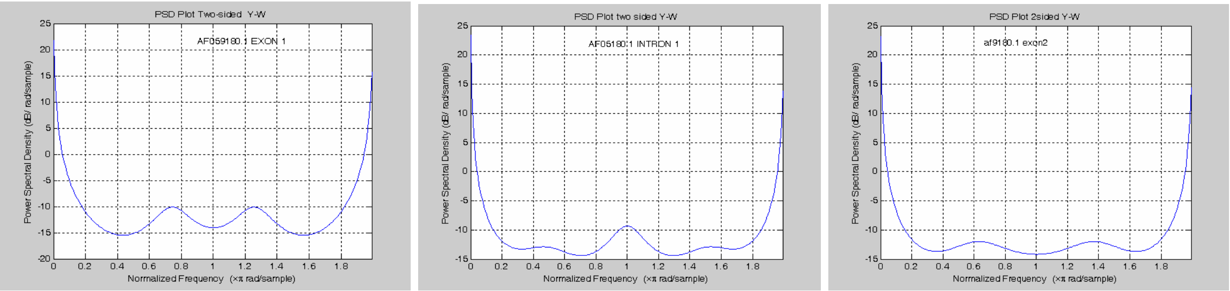
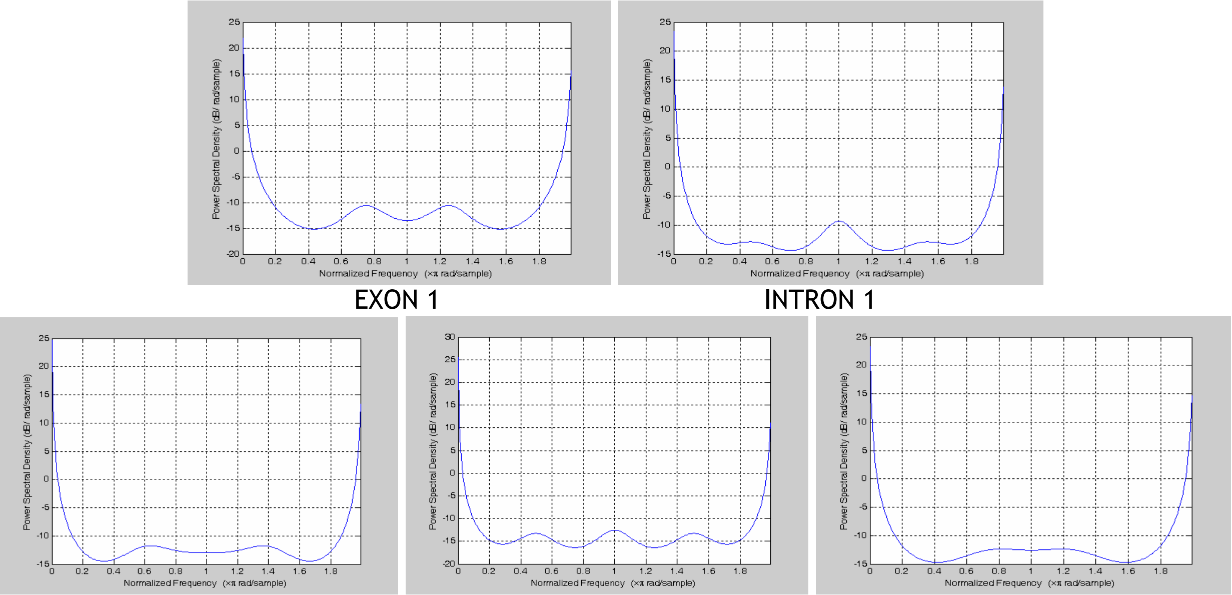


Figure 6. Accession No.AF083883 Low Order AR Yule Walker PSD Plot for Exon 1,Intron 1 and Exon 2.



EXON 1

INTRON 1

EXON 2

INTRON 2

EXON 3

Figure 7. Accession No.AF007546 Low Order AR Yule Walker PSD Plots for Exon 1, Intron 1, Exon 2, Intron 2, and Exon 3

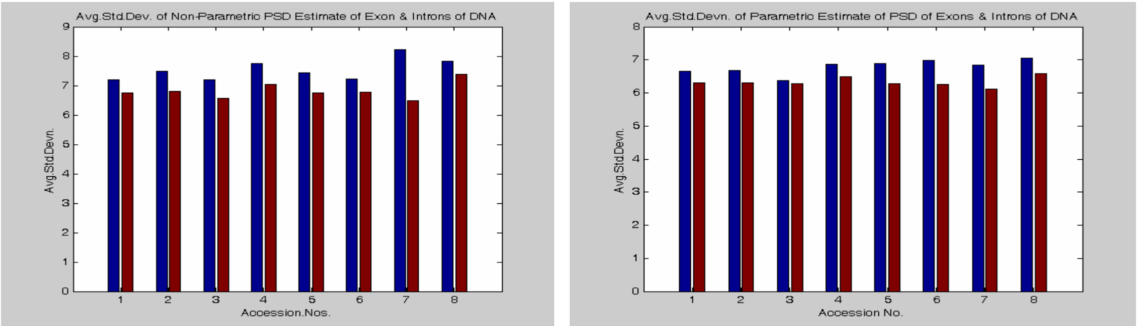


Figure 8. Bar plot of Average Standard Deviation vs. Accession nos.
(by non-parametric and parametric methods)

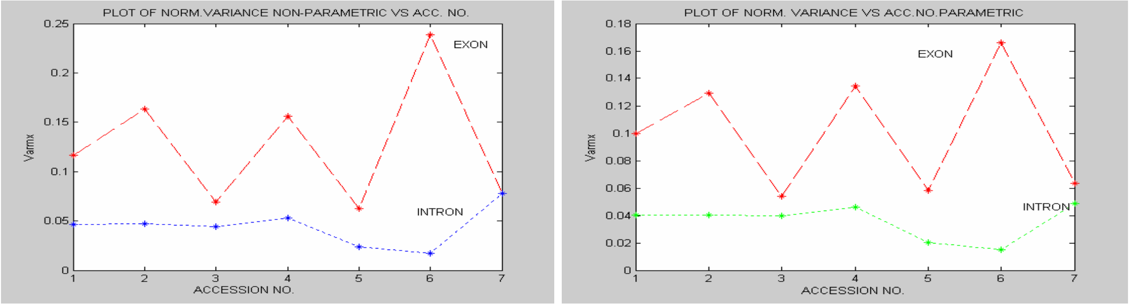


Figure 9. Plot of Normalized variance vs various Accession No. of Homo Sapien genes
for exons and introns (non-parametric and parametric)

Computation of statistical values such as standard deviation, variance and mean executed in course of PSD estimation of exons and introns by parametric as well as non-parametric methods also show distinct results. The bar plot for average standard deviation shows that its value is always larger in case of exons than that of introns for both parametric and non-parametric methods but the results from latter method are more prominent Figure 8. The plots of normalized variances for various Accessions Nos. exhibit similar results Figure 9.

❖ CONCLUSIONS

We are well aware that stochastic or random signals have information bearing Power Spectral Density. Parameterization of a stochastic signal for efficient representation of this information is already in use for speech coding and various other biomedical signal processing applications. In this paper the authors have applied parametric as well as non-parametric Power spectrum estimation techniques to coding and non-coding regions of DNA sequence taken from Homo Sapien genes. A comparative study has been organized in order to distinguish coding from non-coding regions. The non-parametric Power Spectral estimation method is methodologically straight forward and computationally simple. But in case of low Signal to Noise Ratio (SNR) spectral features are difficult to be distinguished and noise artifacts appear in spectral estimates. Parametric spectrum estimate methods have more statistical consistency even on short data segments. Among several parametric models available, AR models presented here are popular because they provide accurate estimation of PSD by solving linear equations. Data sets from various Homo Sapien genes have been investigated. It has been observed from the plots of average PSD for low order Auto Regressive Yule Walker method that the spectral signatures of exons bear a significant pattern as compared to that of introns. The normalized variance values show strong periodicities in case of exons for parametric and non-parametric methods where as introns do not reveal any such property. Future course of investigation may be steered towards other parametric methods such as Burg, Covariance, Maximum Entropy etc. Genes from other species may also be taken into consideration.

❖ REFERENCES

[1.] Anastassiou D., “Frequency–domain analysis of biomolecular sequences”, Bioinformatics 16, 1073-1081.

[2.] Anastassiou D., “DSP in genomics: Processing and frequency domain analysis of character strings,” IEEE,0-7803-7041-2001.

[3.] Chakrabarty Niranjan, Spanias A., Lesmidis L.D. and Tsakalis K., “Autoregressive Modeling and Feature Analysis of DNA Sequences”, EURASIP Journalon Applied Signal Processing 2004:I, 13-28.

[4.] Ficket J.W. and Tung C.S., “Recognition of protein coding regions in DNA sequences”, Nucleic Acids Research, Vol.10, No.17, pp.5303-5318, July 1982.

[5.] Fickett J.W. and Tung C.S., “Assessment of protein coding regions in DNA sequences”, Nucleic Acid Res, 10, 5303-5318, 2000.

[6.] Hayes M.H., “Statistical digital signal processing and modeling”, John Wiley & Sons, Inc., New York, USA, 1996.

[7.] Herzel H. and Grobe B., “Measuring correlations in symbol sequences”, Phys. A, Vol. 216, pp.518-542, 1995.

[8.] Li W. Kaneko & K, “Long range correlation and partial 1/f spectrum in a non- coding DNA sequence,” Europhys. Lett., Vol.17, No.7, pp. 655-660, January1992.

[9.] Nair Achuthsankar.S. and Sreenadhan S., “A coding measure scheme employing electron-ion-interaction pseudopotential (EIIP)”, Journal of Bioinformation V.1(6);. 2006; I(6);197-202.

[10.] Nair Achuthsankar. S. and Mahalaxmi T., “Are Categorical periodograms and Indicator sequences of genomes spectrally equivalent?” In silico Biology 6, 0019 (2006) Bioinformatic Systemse, v.

[11.] Roy M., Biswas S. and Barman (Mandal), S. “Identification and analysis of coding and non-coding regions of a DNA sequence by Positional Frequency Distribution of Nucleotides (PFDN) algorithm” in International Conference on Computers and Devices for Communication CODEC-09), December 14-15, 2009.

[12.] Tiwari S., Ramachandran S., Bhattachary A., Bhattacharya S,. and Ramaswamy R., “Prediction of probable genes by fourier analysis of genomic sequences,” CABIOS, vol 3,no.3.263-270,1997.

[13.] Tuqan J. and Rushdi A., “A DSP based approach for finding the codon bias in DNA sequences”, IEEE journal on signal processing, vol.2.No. 3, June, 2008.

[14.] Vaidyanathan P.P. and Yoon B.J., “The role of signal-processing concepts in genomics and proteomics”, Journal of the Franklin Institute, special issue on Genomics, 2004.

[15.] Voss R.F., “Evolution of long-range fractal correlations and 1/f noise in DNA base sequences”, Phy.Rev.Lett., Vol.68, No.25 , pp.3805-3808, 1992.

[16.] Yin C., Stephen S. and Yau T., “Prediction of protein coding regions by the 3 base periodicy analysis of a DNA sequence”, Journal of Theoretical Biology 247, 687-694, 2007.

[17.] Zhao Lan., “Application of Specral Analysis to DNA sequences“, CSD TR #06-003, January 2006.

[18.] Zhou H. and Yan Hong, “Auto Regrssive Models for Spectral Analysis of short tandem repeats in DNA sequences”, 2006, IEEE International Conference on Systems, Man and Cyberne, Oct.8-11, 2006, Taipei, Tiwan. tics.

[19.] www.ncbi.nlm.nih.gov/

ANNALS OF FACULTY ENGINEERING HUNEDOARA
– INTERNATIONAL JOURNAL OF ENGINEERING
copyright © University Politehnica Timisoara,
Faculty of Engineering Hunedoara,
5, Revolutiei, 331128, Hunedoara,
ROMANIA
http://annals.fih.upt.ro