Journal of Engineering and Applied Sciences 8 (6): 198-207, 2013
ISSN: 1816-949X

High Resolution Frequency Estimation by Minimum Norm Solution for Effective Gene Prediction

[1]M. Roy and [2]S. Barman
[1]The Calcutta Technical School, Govt. of West Bengal,
110, S.N. Banerjee Road, 700013 Kolkata, India
[2]Institute of Radio Physics and Electronics, University of Calcutta,
92 A.P.C. Road, 700-009 Kolkata, India

Abstract: The recent techniques of spectrum estimation are based on linear algebraic concepts of subspaces. In this study, the researchers have used noise subspace method for finding hidden periodicities in DNA. With the vast growth of genomic sequences, the demand to identify accurately the protein coding regions in DNA is increasingly rising. In the past, several techniques involving various cross-fields have come up, among which application of digital signal processing tools is of prime importance. It is known that coding segments have a 3-base periodicity while non-protein coding regions do not have this unique feature. One of the most important spectrum analysis technique based on the concept of subspace is the minimum norm method. The minimum norm estimator developed in this study shows sharp period-3 peaks in coding regions completely eliminating background noise. Comparison of proposed method with existing Sliding Discrete Fourier Transform (SDFT) method popularly known as periodogram has been drawn on several genes from various organisms showing that the proposed method has effective approach towards gene prediction. Resolution, quality factor, sensitivity, specificity, miss rate, wrong rate and computation time are used to establish superiority of minimum norm gene prediction method over existing method.

Key words: Periodogram, de-oxyribo nucleic acid, minimum norm solution, eigen-vector, eigen value

INTRODUCTION

It is a well known fact that the most significant scientific and technological endeavour of 21st century is related to genomics. Therefore, researchers from various cross-fields have concentrated in the field of genomic analysis in order to extract the vast information content hidden in it. DNA (De-oxyribo Nucleic Acid) is the hereditary material present in all living organisms. In eukaryotic organisms, genes (sequences of DNA) consist of exons (coding segments) and introns (non-coding segments). It has been established that genetic information is stored in the particular order of 4 kinds of nucleotide bases: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) which comprise the DNA bio-molecule along with sugar-phosphate backbone. 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. Since, DNA codes information of proteins, various statistical and computational techniques have been explored to study the information content carried by DNA and distinguish the exons from introns.

Genomic information is discrete in nature because it is made up of a finite number of nucleotides in the form of alphabets. Digital Signal Processing (DSP) techniques can be used as an effective tool to analyze the DNA in order to capture its periodic characteristics. Estimation of spectrum of discretely sampled processes is generally based on procedures employing the Fast Fourier Transform (FFT). This approach is computationally efficient and produces reasonable results but inspite of the advantages, it has certain performance limitations. The most important limitation lies in its frequency resolution. Moreover, spectral estimation by Fourier method generates various harmonics which often lead to false prediction of coding regions. Among the recently introduced techniques an eigen decomposition based method known as the minimum norm solution is found to be of great interest. The researchers in this study addressed the problems posed by standard FFT method and proposed a minimum norm algorithm

Corresponding Author: S. Barman, Institute of Radio Physics and Electronics, University of Calcutta, 92 A.P.C. Road, 700-009 Kolkata, India

based on the concept of subspace frequency estimation for efficient prediction of coding regions in DNA sequence.

Application of digital signal processing methods for finding periodicities in DNA sequences has been explored by various researchers (Anastassiou, 2000, 2001; Vaidyanathan and Yoon, 2004; Zhao, 2006). It is established that 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 (Fickett, 1982; Tiwari *et al*., 1997; Yin and Yue, 2007). Peng *et al.* (1995) discussed in their study the statistical properties of gene. Implementation of digital filters to extract period-3 components and effectively eliminate background 1/f noise present in DNA sequence has given good results (Nair and Sreenadhan, 2006; Tuqan and Rushdi, 2008; Sahu and Panda, 2011). Roy *et al.* (2009) introduced Positional Frequency Distribution of Nucleotides (PFDN), an algorithm for prediction of coding regions. Parametric techniques of gene prediction where autoregressive all-pole models were used for identifying coding and non-coding regions provided better results (Roy and Barman, 2011). An exclusive survey of various gene prediction techniques are presented by Manaswini and Sahu (2010). Fundamental theory of principal component analysis is elaborated by Shlens (2003) and its application is discussed by Ubeyli and Guler (2003).

The researchers in this study have compared and analyzed power spectral peaks obtained by modified periodogram method with that by minimum norm solution method for identification of coding regions in DNA sequence (Hayes, 1996; Haykin, 2008; Stoica and Moses, 2011; Praokis and Manolakis, 2008). The algorithm is tested successfully on several sample databases, especially from Celegan organism. Celegan Cosmid *F56F11.4a* gene from Chromosome-III having Accession number AF099922.1 is presented in detail (NCBI Gen Bank).

MATERIALS AND METHODS

The PSD estimation of DNA sequence requires conversion of DNA character strings into numerical form. Different researchers have adopted different mapping methods for this purpose. Here, the researchers have applied a single sequence quaternary mapping rule assigning numerical values, a = -1, c = - j, g = 1 and t = j. MATLAB 7.1 environment is used to show performance of the estimators.

Spectral estimation by non-parametric method can be classified as direct and indirect. These two methods are equivalent and are popularly known as periodogram method. The direct method takes Discrete Fourier Transform (DFT) of the signal and then averages the square of its magnitude. The indirect method is based on the idea of first estimating the auto-correlation of data sequence and then taking its Fourier Transform (FT).

Spectral analysis by periodogram method: In direct method Periodogram $P_{per}(f_k)$ for signal x (n) can be computed by DFT or more efficiently by Fast Fourier Transform (FFT) for N data points as shown:

$$P_{per}\left(\frac{k}{N}\right)=\frac{1}{N}\left|\sum_{n=0}^{N-1}x(n)e^{-j2\Pi nk/N}\right|^2 \quad (1)$$

Where, $f_k = k/N$, for k = 0, 1, 2, ------ N-1. To improve performance in the periodogram method, first the N-point data sequence is sub-divided into K overlapping segments of length M each then periodogram is computed and averaged with Bartlett windowing.

Spectral analysis by eigen decomposition: Eigen decomposition uses vectors that lie in the signal or noise sub-space. Eigen decomposition of M×M autocorrelation matrix R_x is given as follows:

$$R_x=\sum_{i=1}^{p}\lambda_i v_i v_i^H+\sum_{i=p+1}^{M}\lambda_i v_i v_i^H \quad (2)$$

The set of eigen vectors $\{v_1, v_2, ..., v_p\}$, associated with largest eigen values span the signal subspace and are called principal eigen vectors. The second subset of eigen-vectors $\{v_{p+1}, v_{p+2}, ..., v_M\}$ span the noise subspace and have σ_n^2 as their eigen value. Since, the signal and noise eigen vectors are orthogonal, it follows that the signal subspace and the noise subspace are also orthogonal. After eigen decomposition of the autocorrelation matrix, the eigen values are arranged in decreasing order $\lambda_1 \geq \lambda_2 \geq \lambda_3 .... \geq \lambda_M$ as depicted in Fig. 1. From this plot of eigen values one can distinguish initial steep slope representing signal and a more or less flat floor representing noise level.

There are three generic steps of pseudo-spectrum estimation by noise subspace method:

- Construction of autocorrelation matrix from data vector
- Derivation of noise subspace by eigen decomposition
- Identification of signal components from noise subspace with the help of frequency estimation function

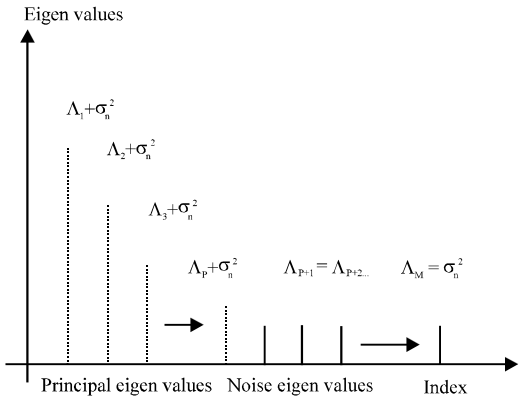

Fig. 1: Decomposition of the eigen values of noisy signal into the principal and noise eigen values

Frequency estimation by minimum norm solution: Frequency estimation is the process of estimating the process of complex frequency components of a signal in the existence of noise. The most common frequency estimation method involves identifying the noise subspace to extract these components (Lobos *et al*., 2000). The minimum norm algorithm developed in this study uses a single vector $\vec{a}$ that is constrained to lie on the noise subspace and the complex exponential frequencies are estimated from the peaks of the frequency estimation function:

$$\hat{P}_{MN}\left(e^{jw}\right)=\frac{1}{\left|\vec{e}^{H}\vec{a}\right|^2} \tag{3}$$

Where, $\{\vec{e}\}$ is an auxiliary vector given by:

$$\vec{e}=\left[1\ e^{jw}\ e^{j2w}\ e^{j3w}...\ e^{j(N-1)w}\right] \tag{4}$$

With $\vec{a}$ constrained to lie in the noise subspace, if the autocorrelation function is known exactly then $|\vec{e}^{H}\vec{a}|^2$ will have nulls at the frequencies of each complex exponentials. Therefore, z-transform of coefficients of $\vec{a}$ may be factored as:

$$A(z)=\sum_{k=0}^{M-1}a(k)z^{-k}=\prod_{k=1}^{p}\left(1-e^{jwk}z^{-1}\right)\prod_{k=p+1}^{M-1}\left(1-z_k z^{-1}\right) \tag{5}$$

Where, z_k for k = p+1 ... M-1 are the spurious roots that do not in general lie on the unit circle. The minimum norm method attempts to eliminate the effects of spurious zeros by pushing them inside the unit circle leaving the desired zeros on the unit circle. The problem, then is to determine which vector in the noise sub-space minimizes the effects of spurious zeros on the peaks of $\hat{P}_{MN}\left(e^{jw}\right)$.

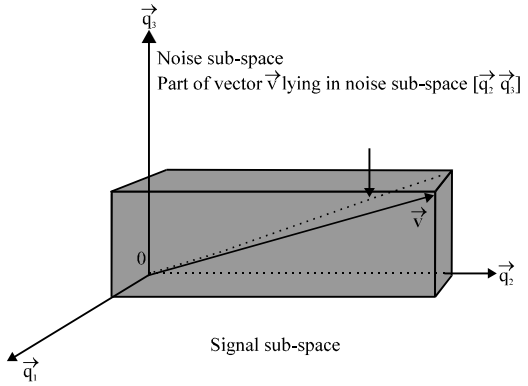

Fig. 2: Projection of signal vector v on noise sub-space in a three dimensional vector space

The approach used in minimum norm algorithm is to find a vector $\vec{a}$ that satisfies the following three constraints:

- The vector $\vec{a}$ lies on the noise sub-space ensuring that p roots of A (z) are on the unit circle
- The vector $\vec{a}$ has minimum Euclidean norm ensuring that spurious roots of A (z) lie inside unit circle
- The first element of $\vec{a}$ is unity, i.e., minimum norm solution is not the zero vector

To solve this constrained minimization problem, researchers begin by noting the constraint that $\vec{a}$ lies on the noise subspace which is given by the following equation:

$$\vec{a}=P_n\vec{v} \tag{6}$$

Where, $P_n = V_n V_n^H$ is the projection matrix that projects an arbitrary vector $\vec{v}$ on the noise subspace (Fig. 2) (Haykin, 2008). Minimum norm method involves projection of signal vector $\vec{v}$ on to the entire noise space. The third constraint is expressed as:

$$\vec{a}^{H}\vec{u}_1=1 \tag{7}$$

Where, $\vec{u}_1$ $[1, 0, 0...0]^T$. This may be combined with the constraint in Eq. 6 giving:

$$\vec{v}^{H}\left(P_n^{H}\vec{u}_1\right)=1 \tag{8}$$

The norm of $\vec{a}$ may be written as:

$$\left\|\vec{a}\right\|^2=\left\|P_n\vec{v}\right\|^2=\vec{v}^{H}\left(P_n^{H}P_n\right)\vec{v} \tag{9}$$

Since, projection matrix P_n is Hermitian, therefore $P_n = P_n^H$ and idempotent, therefore $P_n^2 = P_n$. Hence, researchers get:

$$\|\vec{a}\|^2 = \vec{v}^H P_n \vec{v} \quad (10)$$

Minimizing $\vec{a}$ is equivalent to finding vector $\vec{v}$ that minimizes the quadratic form of $\vec{v}^H P_n \vec{v}$. Reformulating the constrained minimization problem:

$$\min \vec{v}^H P_n \vec{v} \quad \text{subject to} \quad \vec{v}^H \left(P_n^H \vec{u}_1\right) = 1 \quad (11)$$

Once, solution of Eq. 8 is found, the minimum norm solution is formed by projecting $\vec{v}$ onto noise sub-space using Eq. 6 and using optimization theory the minimum norm solution is found to be:

$$\vec{a} = P_n \vec{v} = \lambda P_n \vec{u}_1 = \frac{\left(P_n \vec{u}_1\right)}{\left(\vec{u}_1^H P_n \vec{u}\right)} \quad (12)$$

Which is the projection of the unit vector onto noise sub-space normalized so that the first coefficient is unity. Here:

$$\lambda = \frac{1}{\left(\vec{u}_1^H P_n \vec{u}\right)} \quad (13)$$

In terms of eigen vectors of the autocorrelation matrix, the minimum norm solution is given using quadratic (Q_R) factorizing by the following equation:

$$\vec{a} = \frac{\left(\left(V_n V_n^H\right)\vec{u}_1\right)}{\left(\vec{u}_1^H \left(V_n V_n^H\right)\vec{u}_1\right)} \quad (14)$$

RESULTS AND DISCUSSION

The proposed algorithm has been tested on various genes to predict location of coding regions of varying lengths and simulation results are compared with that of periodogram method on the same DNA data. According to period-3, property of DNA a prominent peak should be visible in the PSD plot of each exon segment. Specific coding regions of Celegan Cosmid *F56F11.4a* gene are mentioned in Table 1 and the statistical parameters and computation times of both periodogram and minimum norm methods for genes *F56F11.4a*, *T12B5.1*, *C30C11* and *D13156* are indicated in Table 2 and Fig. 3.

It is seen that this new approach removes the entire noise and reveals the hidden periodicities prominently. A comparison has been drawn with periodogram method with Bartlett (triangular) sliding window with 50% overlap and suitable segment lengths M and number of segments

Table 1: Celegan *F56F11.4a* gene coding regions

Exon	Start-end (bp)	Exon length (bp)
1	7948-8059	111
2	9548-9877	330
3	11134-11397	264
4	12485-12664	180
5	14275-14625	351

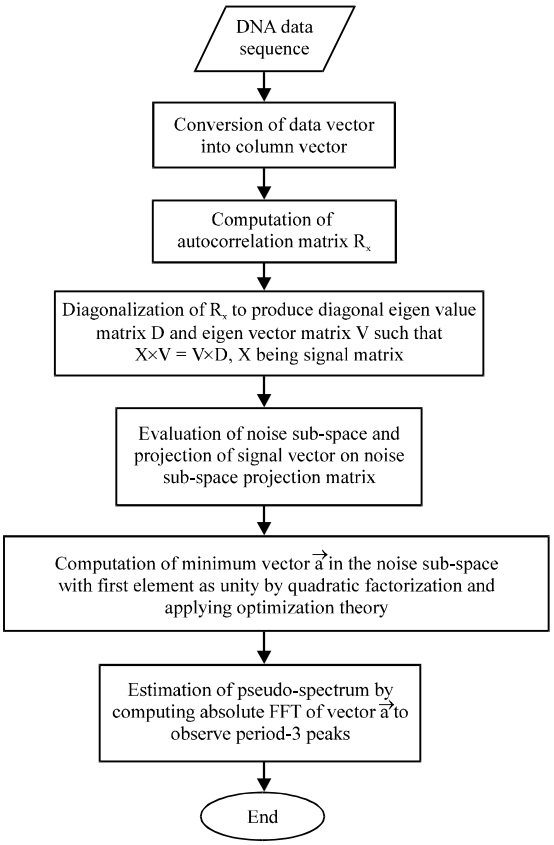

Fig. 3: Flowchart of algorithm for proposed minimum norm solution method for estimating period-3 peaks

K. The choice of window's length M should be done subjectively based on a trade-off between spectral resolution and statistical variance. If M is too small important features may be smoothed out while if M is too large the behavior becomes more like periodogram with erratic variation. Therefore, a compromise value is chosen between range 1/25<M/N<1/3 where N is nucleotide sequence length. Quality factor which measures the ratio of variance to square of mean of PSD has been used as metric for comparison between the two methods as shown in Table 2. The plot of quality factor of various genes is given in Fig. 4. It is observed that quality factor of spectrum by minimum norm method is much higher than periodogram method. Table 2 also indicates that computation time required in minimum norm method is more than periodogram method.

Table 2: Summary of statistical parameters and computation time of periodogram and minimum norm methods for various genes

	Sliding DFT method				Minimum norm method			
Genes	Q.F. (mean)^2/var	CPU time (sec)	Window length (m)	K No. of segments	Q.F. (mean)^2/var	CPU time (sec)	Model order (p)	Percent rise in Q.F.
F56F11.4a	4.83	0.24	351	23	121.89	104.87	20	2.42e+3
T12B5.1G-1	6.31	0.14	252	7	347.96	48.72	8	5.41e+3
T12B5.1G-2	5.58	0.14	252	8	305.50	50.38	16	5.37e+3
T12B5.1G-3	3.54	0.12	252	4	742.96	06.69	2	2.09e+4
T12B5.1G-4	8.38	0.15	252	9	221.09	54.15	17	2.54e+3
T12B5G5G-5	5.88	0.14	252	6	227.29	07.76	17	3.76e+3
C30C11G-1	10.42	0.18	252	12	498.40	11.38	7	4.68e+3
C30C11G-2	3.92	0.10	210	4	107.79	06.20	17	2.65e+3
D13156	4.84	0.15	351	5	246.08	37.38	17	4.98e+3

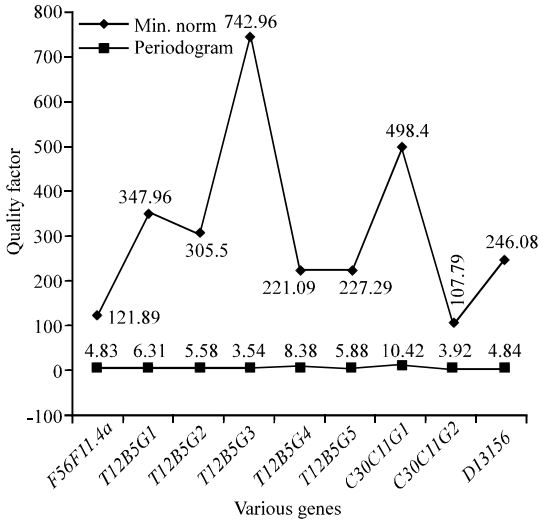

Fig. 4: Quality factor for various genes by minimum norm and periodogram methods

Performance comparison of proposed method with existing method: The performance analysis of the methods can be made by prediction measures, such as Sensitivity (S_N), Specificity (S_P), Miss Rate (M_R) and Wrong Rate (W_R). Their definitions are given below:

$$S_N = \frac{T_P}{(T_P + F_N)} \quad (15)$$

$$S_P = \frac{T_P}{(T_P + F_P)} \quad (16)$$

$$M_R = \frac{M_E}{A_E} \quad (17)$$

$$W_R = \frac{W_E}{P_E} \quad (18)$$

Where:

M_E = Missing Exons
A_E = Actual Exons
W_E = Wrong Exons
P_E = Predicted Exons
T_P = True Positive
F_P = False Positive
F_N = False Negative

T_P corresponds to those genes that are accurately predicted by the algorithm and also exist in the GenBank annotation. F_P corresponds to the exon regions identified by the given algorithm but are not specified in the standard annotation. F_N is coding region that is present in the GenBank annotation but is not predicted as a coding segment by the algorithm used. The average of S_N and S_P gives the overall exon sensitivity and specificity. Table 3 summarizes the simulation results of the eight genes used as test data. It is evident from tabulated data that S_N, S_P and the average of S_N and S_P of proposed method are higher than existing method in all the cases where as miss and wrong rate are much lower indicating superior performance of the proposed algorithm over existing method (Meher *et al*., 2011).

In the beginning, both periodogram technique and proposed minimum norm algorithm are applied to Celegan Cosmid *F56F11.4a* gene. The periodogram result is shown in Fig. 4 and the proposed algorithm result is plotted in Fig. 5. It is evident from Fig. 6 that 5, period-3 spectral peaks without any noise component are visible in the specific coding regions as per data mentioned in Table 1.

Figure 7 and 8 show the results of application of conventional periodogram method and proposed minimum norm solution method to 32488 bp length Celegans Cosmid T12B5.1 DNA (Accession No. FO081674.1 AF100307). The plots show 3 exons in gene-1 between 17332-17402, 17645-18266 and 18311-18505 bp. In Fig. 6, the exon peaks are present along with other peaks, hence prediction becomes ambiguous. In Fig. 7 obtained by the proposed algorithm, there are 3 sharp period-3 peaks present in proper location absolutely devoid of

Table 3: Performance analysis summary of data for minimum norm and periodogram methods

			Prediction			Measures	
Genes	DSP methods	Threshold value	S_N	S_P	$(S_N+S_P)/2$	M_R	W_R
F56F11.4a	Periodogram	1.75	0.40	1.00	0.70	0.60	0.00
	Periodogram	1.50	0.80	0.66	0.73	0.20	0.40
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
T12B5 Gene-1	Periodogram	1.75	1.00	0.43	0.71	0.00	0.55
	Periodogram	1.50	1.00	0.33	0.66	0.00	0.66
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
T12B5 Gene-2	Periodogram	1.75	1.00	0.60	0.80	0.00	0.40
	Periodogram	1.50	1.00	0.50	0.75	0.00	0.50
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
T12B5 Gene-3	Periodogram	1.75	1.00	0.15	0.57	0.00	0.84
	Periodogram	1.50	1.00	0.12	0.56	0.00	0.87
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
T12B5 Gene-4	Periodogram	1.75	0.50	0.40	0.45	0.50	0.60
	Periodogram	1.50	0.75	0.33	0.54	0.25	0.66
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
T12B5 Gene-5	Periodogram	1.75	0.66	0.22	0.44	0.33	0.77
	Periodogram	1.50	1.00	0.25	0.62	0.00	0.75
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
C30C11 Gene-1	Periodogram	1.75	0.50	0.40	0.45	0.50	0.60
	Periodogram	1.50	1.00	0.40	0.70	0.00	0.60
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
C30C11 Gene-2	Periodogram	1.75	1.00	0.33	0.66	0.00	0.66
	Periodogram	1.50	1.00	0.21	0.60	0.00	0.78
	Min. norm	*	1.00	1.00	1.00	0.00	0.00
D13156	Periodogram	1.75	1.00	0.22	0.61	0.00	0.77
	Periodogram	1.50	1.00	0.15	0.57	0.00	0.86
	Min. norm	*	1.00	0.50	0.75	0.00	0.50

*Threshold value not required

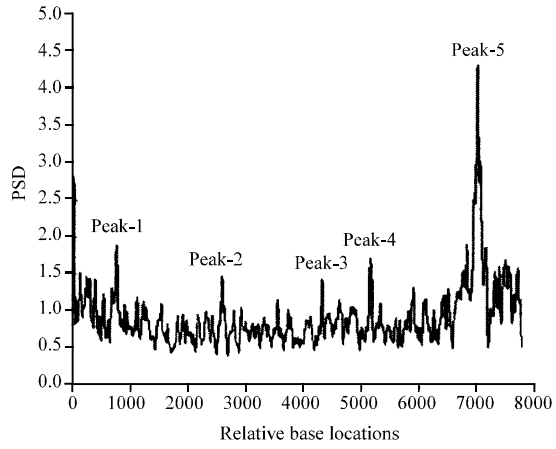

Fig. 5: Plot of PSD by periodogram method for the *F56F11.4a* gene

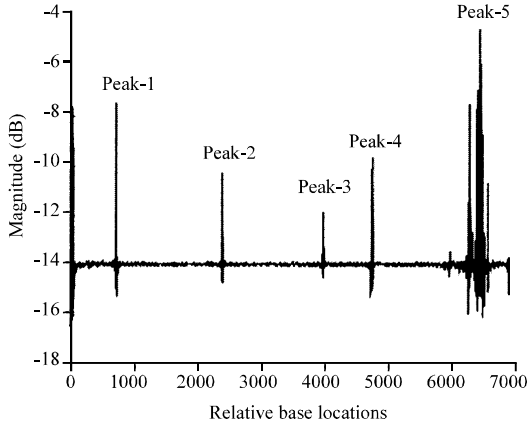

Fig. 6: Plot of period-3 peaks by minimum norm solution for *F56F11.4a* gene

noise without any scope of ambiguity. Similar results are seen in Fig. 9 and 10 for gene-2 with 3 exons between 18994-19064, 19349-19997 and 20059-20253 bp. The technique was verified successfully for the remaining three genes also. As a third example, application of both the methods to DNA C30C11 (Accession No. FO080722.7 L09634) from Celegan Chromosome-III having length 30866 bp was considered. Figure 11 and 12 mention spectral peaks by periodogram and minimum norm solution methods, respectively for gene-1 having exons between 4874-4985, 5034-5408, 5452-6179 and 6227-6526 bp. In Fig. 12, it is observed that peak-2 is shifted right from actual position. Figure 13 and 14 indicate accurate results for gene-2 with exon segments between 7320-7503, 7555-7757 and 7804-7923 bp.

All these plots showing results of both the methods reflect the superiority of proposed technique over the conventional method because the peaks obtained with proposed algorithm are sharp, unambiguous and without any noise. The threshold

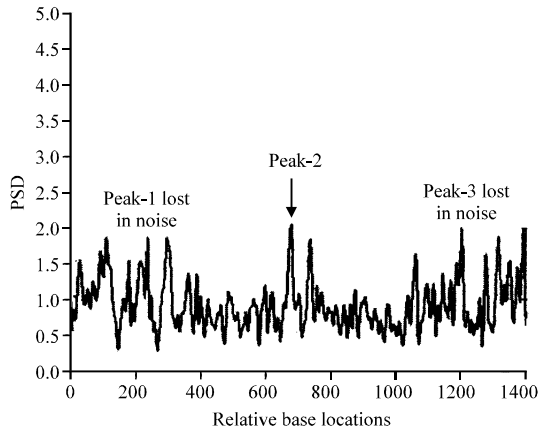

Fig. 7: Plot of PSD by periodogram method for *T12B5.1* gene-1

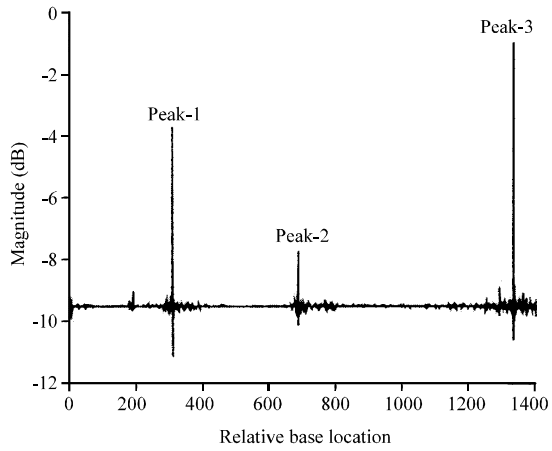

Fig. 8: Plot of period-3 peaks by minimum norm solution for *T12B5.1* gene-1

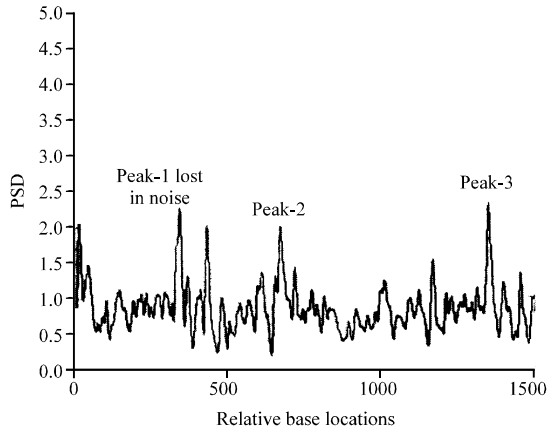

Fig. 9: Plot of PSD by periodogram method for *T125B.1* gene-2

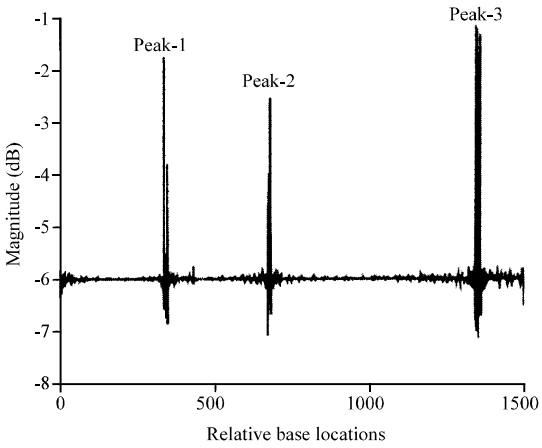

Fig. 10: Plot of period-3 peaks by minimum norm solution for *T12B5.1* gene-2

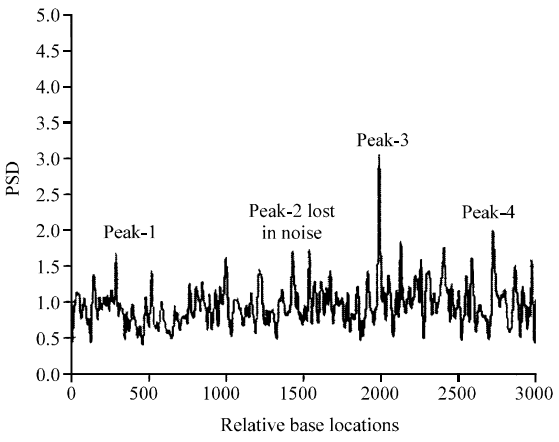

Fig. 11: Plot of PSD by periodogram method for Celegan *C30C11* gene-1

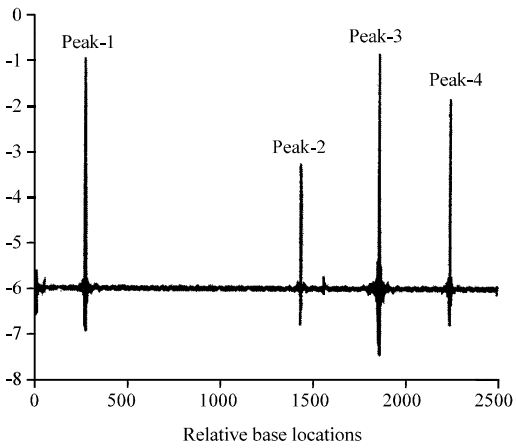

Fig. 12: Plot of period-3 peaks by minimum norm solution for *C30C11* gene-1

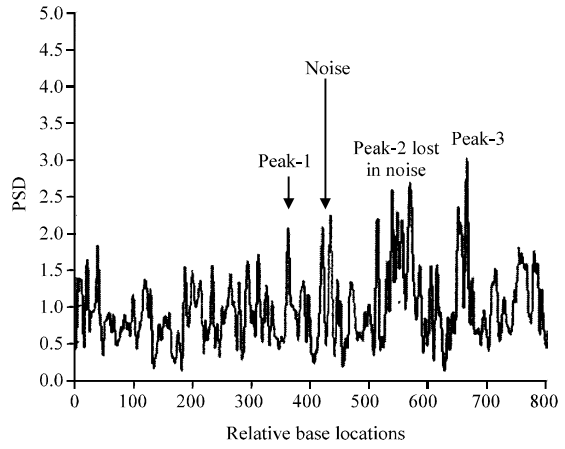

Fig. 13: Plot of PSD by periodogram method for *C30C11* gene-2

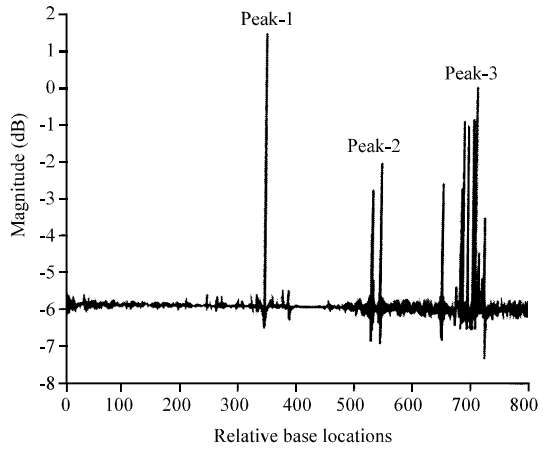

Fig. 14: Plot of period-3 peaks by minimum norm solution for *C30C11* gene-2

values of periodogram method for performance analysis have been chosen juditiously as 1.75 and 1.5, respectively. Table 3 indicates list of genes studied and performance analysis of periodogram and minimum norm solution approaches. In all the examples cited the proposed method shows better results than the existing method giving higher value of sensitivity, specificity and their average as well as low miss and wrong rates.

Model order selection approach based on eigen ratio: A key issue in developing eigen decomposition model is selection of proper model order p. In order to estimate minimum norm solution based pseudo-spectrum, the dimension M-p of the noise subspace must be determined accurately. If value of p taken is less, then few prominent peaks may go unnoticed. On the other hand if p is more than required value, undesired peaks are introduced in the

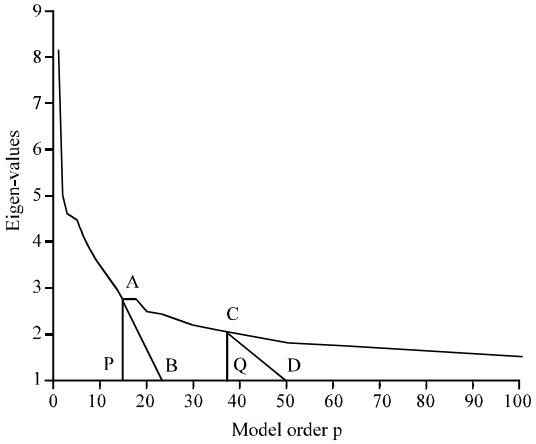

Fig. 15: Plot of eigen value vs. model order for *F56F11.4a* gene

plot leading to false prediction. The most common approach is to calculate and sort the eigen values of the correlation matrix R_x of the noisy signal as mentioned here. The eigen values plotted in decreasing order is known as scree plot. The prime eigen values of dimension p with steep slope correspond to the signal subspace. The set of smallest eigen values having dimension M-p with values equal to noise variation σ^2 is more or less flat in nature (Fig. 1). Decrease in negativity of the derivative from higher value to lower value is determined by slope of tangents drawn from the scree plot to the x-axis. About 2 points are chosen carefully on the scree plot, such that the first is on steep slope and second is on less steep portion of the eigen curve. The values of model order p intercepted by the two projections drawn vertically downward from the point of the tangent touching the eigen curve (scree plot) to the x-axis are identified. A big gap or elbow is looked for within this segment to be vtreated as threshold between signal and noise sub-spaces with the help of the following technique (Fig. 15 and 16).

A very simple method based on eigen ratio adopted by the researcher has been discussed in this study (Liavas and Regalia, 2001). As shown in Fig. 17 and 18, the researchers have plotted eigen value ratio λ_p/λ_{p+1} vs p. It is observed that there exists an eigen value gap of high magnitude between orders p = 20, 21, 16 and 17, respectively. Satisfactory estimates of rank of R_x by suggested method was found to be 20 for *F56F11.4a* gene, 16 for *T12B5.1* gene-2 and 7 for *C30C11* gene-1, respectively. Thus, it may be considered that eigen values $\lambda_{21,\ 17,\ 8}$ onwards are the noise eigen values in the three cases, respectively.

Spectral content measure techniques based on sliding DFT was compared with proposed technique.

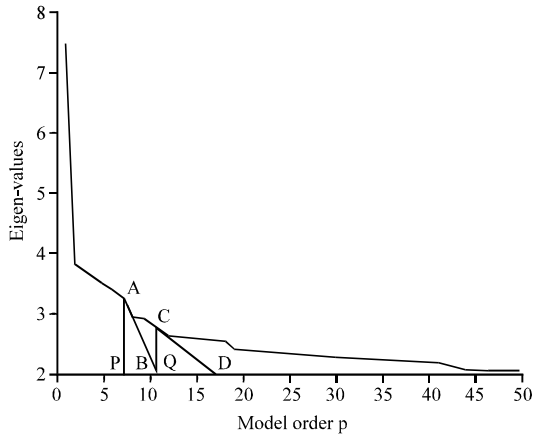

Fig. 16: Plot of eigen value vs. model order for *C30C11* gene-1

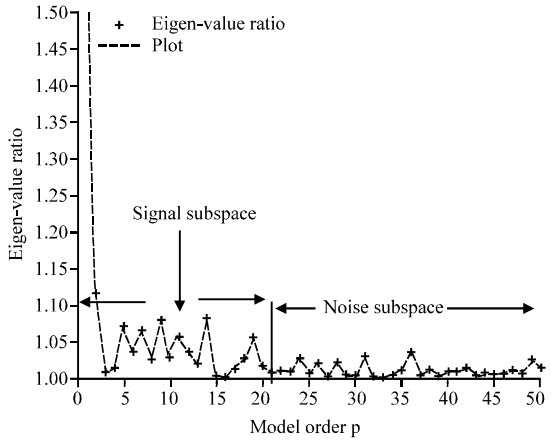

Fig. 17: Plot of eigen value ratio vs. model order *F56F11.4a* gene

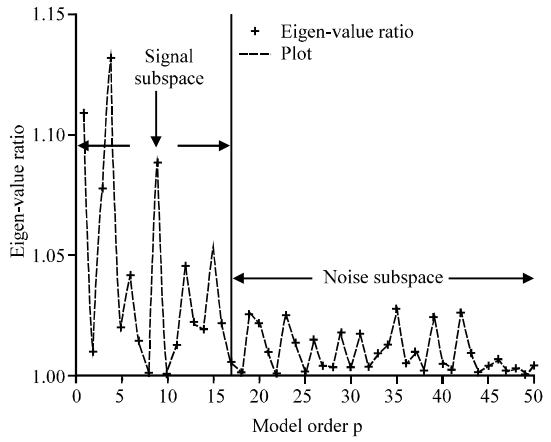

Fig. 18: Plot of eigen value ratio vs. model order for *T12B5.1* gene-2

Tiwari *et al.* (1997) employed Fourier technique to analyze three-base periodicity in order to recognize coding regions in genomic DNA. They observed that some genes do not exhibit period-3 property at all in *S. cerevisiae*. Anastassiou (2000, 2001) was inspired by the research of Tiwari *et al.* (1997) and introduced computational and visual tools for analyzing bio-molecular sequences. Researcher provided optimization procedure for improving performance of traditional Fourier technique. Vaidyanathan and Yoon (2004) designed multistage narrowband bandpass filter for reducing background 1/f noise. Sahu and Panda (2011) improved computational efficiency by employing SDFT with the help of Goertzel algorithm but the method is constrained by frequency resolution and spectral leakage effects.

The minimum norm algorithm presented in this study provides a novel approach. The first important feature of the proposed algorithm is that it produces extremely sharp period-3 peaks in the protein coding regions. The second important feature is that it eliminates noise completely, hence there is no requirement of setting threshold value. Moreover, this method offers very high sensitivity and specificity at the same time very low miss rate and wrong rate compared to other available techniques. The proposed algorithm though offers high predictive accuracy compared to existing methods, it has certain limitations on its part. Model order selection which is a key issue needs to be done judiciously for accurate exon detection. The time of execution is more compared to existing methods, since it depends on autocorrelation lag size which is pre-determined depending on length of nucleotide sequence being tested.

CONCLUSION

DNA sequence analysis through power spectrum estimation by traditional non-parametric methods is in use for long. These are methodologically straight forward, computationally simple and easy to understand but due to low SNR spectral features are difficult to distinguish as noise artifacts appear in spectral estimates. Therefore, effective identification of protein-coding region becomes difficult. The application of minimum norm frequency estimator to capture period-3 peaks in coding regions has been introduced here. Researchers used a constrained vector that lies on the noise subspace and completely filtered out the spurious peaks. Selection of proper model order is a fundamental issue in application of eigen decomposition approach. The eigen ratio gap or elbow located on the scree plot is treated as threshold between signal and noise spaces. Use of eigen decomposition based methods to various DNA sequence has given

amazing results as compared to standard classical methods in terms of resolution, quality factor, sensitivity, specificity, miss rate and wrong rate. It was observed that high resolution pseudo-spectrum estimation, such as minimum norm could be effectively used for identification of protein coding regions in DNA. Unfortunately, the computational effort of this high resolution method is significantly higher than FFT processing. Since, the main objective is to detect protein coding regions accurately which is fulfilled by the proposed method, increase in computation time may be compromised. Hence, it can be concluded that identification of protein-coding regions in DNA can be done effectively in a much superior way by applying minimum norm technique compared to periodogram power spectrum estimator.

REFERENCES

Anastassiou, D., 2000. Frequency-domain analysis of biomolecular sequences. Bioinformatics, 16: 1073-1081.

Anastassiou, D., 2001. DSP in genomics: Processing and frequency-domain analysis of character strings. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, May 07-11, 2001, Salt Lake City, USA., pp: 1053-1056.

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequences. Nucleic cids Res., 10: 5303-5318.

Hayes, M.H., 1996. Statistical Digital Signal Processing and Modeling. John Wiley and Sons Inc., New York, USA., pp: 393-474.

Haykin, S., 2008. Adaptive Filter Theory. 4th Edn., Dorling Kindersley Pvt. Ltd., India, pp: 809-822.

Liavas, A.P. and P.A. Regalia, 2001. On the behavior of information theoretic criteria for model order selection. IEEE Trans. Signal Process., 49: 1689-1695.

Lobos, T., Z. Leonowics and J. Rezmer, 2000. Harmonics and inter-harmonics estimation using advanced signal processing methods. Proceedings of the 9th International Conference on Harmonics and Quality Power, Volume 1, October 1-4, 2000, Orlando, Florida, USA., pp: 335-340.

Manaswini, P. and R.K. Sahu, 2010. An exclusive survey on gene prediction methodologies. Int. J. Comput. Sci. Inform. Secur., 8: 88-103.

Meher, J., P.K. Meher and G. Dash, 2011. Improved comb filter based approach for effective prediction of protein coding regions in DNA sequences. J. Signal Inform. Process., 2: 88-99.

Nair, A.S. and S. Sreenadhan, 2006. An improved digital filtering technique using nucleotide frequency indicators for locating exons. J. Comput. Soc. India, 36: 54-60.

Peng, C.K., S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M. Simons and H.E. Stanley, 1995. Statistical properties of DNA sequences. Physica A Stat. Mech. Appl., 221: 180-192.

Praokis, J.G. and D.G. Manolakis, 2008. Digital Signal Processing: Principles, Algorithms and Applications. 4th Edn., PHI Learning Pvt. Ltd., India, pp: 960-985.

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. Proceedings of the 4th International Conference on Computers and Devices for Communication, December 14-16, 2009, Kolkata, India, pp: 1-4.

Roy, M. and S. Barman, 2011. Spectral analysis of coding and non-coding regions of a DNA sequence by parametric and nonparametric methods: A comparative approach. Ann. Faculty Eng. Hunedoara Int. J., 9: 57-62.

Sahu, S.S. and G. Panda, 2011. Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach. Genomics Proteomics Bioinform., 9: 45-55.

Shlens, J., 2003. A tutorial on principal component analysis, derivation, discussion and singular value decomposition. Version 1, March 25, 2003, http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf.

Stoica, S. and R. Moses, 2011. Spectral Analysis of Signals. PHI Pvt. Learning Ltd., India, pp: 23-67.

Tiwari, S., S. Ramachandran, A. Bhattacharya, S. Bhattacharya and R. Ramaswami, 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput. Applied Biosci., 13: 263-270.

Tuqan, J. and A. Rushdi, 2008. A DSP approach for finding the codon bias in DNA sequences. IEEE J. Sel. Top. Signal Process., 2: 343-356.

Ubeyli, E.D. and I. Guler, 2003. Comparison of eigenvector methods with classical and model-based methods in analysis of internal carotid arterial Doppler signals. Comput. Biol. Med., 33: 473-493.

Vaidyanathan, P.P. and B.J. Yoon, 2004. The role of signal processing concepts in genomics and proteomics. J. Franklin Instit., 341: 111-135.

Yin, C. and S.S. Yue, 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol., 247: 687-694.

Zhao, L., 2006. Application of spectral analysis to DNA sequences. Purdue University, Purdue e-Pubs, Report Number: 06-003.