Gene 575 (2016) 488–497

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Improved gene prediction by principal component analysis based autoregressive Yule–Walker method

Manidipa Roy, Soma Barman *

Institute of Radio Physics and Electronics, University of Calcutta, 92, APC Road, Kolkata 700009, India

ARTICLE INFO

Article history:
Received 11 March 2015
Received in revised form 25 August 2015
Accepted 11 September 2015
Available online 16 September 2015

Keywords:
Spectral analysis
Principal component analysis
Wavelet packet transform
Eigenvalue-ratio

ABSTRACT

Spectral analysis using Fourier techniques is popular with gene prediction because of its simplicity. Model-based autoregressive (AR) spectral estimation gives better resolution even for small DNA segments but selection of appropriate model order is a critical issue. In this article a technique has been proposed where Yule–Walker autoregressive (YW-AR) process is combined with principal component analysis (PCA) for reduction in dimensionality. The spectral peaks of DNA signal are used to detect protein-coding regions based on the 1/3 frequency component. Here optimal model order selection is no more critical as noise is removed by PCA prior to power spectral density (PSD) estimation. Eigenvalue-ratio is used to find the threshold between signal and noise subspaces for data reduction. Superiority of proposed method over fast Fourier Transform (FFT) method and autoregressive method combined with wavelet packet transform (WPT) is established with the help of receiver operating characteristics (ROC) and discrimination measure (DM) respectively.

1. Introduction

Application of spectral analysis to find periodicities in DNA sequences is presently being explored by various researchers. Since success rate of ab initio gene prediction is still very low it is an open research problem (Guigo et al., 2006) in genomics. Gene prediction refers to detection of protein coding regions of genes in a long DNA sequence. Coding regions being the actual information bearing part, its identification and study is of utmost importance. 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. The DNA sequence can be divided into genes and inter-genic spaces. Genes can again be subdivided into exons and introns. It has been established that base sequences in the coding regions of DNA molecules exhibit a period-3 property because of the codon structure involved in the translation of nucleotide bases into amino acids (D. Anastassiou, 2000; Vaidyanathan and Yoon, 2004). The discrete nature of DNA has facilitated implementation of DSP techniques to extract period-3 components by effectively eliminating background noise (Yin and Yau, 2007; Tuqan and Rushdi, 2008). Sussillo et al. (2004) performed frequency-domain analysis in the genomes of various organisms e.g. *Saccharomy cescerevisiae* and *Drosophila melanogaster* using tricolor spectrogram identifying several types of distinct visual patterns characterizing specific DNA regions. Rao and Shepherd (2004) established that when a high-resolution spectrum is desired the parametric method provides a better alternative even for small duration signals, provided accurate model order is estimated. A model based exon detection approach using statistically optimal null filter (SONF) was proposed by Kakumani et al. (2008) for short exons and successive exons separated by short introns. But the SONF method lacked in eliminating spurious peaks. Chakrabarty et al. (2004) proposed a gene prediction technique where AR models of coding and non-coding DNA regions were compared based on AR residual errors. They pointed out that AR spectrum analysis was not suitable as high order models developed spurious spectral peaks. Rosen (2007) further explored AR method under noise conditions and showed that an extremely large number of coefficients are needed to model DNA sequences due to the spurious peaks. Sahu and Panda (2010) applied adaptive autoregressive technique in gene prediction to increase its efficiency.

The problems stated by Chakrabarty et al., Rao et al. and Rosen were addressed in the proposed work by applying PCA prior to AR modeling. In this paper the authors have adopted a fixed model order technique for proper choice of model order. A novel model based technique is used in conjunction with PCA to get larger period-3 peaks with better resolution and less number of spurious peaks than existing methods. The advantage of the proposed method is that optimal model order selection is no more critical as noise floor is removed prior to PSD estimation. Since PCA focuses only on principal components with large eigenvalues, prediction performance is significantly improved (M.H. Hayes, 1996). A compact mapping rule assigning numerical values a = −1, c = −j, g = 1 and t = j to nucleotide bases has been used for converting symbolic DNA

Abbreviations: AR, Autoregressive; YW-AR, Yule–Walker autoregressive; PCA, Principal Component Analysis; PSD, Power Spectral Density; FFT, Fast Fourier Transform; WPT, Wavelet Packet Transform; ROC, Receiver Operating Characteristics; DM, Discrimination Measure.

* Corresponding author.
E-mail address: barmanmandal@gmail.com (S. Barman).

http://dx.doi.org/10.1016/j.gene.2015.09.023
0378-1119/

sequence into numerical form (Rao and Shepherd, 2004; Kwan et al., 2012). The simulated system model is realized using MATLAB (version R2009b) environment to identify protein coding regions in eukaryotic genes. DNA (de-oxyribo-nucleic acid) databases available in National Center for Biotechnology Information (NCBI Gen Bank) are used for the study (http://www.ncbi.nlm.nih.gov).

A brief overview of spectral analysis by FFT and Yule–Walker autoregressive modeling is furnished in Section-2.1 and Section-2.2 respectively. In Section-2.3 eigen-decomposition and principal component analysis are elaborated. Principal component autoregressive spectral estimation is discussed in Section-2.4. In the result and discussion section (Section-3), power spectral density (PSD) plots obtained by different methods are presented. In Section-4 evaluation criteria for various datasets is discussed with the help of ROC curves and discrimination measure (DM). In Section-5 validation of the proposed method with AR-WPT method is presented. Lastly in Section-6 a conclusion based on discussion in previous two sections is drawn.

2. Materials & methods

2.1. Spectral analysis of DNA sequence by FFT method

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

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

where $f_k = k/N$, for $k = 0,1,2,\text{——}N-1$.

To enhance performance of periodogram which is known as modified periodogram method, at first the N-point data sequence is subdivided into k overlapping segments of length M each subsequently periodogram is computed and averaged with Bartlett windowing (Praokis and Manolakis, 1992).

2.2. Autoregressive spectrum estimation

AR method is popular because estimation of AR parameters is achieved easily by solving linear Yule–Walker equations. Parametric spectrum estimation consists of two basic steps. Given a data sequence $x[n]$ for $0 \le n \le N$, first the parameters of the model are estimated, then PSD is computed from these estimated parameters. An autoregressive process $x(n)$ may be represented as the output of an all-pole filter that is driven by unit variance white noise. The estimated power spectrum of p^{th} order autoregressive process is given by the following equation:

$$\hat{P}_{AR}\left(e^{jw}\right) = \frac{|\hat{b}(0)|^2}{|1 + \sum_{k=1}^{p} \hat{a}_p(k)e^{-jwk}|^2} \quad (2)$$

where variance $|\hat{b}(0)|^2$ and coefficients $\hat{a}_p(k)$ are estimated from given data. Since in the above equation variance $|\hat{b}(0)|^2$ is constant, the only values that are needed for calculating the shape of PSD are the coefficients $\hat{a}_p(k)$. The estimation of $\hat{P}_{AR}(e^{jw})$ may be facilitated if something is known about the process in addition to signal values. The accuracy of estimation depends on how accurately the model parameters can be estimated and whether the AR model is consistent with generated data. Out of various all-pole modeling methods Yule–Walker method (AR-YW) is explored in this paper. For model order selection, normalized mean-squared prediction error $\nu_{(p)}$ is plotted as a function of model order p where $\varepsilon_{(p)}$ denotes p^{th} prediction error squared.

$$\nu_{(p)} = \varepsilon_{(p)}/\varepsilon_{(0)} \quad (3)$$

From the plot it is observed that there is often a value of p above which increasing p has little effect on $\nu_{(p)}$. This value is an efficient choice for optimum order p (Oppenheim and Schafer, 2013).

2.3. Eigen-decomposition and principal component analysis

The broad application of principal component analysis is in pattern classification, where it has been used here for identifying the feature vectors that are most significant. Principal component spectrum estimation uses vectors that lie in the signal subspace. The eigenvectors and eigenvalues of the correlation matrix of the noisy signal are partitioned into two disjoint subsets. The set of eigenvectors $\{v_1, v_2 \ldots v_p\}$, associated with largest eigenvalues span the signal subspace and are called principal eigenvectors. The second subset of eigenvectors $\{v_{p+1}, v_{p+2} \ldots v_M\}$ span the noise subspace and have σ_n^2 as eigenvalue. The signal and noise eigenvectors as well as the signal and the noise subspaces are orthogonal. After eigen-decomposition of the correlation matrix, eigenvalues are arranged in descending order as shown in Fig. 1.

Then rank p constraint is imposed on the correlation matrix effectively filtering out the noise subspace thus enhancing the signal components (J. Shlens, 2003). The mathematical background of principal component spectrum estimation is given here. Let R_{xx} be an $M \times M$ autocorrelation matrix of the signal consisting of p complex exponentials in white noise. The eigen-decomposition of R_{xx} is given by:

$$R_{xx} = \sum_{i=1}^{M} \lambda_i \vec{v}_i \vec{v}_i^H = \underbrace{\sum_{i=1}^{p} \lambda_i \vec{v}_i \vec{v}_i^H}_{\text{signal}} + \underbrace{\sum_{i=p+1}^{M} \lambda_i \vec{v}_i \vec{v}_i^H}_{\text{noise}}. \quad (4)$$

On effectively filtering out noise portion the estimate of spectral component due to signal alone is given by:

$$R_s = \sum_{i=1}^{p} \lambda_i \vec{v}_i \vec{v}_i^H. \quad (5)$$

2.4. Principal component autoregressive (PC-AR) spectral estimation approach

The authors applied principal component analysis to Yule–Walker AR technique for power spectrum estimation. Autoregressive spectrum estimation using autocorrelation finds solution to the linear equations given as:

$$R_{xx}\vec{a}_M = \varepsilon_M \vec{u}_1 \quad (6)$$

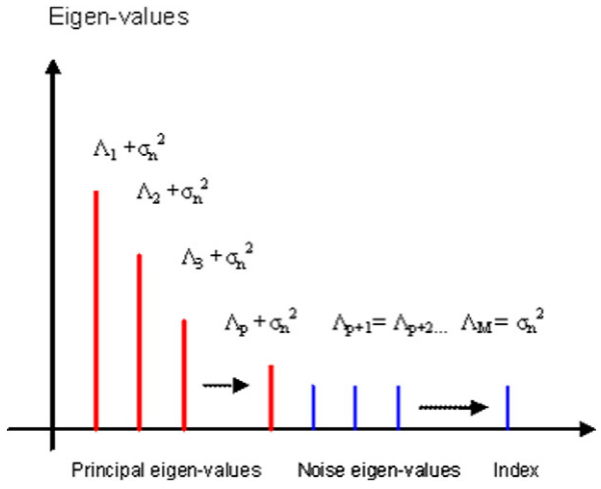

Fig. 1. Decomposition of eigenvalues of noisy signal into principal and noise eigenvalues Λ_i and σ_n^2.

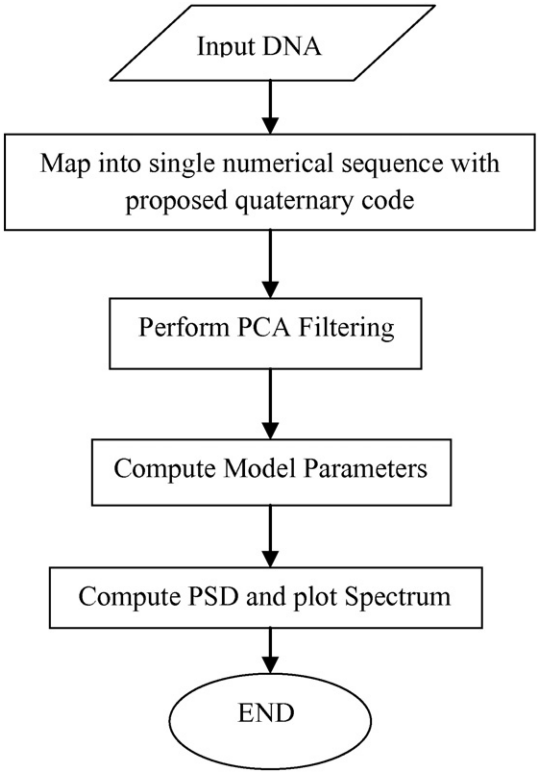

Fig. 2. Flowchart showing algorithm of proposed Principal Component Autoregressive Yule–Walker technique for estimation of period-3 peaks in DNA sequence.

where R_{xx} is an $(M+1)\times(M+1)$ autocorrelation matrix and $\vec{u}_1$ is unit vector which give the solution:

$$\vec{a}_M = \varepsilon_M R_{xx}^{-1}\vec{u}_1. \tag{7}$$

An estimate of the spectrum is given as follows:

$$\hat{P}_{AR}\left(e^{jw}\right) = \frac{|b(0)|^2}{|e^H\vec{a}_M|^2} \tag{8}$$

where $|b(0)|^2 = \varepsilon_M$ is an appropriately chosen constant. If data $x(n)$ consists of p complex exponentials in noise, then a Principal Component solution to Eq. (5) can be formed as:

$$\vec{a}_{pc} = \varepsilon_M \sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\vec{v}_i\vec{v}_i^H\right)\vec{u}_1. \tag{9}$$

The Principal Component Autoregressive spectrum estimate is given by:

$$\hat{P}_{PC\text{-}AR}\left(e^{jw}\right) = \frac{1}{|\sum_{i=1}^{p}\alpha_i e^H\vec{v}_i|^2} \tag{10}$$

where $\alpha_i = v_i^*(0)/\lambda_i$, $v_i(0)$ being the first element of normalized eigenvector $\vec{v}_i$ (M.H.Hayes, 1996).

A simple and practical technique to find the threshold between signal and noise subspaces based on eigenvalue-ratio is employed giving accurate spectral peaks in exon regions only (Liavas and Regalia, 2001). The plot of eigenvalue-ratio λ_p/λ_{p+1} vs order p shows an eigenvalue gap of high magnitude at the threshold separating signal subspace from noise subspace. The algorithm of proposed PC-AR technique for estimation of period-3 peaks is shown as a flowchart in Fig. 2.

3. Results & discussions

In this work authors proposed PC-AR method for locating genes in DNA sequence, applying Yule–Walker autoregressive PSD estimation in combination with principal component analysis of correlation matrix of data. A comparison with FFT (modified periodogram), standard Yule–Walker autoregressive technique and autoregressive method combined with wavelet packet transform (WPT) for de-noising is presented. According to period-3 property of DNA there are prominent visible peaks in PSD plots in the coding areas.

The nucleotide bases from several organisms mentioned in Table 1 are used as raw data for analysis. *Plasmodium knowlesi* gene (Accession no. AF065986) was chosen randomly in course of literature survey. Previous researchers Kakumani et al. (2008) and Tuqan and Rushdi (2008) used the other datasets in their work. As our work is focused on model based approach like Kakumani et al. and 3-base codon periodicity like Tuqan and Rushdi, we chose them as test data. PSD plots by FFT, AR-YW and PC-AR techniques are shown in Figs. 3, 4 and 5 respectively. Fig. 3 shows PSD spectrum by FFT (modified periodogram) method. In the AR-YW method (model based) choice of proper model order selection is very important. The fixed model order method used by the authors resolves the problem of varying predicted model orders while still allowing good spectra to be produced. The model order value at which sharp decline in normalized error variance becomes restricted is chosen as optimum model order for finding AR coefficients. A plot of normalized error variance $\nu_{(p)}$ versus model order p, starting from low to high in steps of 2 is shown in Fig. 6. It indicates that rate of decrease in $\nu_{(p)}$ is less between $p = 20$ to $p = 28$. As per the concept of fixed model order technique the lowest value of $p = 20$ is chosen as the optimum model order to minimize computation time. In the AR-YW spectrum depicted in Fig. 4 with optimum model order value $p = 20$, good performance is discerned.

Table 1
Coding regions of various genes.

S.N.	Name of organism	Accession no.	DNA test sample (start–end) Data length (bp)	Coding sequence
1.	*Plasmodium knowlesi*	AF065986	(420–2340) 1920	545..629,987..1072, 1486..2074, 2196..2224
2.	*Callithrix pygmaea* CBUEGLOBIM	L25361	(1–1580) 1580	144..235,364..586,1399..1527
3.	*C. elegans* Cosmid T12B5.1 Gene-2	FO081674 AF100307	(18,901–20,500) 1600	18994..19064,19339..19997, 20059..20258
4.	*Mus musculus* Domesticus MUSBNP	D16497	(1–1468) 1468	210..335,530..752,1196..1212
5.	*C. elegans* Cosmid C30C11 Gene-1	FO080722 L09634	(4001–7020) 3020	4874..4985,5034..5408, 5452..6179, 6227..6526

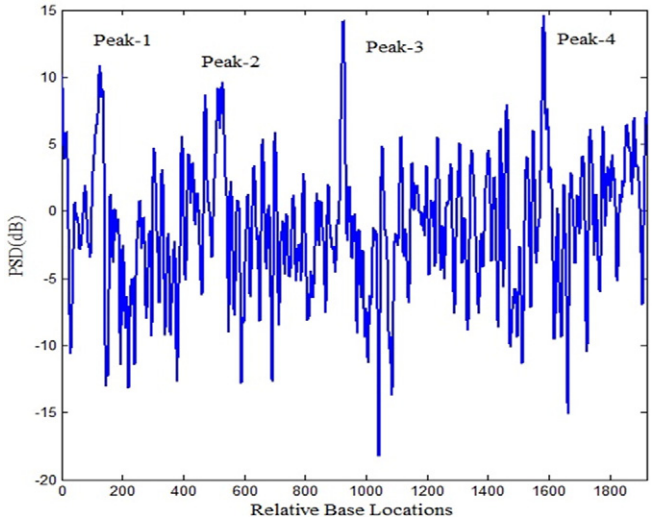

Fig. 3. PSD plot by FFT method for *Plasmodium knowlesi* gene (Accession no. AF0659860) showing four period-3 peaks in noise.

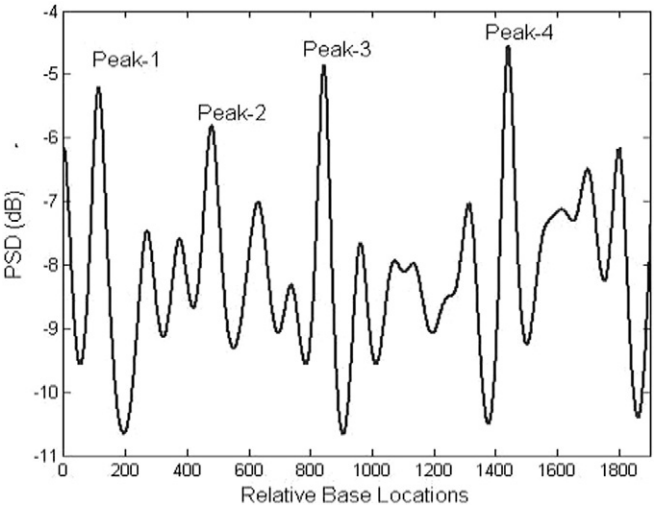

Fig. 4. PSD plot by AR-YW method with coefficient p = 20 for *Plasmodium knowlesi* gene showing the period-3 peaks in less noise.

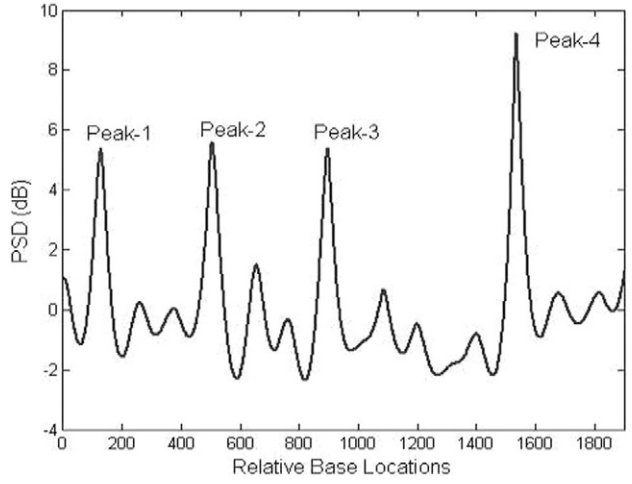

Fig. 5. PSD plot by PC-AR method with coefficient p = 20 and threshold value pth = 12 for Plasmodium knowlesi gene showing prominent period-3 peaks in least noise.

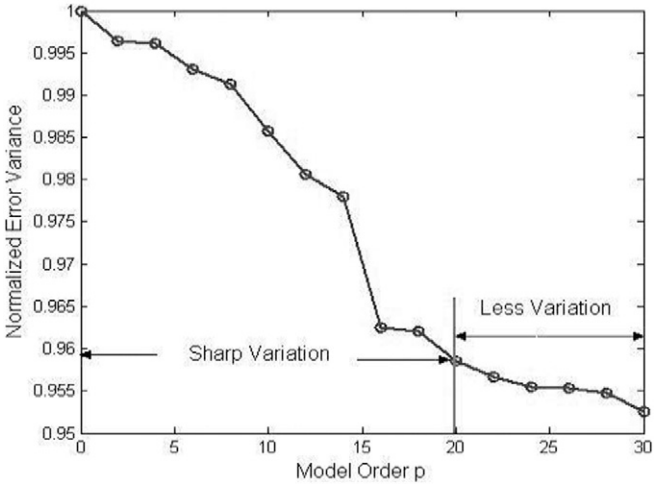

Fig. 6. Model order selection by fixed method in AR-YW technique for *Plasmodium knowlesi* gene indicating less variation in error variance starting from model order p = 20.

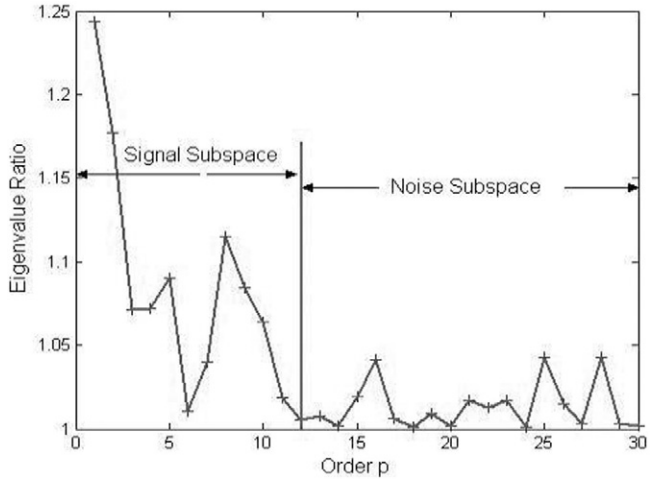

Fig. 7. Eigen decomposition by eigen-ratio method in PC-AR technique for *Plasmodium knowlesi* gene showing threshold value $p_{th} = 12$.

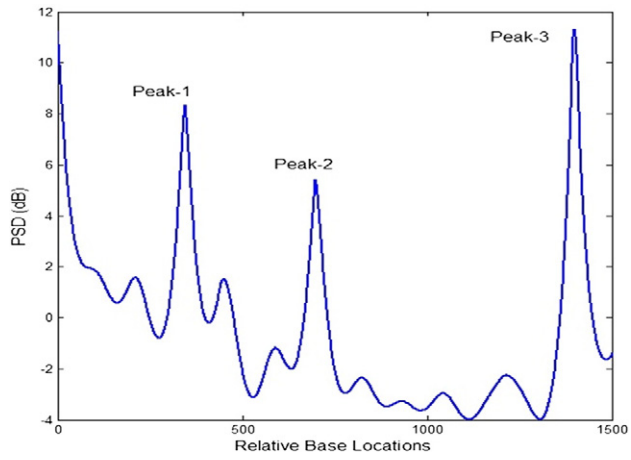

Fig. 8. PSD plot by PC-AR method for T12B5.1 Gene-2 (Accession no. AF100307) with coefficient p = 20 and threshold value $p_{th} = 15$ showing sharp period-3 peaks in coding regions.

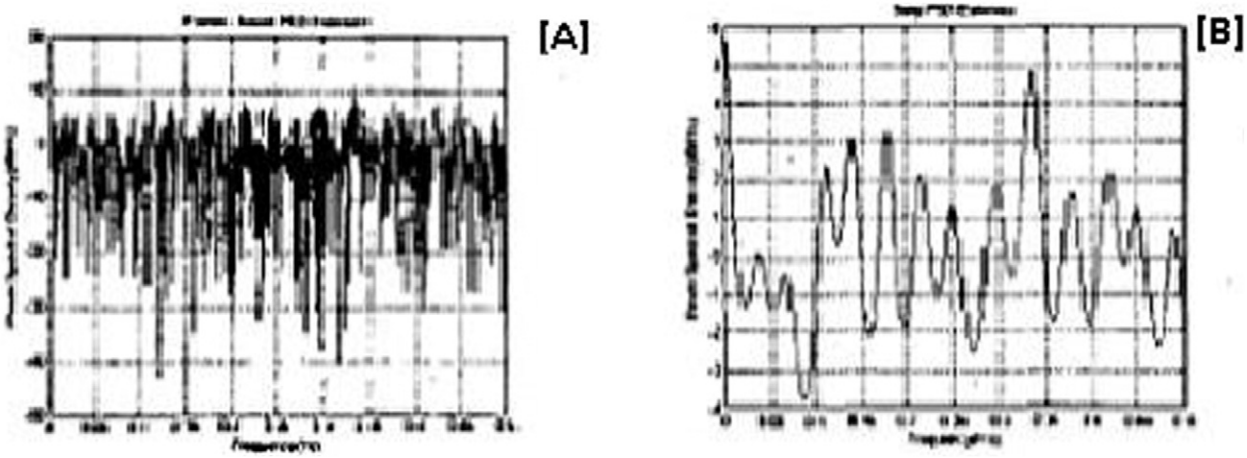

Fig. 9. [A] Estimated FFT power spectrum of HUMIGHAF gene accession no. J00231 [B] Power spectrum by AR Covariance method for HUMIGHAF gene accession no. J00231 Reproduced from Rao and Shepherd (2004).

In the PC-AR process, the dimension of the DNA dataset is reduced by projecting the data onto a few prominent eigenvectors with large eigenvalues by eigen-decomposition. Here eigen-decomposition has been accomplished by eigen-ratio method as shown in Fig. 7. As already stated in the AR-YW method, as the AR order increases there are more coefficients to map both signal and noise components hence selection of optimum model order becomes extremely important. In the proposed method selection of optimum model order becomes less important, because the noise vectors are discarded prior to computation of AR coefficients. The matrix procedure based upon principal components only takes into account the initial eigenvectors describing only the signal components while the latter eigenvectors capturing the noise contributions are eliminated (Fig. 1). Therefore rather than struggling to find an optimum model order, an explicit signal and noise subspace partition is made and only signal subspace is considered for computation. Spectral plot obtained by proposed PC-AR spectrum estimator shown in Fig. 5 with model order $p = 20$ and threshold order $p_{th} = 12$ indicates very sharp peaks in coding regions. Here 'p' corresponds to AR-YW filter coefficient and 'p_{th}' corresponds to threshold between signal and noise subspace. Fig. 8 shows PSD plot by PC-AR method with $p = 20$ and $p_{th} = 15$, applied to T12B5.1 Gene-2 where coding region-1 is only 70 base-pairs long. It is observed that PC-AR technique can detect coding segments smaller than 100 bp length with equal accuracy.

Rao and Shepherd (2004) in their article analyzed a sequence with length 1086 bp from Human lg gamma3 heavy chain disease (HUMIGHAF Accession no. J00231) as a typical example. The sequence has one coding segment between 23 bp to 964 bp. The experimental result obtained by Rao and Shepherd are reproduced in Fig. 9[A] and [B]. No obvious sharp peak is discernible at $f = 1/3$ in the power spectrum by FFT method shown in Fig. 9[A]. The improved AR covariance method

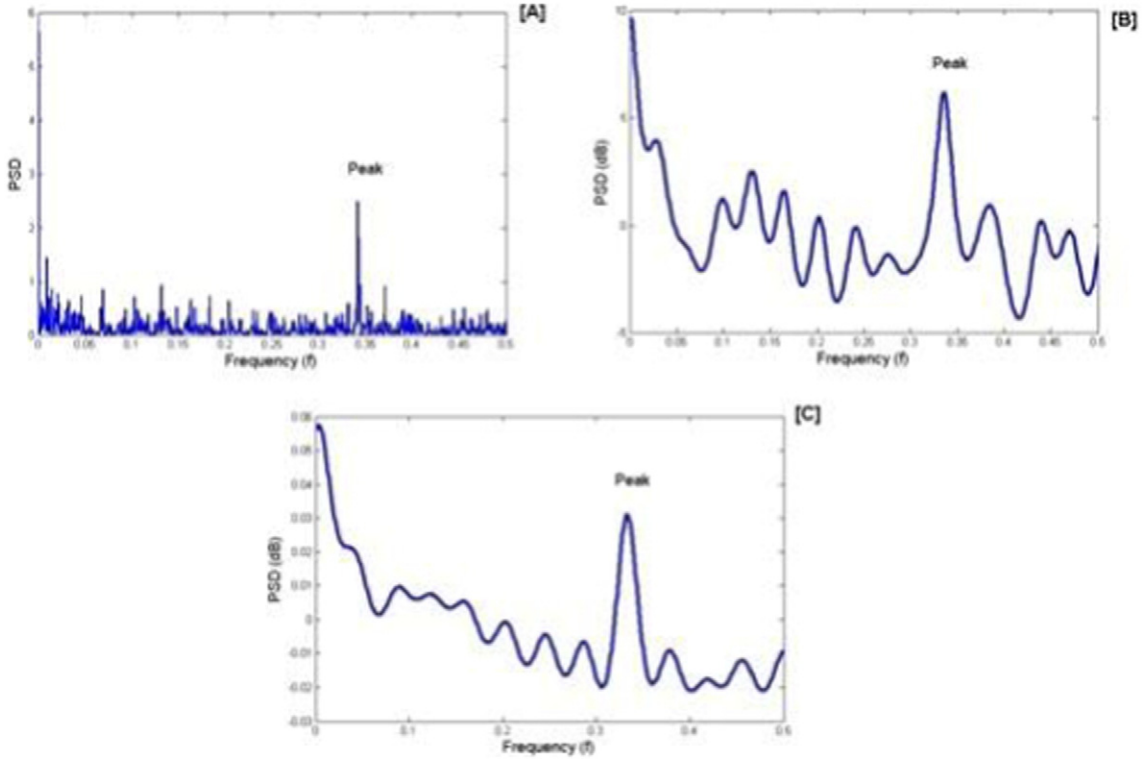

Fig. 10. [A] Power spectrum of HUMIGHAF gene accession no. J00231 by periodogram method [B]] Power spectrum of HUMIGHAF gene accession no. J00231 by AR Yule–Walker method [C] Power spectrum of HUMIGHAF gene accession no. J00231 by PC-AR Yule–Walker method

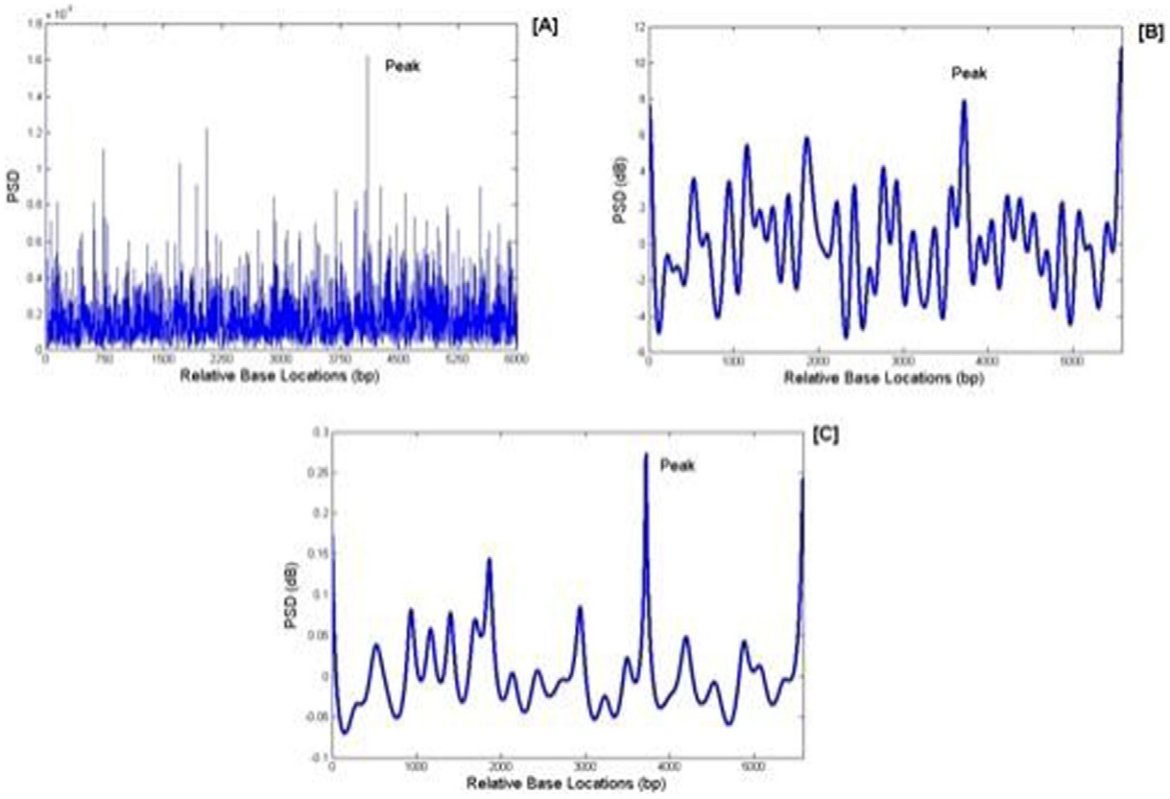

Fig. 11. [A] Spectral plot for YRF1-6 Gene by FFT method showing period-3 peak in noise [B] Spectral plot for YRF1-6 Gene by AR-YW method showing period-3 peak among spurious peaks [C] Spectral plot for YRF1-6 Gene by proposed PC-AR method showing one prominent peak Sample from *S. cerevisiae* Chromosome XIV, YRF1-6/YNL339C Gene (5580 bp).

proposed by Rao and Shepherd, with coefficient p = 30 reproduced in Fig. 9[B] has a 3-periodicity peak at f = 1/3 among other spurious peaks. The same dataset is used in the present work as a test sample by applying periodogram, AR Yule–Walker (p = 30) and PC-AR (p = 30, p_{th} = 5) methods shown in Fig. 10[A], [B] and [C] respectively. It is observed that all the figures display sharp period-3 peaks. The periodogram method has less resolution than AR methods. The proposed PC-AR method has least spurious peaks as noise is eliminated by principal component analysis prior to AR modeling.

Dataset from *S. cerevisiae* Chromosome XIV, YRF1-6/YNL339C Gene used by Sussillo et al. (2004) as test sample is analyzed in Fig. 11[A], [B], and [C]. There is one coding region of 5580 bp length in the said gene. In Fig. 11[A] by FFT method the period-3 peak is visible in noise background, similarly in Fig. 11[B] by autoregressive Yule–Walker method (p = 25), the period-3 peak is discerned among other spurious peaks. Whereas in Fig. 11[C] one clear and prominent period-3 peak is observed in proper location by the proposed Principal Component autoregressive Yule–Walker technique for p = 35 and p_{th} = 10. It is evident from the spectral plot of Fig. 11[C] that the proposed PC-AR method outperforms the FFT and the standard Yule–Walker methods to a great extent providing higher amplitude period-3 peak in less noise. In order to assess the performance of the proposed method, discrimination measures (DM) at optimum model order value has been estimated to be 1.25 for AR-YW method and 1.8 for proposed PC-AR method.

Table 2
Summary of performance analysis of data from Table 1 by FFT method.

Gene	Threshold	Prediction			Measures	
	(dB)	S_N	S_P	$(S_N + S_P)/2$	M_R	W_R
Plasmodium knowlesi	10	0.75	1.00	0.87	0.25	0.00
LOCUS AF065986	8	1.00	0.80	0.90	0.00	0.20
	6	1.00	0.66	0.83	0.00	0.34
Callithrix pygmaea	10	0.66	1.00	0.83	0.34	0.00
LOCUS CBUEGLOBIM	8	0.66	1.00	0.83	0.34	0.00
	6	1.00	0.43	0.71	0.00	0.57
Caenorhabditis elegans	10	0.33	1.00	0.66	0.67	0.00
Cosmid T12B5.1 G-2	8	1.00	0.50	0.75	0.00	0.50
	6	1.00	0.30	0.65	0.00	0.70
Mus musculus Domesticus	10	0.66	1.00	0.83	0.34	0.00
LOCUS MUSBNP	8	0.66	0.66	0.66	0.34	0.34
	6	1.00	0.43	0.71	0.00	0.57
Caenorhabditis elegans	10	0.25	1.00	0.62	0.75	0.00
C30C11 Gene-1	8	0.50	1.00	0.75	0.50	0.00
	6	1.00	1.00	1.00	0.00	0.00

Table 3
Summary of performance analysis of data from Table 1 by AR-YW method.

Gene	Threshold	Prediction			Measures	
	(dB)	S_N	S_P	$(S_N + S_P)/2$	M_R	W_R
Plasmodium knowlesi	-6	1.00	1.00	1.00	0.00	0.00
LOCUS AF065986	-7	1.00	0.66	0.83	0.00	0.34
	-8	1.00	0.33	0.66	0.00	0.67
Callithrix pygmaea	-6	0.33	1.00	0.66	0.67	0.00
LOCUS CBUEGLOBIM	-7	0.66	0.40	0.53	0.34	0.60
	-8	1.00	0.30	0.65	0.00	0.70
Caenorhabditis elegans	-6	0.66	1.00	0.83	0.34	0.00
Cosmid T12B5.1 G-2	-7	1.00	0.60	0.80	0.00	0.40
	-8	1.00	0.50	0.75	0.00	0.50
Mus musculus Domesticus	-6	0.33	0.50	0.41	0.67	0.50
LOCUS MUSBNP	-7	0.66	0.50	0.58	0.34	0.50
	-8	1.00	0.50	0.75	0.00	0.50
Caenorhabditis elegans	-6	0.25	1.00	0.62	0.75	0.00
C30C11 Gene-1	-7	1.00	0.80	0.90	0.00	0.20
	-8	1.00	0.40	0.70	0.00	0.60

Table 4
Summary of performance analysis of data from Table-1 by proposed PC-AR method.

Gene	Threshold	Prediction			Measures	
	(dB)	S_N	S_P	$(S_N + S_P)/2$	M_R	W_R
Plasmodium knowlesi	4	1.00	1.00	1.00	0.00	0.00
LOCUS AF065986	2	1.00	0.80	0.90	0.00	0.20
	0	1.00	0.36	0.68	0.00	0.54
Callithrix pygmaea	4	1.00	1.00	1.00	0.00	0.00
LOCUS CBUEGLOBIM	2	1.00	0.43	0.71	0.00	0.57
	0	1.00	0.43	0.71	0.00	0.57
Caenorhabditis elegans	4	1.00	1.00	1.00	0.00	0.00
Cosmid T12B5.1 G-2	2	1.00	0.60	0.80	0.00	0.40
	0	1.00	0.43	0.71	0.00	0.57
Mus musculus Domesticus	4	0.66	1.00	0.83	0.34	0.00
LOCUS MUSBNP	2	0.66	1.00	0.83	0.34	0.00
	0	1.00	0.50	0.75	0.00	0.50
Caenorhabditis elegans	4	0.75	1.00	0.87	0.25	0.00
C30C11 Gene-1	2	1.00	0.66	0.83	0.00	0.34
	0	1.00	0.33	0.66	0.00	0.67

4. Evaluation criterion

4.1. Performance comparison of proposed method with existing method

4.1.1. Performance comparison by receiver operating characteristics

The definitions of sensitivity (S_N), specificity (S_P), miss rate (M_R) and wrong rate (W_R):

$$\text{Sensitivity} \quad S_N = T_P/(T_P + F_N) \tag{11}$$

$$\text{Specificity} \quad S_P = TN/(T_N + F_P) \tag{12}$$

$$\text{Miss rate} \quad M_R = M_E/A_E \tag{13}$$

$$\text{Wrong rate} \quad W_R = W_E/P_E \tag{14}$$

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 and 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 which are 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. Sensitivity (S_N) is the proportion of coding nucleotides that have been correctly predicted as coding. Specificity (S_P) is the proportion of non-coding nucleotides that have been correctly predicted as non-coding (T_N). Since frequency of non-coding nucleotides in DNA is much greater than frequency of coding nucleotides i.e. $T_N \gg F_P$ giving very high non-informative value of S_P. Therefore traditionally specificity (S_P) is the proportion of predicted coding nucleotides that are actually coding as mentioned in Eq. (15).

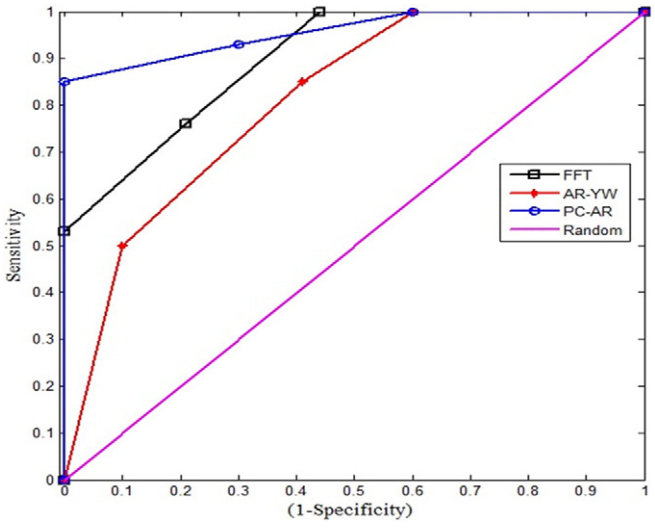

Fig. 12. Plot of random classifier and ROC curves for various genes from Table 1 by FFT, AR-YW and PC-AR techniques.

Table 5
Area under ROC curves associated with the predictors.

Area under ROC curve			
Dataset	Miscellaneous	Burset and Guigo	HMR 195
AR-YW classifier	0.81	0.70	0.71
FFT	0.89	0.76	0.78
PC-AR classifier	0.96	0.82	0.85
Random classifier	0.50	0.50	0.50

$$\text{Specificity} \quad S_P = T_P/(T_P + F_P) \tag{15}$$

The average value of SN and SP gives the overall exon sensitivity and specificity Meher et al., 2011; Roy and Barman, 2014). The plots of existing and proposed methods reflect the superiority of proposed technique over the conventional methods because the peaks obtained with proposed algorithm are sharp, well defined unambiguous and noise free. The threshold values for performance analysis of FFT, AR-YW and PC-AR methods have been chosen judiciously. Analysis summary of FFT (modified periodogram), AR-YW and PC-AR approaches are shown in Tables 2, 3 and 4 respectively.

In all the above examples cited the proposed method shows better result than the existing methods giving higher value of sensitivity, specificity and their average as well as lower value of miss rate and wrong rate. In order to assess and evaluate the performance of the proposed method over the existing methods the ROC curves are obtained. It is a representation of the prediction accuracy of separation of exons and introns in the gene. The ROC curve relates the true positive hits as a function of false positive hits of the exon intron separation for varying threshold values. To draw the ROC curves we initially set the threshold value to the smallest PSD peak.

All the values greater than the threshold were considered to be protein-coding areas while the lower values were considered to be intronic or non-coding. Then T_P, F_P and F_N values were computed. Sensitivity and specificity were calculated as per Eqs. (11) and (15). Then higher thresholds were set consecutively. To compare the performance of PC-AR with FFT and AR-YW, ROC showing true positive ratio (Sensitivity) versus false positive ratio (1-Specificity) were plotted for the sample datasets from Table 1 as shown in Fig. 12. The closer the ROC curve to the diagonal, the less effective is the method in discriminating exons from introns. More steep is the curve towards the vertical axis and then across, the better is the method. ROC curves of all genes tested

Table 6
Optimum DM of various genes by FFT/AR-YW/PC-AR method.

Organism	DM FFT	DM AR-YW	DM PC-AR
Plasmodium knowlesi LOCUS AF065986	1.12	1.04	1.78
Callithrix Pygmaea LOCUS CBUEGLOBIM	1.00	0.83	1.42
Caenorhabditis elegans Cosmid T12B5.1 G-2	1.55	0.89	4.66
Mus musculus Domesticus LOCUS MUSBNP	0.95	0.90	2.2
Caenorhabditis elegans C30C11 Gene-1	1.10	0.92	1.3

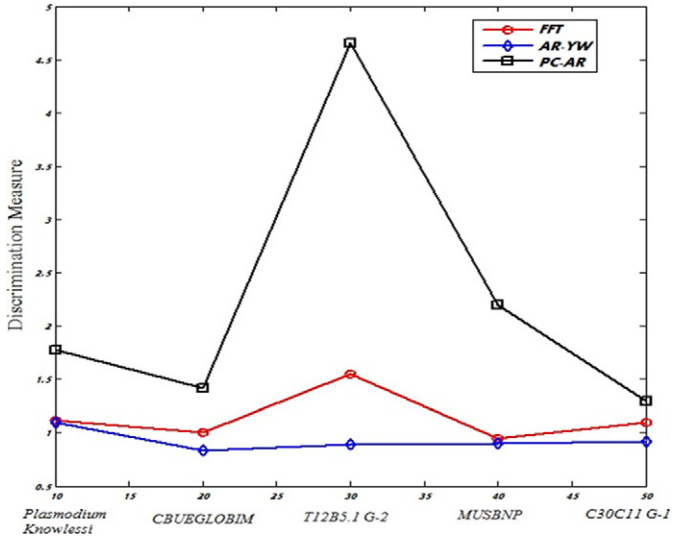

Fig. 13. Plot of discrimination measure from Table 6 by FFT, AR-YW and PC-AR methods for various genes from Table 1.

by both the classifiers exhibited superiority of PC-AR method over traditional methods. A more precise way of evaluating the performance by ROC is to calculate AUC. The closer the area to 0.5, the poorer is the method and closer to 1.0, the better is the method. Areas under the proposed and existing methods for all datasets are given in Table 5. It reveals that the PC-AR method of exon prediction outperforms FFT and AR-YW method.

4.1.2. Performance comparison by discrimination measure

Discrimination measure (DM) is another criterion to differentiate between coding and non-coding regions based on period-3 property of estimated power spectrum defined as:

$$\mathrm{DM} = \frac{\text{Least Peak Amplitude in Coding Regions}}{\text{Highest Peak Amplitude in Non-coding Regions}}. \quad (16)$$

Table 7
Summary of optimal performance of Burset and Guigo dataset by FFT/AR-YW/PC-AR method.

Gene name and accession no.	DSP methods	Prediction		Measures	
		SN	SP	MR	WR
PP32R1, AF008216, *Homo sapiens*	FFT	1.00	1.00	0.00	0.00
	AR-YW	1.00	0.66	0.00	0.34
	PC-AR	1.00	1.00	0.00	0.00
ALOEGLOBIN L25370, *Alouatta Belzebul* epsilon–globin gene	FFT	0.66	1.00	0.34	0.00
	AR-YW	0.66	0.43	0.34	0.57
	PC-AR	1.00	0.75	0.00	0.25
Humbetgloa, L26462, human betaglobin	FFT	1.00	0.66	0.00	0.34
	AR-YW	1.00	0.60	0.00	0.40
	PC-AR	1.00	1.00	0.00	0.00
AGU04852, U04852, *Ateles geoffroyi* haptoglobin (Hp) gene	FFT	0.60	0.66	0.40	0.34
	AR-YW	0.60	0.40	0.40	0.60
	PC-AR	0.80	0.50	0.20	0.50
G101 U12024, *Astyanax mexicanus* green opsin gene	FFT	0.66	0.66	0.34	0.34
	AR-YW	0.63	0.38	0.37	0.62
	PC-AR	0.83	0.75	0.17	0.25
HUMCBRG, M62420, carbonyl reductase gene	FFT	0.66	0.66	0.34	0.34
	AR-YW	1.00	0.66	0.00	0.34
	PC-AR	1.00	1.00	0.00	0.00
BOVANPA M13145, bovine atrial natriuretic peptide	FFT	0.66	0.50	0.34	0.50
	AR-YW	0.50	0.75	0.50	0.25
	PC-AR	1.00	1.00	0.00	0.00

Table 8
Summary of optimal performance of HMR 195 dataset by FFT/AR-YW/PC-AR methods.

Gene	DSP methods	Prediction		Measures	
		S_N	S_P	M_R	W_R
FABP3, U17081, human fatty acid binding protein	FFT	0.75	0.75	0.25	0.25
	AR-YW	0.75	0.66	0.25	0.34
	PC-AR	1.00	1.00	0.00	0.00
SIX3, AF092047, *Homo sapiens* homeobox protein	FFT	1.00	1.00	0.00	0.00
	AR-YW	1.00	0.50	0.00	0.50
	PC-AR	1.00	1.00	0.00	0.00
Osteomodulin, AB009589, human gene for osteomodulin	FFT	1.00	1.00	0.00	0.00
	AR-YW	1.00	0.40	0.00	0.60
	PC-AR	1.00	0.66	0.00	0.33
KIP, AB021866, *Homo sapiens* KIP gene	FFT	0.71	0.50	0.29	0.50
	AR-YW	1.00	1.00	0.00	0.00
	PC-AR	1.00	1.00	0.00	0.00
CLDN3, AF007189, *Homo sapiens* Claudin3 ns	FFT	1.00	0.66	0.00	0.34
	AR-YW	0.75	0.60	0.25	0.40
	PC-AR	1.00	1.00	0.00	0.00
mafG, AB009693, *Mus musculus* gene for mafG	FFT	1.00	0.50	0.00	0.50
	AR-YW	1.00	0.50	0.00	0.50
	PC-AR	1.00	1.00	0.00	0.00
Dp19, AF061327, *Homo sapiens* cyclin-dependent kinased4 inhibitor	FFT	0.50	0.50	0.50	0.50
	AR-YW	0.80	0.50	0.20	0.50
	PC-AR	1.00	1.00	0.00	0.00
AF064081, *Mus musculus* alpha-sarcoglycan gene	FFT	1.00	0.75	0.00	0.25
	AR-YW	0.66	0.66	0.34	0.34
	PC-AR	1.00	1.00	0.00	0.00

Discrimination measures at optimum p values by all the methods for various genes are mentioned in Table 6. If (DM > 1) all exons are well defined and there is no scope of false prediction. If (DM < 1) there is at least one exon not having enough signal power to be distinguished from non-coding region. Fig. 13 shows that the DM for PC-AR estimator is much higher than FFT and AR-YW spectral estimators. See Tables 7 and 8.

The proposed technique is further validated upon Barset and Guigo (1996) (http://genome.imim.es/databases/genomics96) and HMR 195 dataset (prepared by Sanja Rogic) (http://www.cs.ubc.ca/~rogic/evaluation) which are often used as standard benchmark. It is observed that the proposed algorithm offers higher prediction accuracy.

The lists of genes whose datasets have been studied and analyzed by various DSP approaches are shown in respective tables. Tables 7 and 8 summarize the simulation result of genes from Burset and Guigo and HMR 195 datasets respectively.

5. Validation of proposed method with wavelet packet transform

Wavelet packet transform (WPT) is a generalization of wavelet decomposition that offers a richer signal analysis. Liu and Luan (2014) used WPT as an effective tool for de-noising gene prediction. They combined autoregressive method and WPT, increasing the accuracy of exonic region identification while reducing noise. Liu and Lian used the K-Quaternary Code I (denoted as Code13) technique to convert sequence into numerical signal. They used *Caenorhabditis elegans Cosmid F56F11.4a* gene as test data. In their work presented in Fig. 14[A] it is

Table 9
Discrimination Measure for *C. elegans* Cosmid F56F11.4a Gene.

DSP technique	DM
AR	0.83
FFT	1.12
PC-AR	1.28
Wavelet	0.98

Table 10
Summary of optimal performance of various datasets chosen at random by FFT/AR-YW/PC-AR methods.

Gene Id, accession no. sequence length, coding segment, bp	DSP methods	Prediction			Measure		
		S_N	S_P	$(S_N + S_P) / 2$	M_R	W_R	DM
S. cerevisiae, NC_001146 3000 (609, 682–609, 786, 609, 871–611, 661)	FFT	1.0	0.28	0.64	0.0	0.72	0.83
	AR-YW	0.5	1.0	0.75	0.5	0.0	1.6
	PC-AR	1.0	0.66	0.83	0.0	0.34	1.8
D. melanogaster, NM_170135 1776 (640–1671)	FFT	1.0	0.5	0.75	0.0	0.5	1.1
	AR-YW	1.0	1.0	1.0	0.0	0.0	1.3
	PC-AR	1.0	1.0	1.0	0.0	0.0	2.5
BOVGAS, M31657 1066 (540–750, 896–999)	FFT	1.0	0.22	0.61	0.0	0.78	0.66
	AR-YW	1.0	0.2	0.60	0.0	0.80	1.1
	PC-AR	1.0	0.66	0.83	0.0	0.34	1.2
DMPROTP1, L17007 624 (122–248, 376–425)	FFT	1.0	1.00	1.00	0.0	0.0	1.4
	AR-YW	1.0	1.00	1.00	0.0	0.0	2.1
	PC-AR	1.0	1.00	1.00	0.0	0.0	3.0
Platisthys Flesus, AF135499 1845 (1–123, 228–467, 857–1295, 1408–1589, 1702–1845)	FFT	0.8	0.44	0.62	0.2	0.56	1.0
	AR-YW	1.0	0.625	0.812	0.0	0.37	2.5
	PC-AR	1.0	0.71	0.85	0.0	0.29	4.0
CALEGLOBIM, L25363 1698 (144–235, 364–586, 1399–1527)	FFT	1.0	0.60	0.8	0.0	0.4	1.3
	AR-YW	1.0	0.75	0.875	0.0	0.25	1.1
	PC-AR	1.0	0.75	0.875	0.0	0.25	1.6
PIGAPAI, L00626 3333 (751–793, 975–1128, 1770–2370)	FFT	1.0	0.37	0.68	0.0	0.63	0.88
	AR-YW	1.0	0.60	0.80	0.0	0.40	2.3
	PC-AR	1.0	0.66	0.83	0.0	0.34	6.0
HUMELAFIN, D13156 1878 (247–325, 1185–1459)	FFT	1.0	0.28	0.64	0.0	0.72	0.75
	AR-YW	1.0	0.66	0.83	0.0	0.34	1.1
	PC-AR	1.0	0.66	0.83	0.0	0.34	2.0
Homo sapien beta-globin, AF007546 2128 (180–271, 402–624, 1475–1603)	FFT	1.0	0.5	0.75	0.0	0.5	0.9
	AR-YW	1.0	0.75	0.875	0.0	0.25	1.0
	PC-AR	1.0	0.75	0.875	0.0	0.25	1.5
Homo sapien mutant beta-globin, F059180 1552 (27–118, 249–426)	FFT	1.0	0.2	0.6	0.0	0.8	0.66
	AR-YW	1.0	0.66	0.83	0.0	0.34	1.1
	PC-AR	1.0	0.66	0.83	0.0	0.34	2.0

observed that peaks were corrupted with background noise. The black lines in Fig. 14[A] show predicted coding segments and red lines show actual coding segment positions. We applied PC-AR algorithm to *C. elegans Cosmid F56F11.4a* gene having 8060 bp length test data starting from 7021 bp location. It has five known coding segments between locations 7948–8059 bp, 9548–9877 bp, 11,134–11,397 bp, 12,485–12,664 bp and 14,275–14,625 bp. In our result shown in Fig. 14[B] the red lines at the base of the five peaks indicate predicted coding regions where as the green lines near the X-axis show the actual coding positions. It is evident that the five period-3 spectral peaks

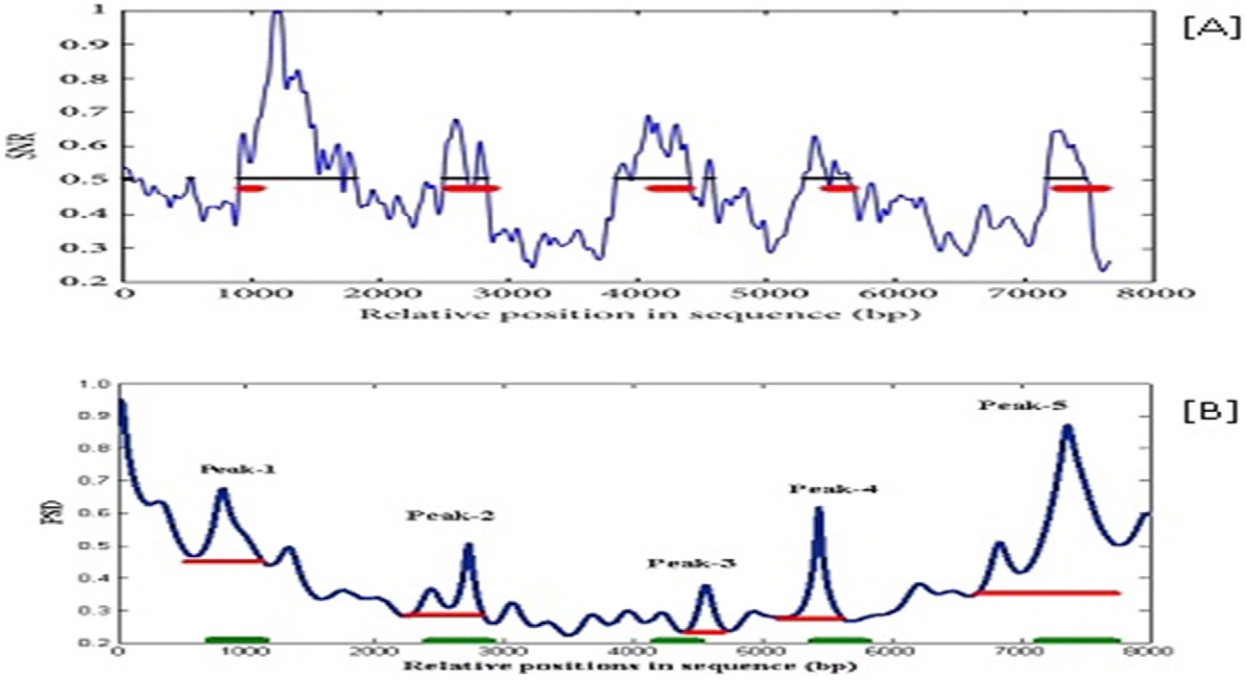

Fig. 14. [A] Shows signal spectrum of *C. elegans* F56F11.4a gene with noise corrupted peaks by applying WPT [Reproduced from article of Liu and Luan (2014)] [B] Shows five distinct noise less peaks in the normalized signal spectrum by proposed PC-AR method for the same gene with better location accuracy.

visible in the specific coding regions are well defined, more or less accurately positioned with minimum noise components. Table 9 indicates discrimination measure for *C. elegans Cosmid F56F11.4a* gene by all the methods discussed, establishing superiority of the PC-AR method over others.

Analysis of NC_001146 gene and other organisms selected randomly from previous literatures are presented in Table 10. In all the examples cited, the proposed method shows higher discriminating measure, sensitivity, specificity with low miss rate and wrong rate compared to FFT and autoregressive Yule–Walker methods.

6. Conclusion

In this article the authors have introduced a technique where Yule–Walker autoregressive process is used in conjunction with principal component analysis to identify protein coding regions in DNA and the performance was compared with existing FFT and AR-YW techniques initially. The novelty of the proposed method is that optimal model order selection is not critical as noise floor is removed by PCA technique prior to PSD estimation. The proposed method was further validated with another model based method using wavelet packet transform for noise reduction. The PSD plots by proposed method display much sharper and well defined peaks in the exon regions compared to the traditional techniques proposed by earlier researchers as well as the newly developed WPT technique. A single indicator sequence based on quaternary code is used for conversion of alphabetic nucleotide sequence into numerals. The superiority of PC-AR method is established with the help of ROC curves, measure of area under characteristics (AUC) and discrimination measure (DM) values.

Competing interests

No competing interest.

Ethical approval

Not required.

References

Anastassiou, D., 2000. Frequency-domain analysis of bio-molecular sequences. Bioinformatics 16 (12), 1073–1081.

Barset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–357 Available: http://genome.imim.es/databases/genomics96.

Chakrabarty, N., Spanias, A., Lesmidis, L.D., Tsakalis, K., 2004. Autoregressive modeling and feature analysis of DNA sequences. EURASIP J. Appl. Signal Process. 1, 13–28.

Guigo, R., Flicek, P., Abri, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E., Reese, M.G., 2006. Review, EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 7 (Suppl. 1), S2.31.

Hayes, M.H., 1996. Statistical Digital Signal Processing and Modeling, Student Edition. John Wiley & Sons, Inc., New York, USA, pp. 440–448http://www.cs.ubc.ca/~rogic/evaluation.

Kakumani, R., Devabhaktuni, V., Ahmad, M.O., 2008. Prediction of Protein-Coding Regions in DNA Sequences Using a Model-Based Approach. IEEE International Symposium on Circuits and Systems, ISCAS, pp. 1918–1921.

Kwan, H.K., Kwan, B.Y.M., Kwan, J.Y.Y., 2012. Novel methodologies for spectral classification of exon and intron sequences. EURASIP J. Appl. Signal Process. http://dx.doi.org/10.1186/1687-6180-2012-50.

Liavas, A.P., Regalia, P.A., 2001. On the behaviour of information theoretic criteria for model order selection. IEEE Trans. Signal Process. 49 (8), 1689–1695.

Liu, G., Luan, Y., 2014. Identification of Protein Coding Regions in the Eukaryotic DNA sequences based on Marple Algorithm and Wavelet Packets Transform. Abstract and Applied Analysis, 2014 (2014), p. 402567 Article ID. Available: http://dx.doi.org/10.1155/2014/402567.

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

National Center for Biotechnology Information (NCBI Gen Bank) [Online] Available: http://www.ncbi.nlm.nih.gov

Oppenheim, A.V., Schafer, R.W., 2013. Discrete-Time Signal Processing. 3rd edition. Dorling Kindersley India Pvt. Ltd., pp. 928–943 Published by.

Praokis, J.G., Manolakis, D.G., 1992. Digital Signal Processing. 2nd edition. Mace Vamillan Publishing Company, pp. 886–896.

Rao, N., Shepherd, S.J., 2004. Detection of 3-periodicity for small genomic sequences based on AR technique. International Conference on Communications, IAC and Systems 2, pp. 1032–1036.

Rosen, G., 2007. Comparison of autoregressive measures for DNA sequence similarity,. IEEE Workshop on Genomic Signal Processing (GENSIPS).

Roy, M., Barman, S., 2014. Effective gene prediction by high resolution frequency estimator based on least-norm solution technique. EURASIP J. Bioinforma. Syst. Biol. http://dx.doi.org/10.1186/1687-4153-2014-2.

Sahu, S.S., Panda, G., 2010. A DSP approach for protein coding region identification in DNA sequence. Int. J. Signal Image Process. 1 (2), 75–79.

Shlens, J., 2003. A Tutorial on Principal Component Analysis, Derivation, Discussion and Singular Value Decomposition. Version-I.

Sussillo, D., Rundaje, A., Anastassiou, D., 2004. Spectrogram analysis of genome. EURASIP J. Appl. Signal Process. 1, 29–42.

Tuqan, J., Rushdi, A., 2008. A DSP based approach for finding the codon bias in DNA sequences. IEEE J. Signal Process. 2 (3), 343–356.

Vaidyanathan, P.P., Yoon, B.J., 2004. The role of signal processing concepts in genomics and proteomics,. J. Frankl. Inst. Eng. Math. Spec. Issue Genomics 351 (1), 111–135.

Yin, C., Yau, S.s.- T., 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol. 247, 687–694.