An efficient signal processing approach in eukaryotic gene prediction

Sitanshu Sekhar Sahu*[1], Ganapati Panda [1]

[1]Department of Electronics and Communication Engineering, National Institute of Technology, Rourkela,Orissa, India

Email: Sitanshu Sekhar Sahu*- sitanshusekhar@gmail.com; Ganapati Panda - ganapati.panda@gmail.com;

*Corresponding author

Abstract

Background: Genomic sequence, structure and function analysis of various organisms have been a challenging problem in bioinformatics. The exponential growth of the repository of the genomic sequences through many scientific and biological communities has put a major thrust into the genome research. The genomic sequence analysis and especially finding the genes and exons within it is a burning problem now-a-days. Accurate and efficient prediction technique of genes and exons has been developing for many years. Various transformation techniques and filtering methods have been used for efficient prediction of exons. Still it needs an improvement in the prediction accuracy.

Results: Aiming at the rapid growth of genome sequence, we propose two new methods based on sliding DFT (SDFT) and adaptive autoregressive modeling for efficient and cost effective prediction of the exons in the gene. Simulation study carried out on many DNA sequences and subsequently reveals that a substantial saving in computation time is achieved by our methods without degrading the performance. The potential of the proposed methods are validated by receiver operating characteristic curve (ROC) analysis.

Conclusion: Two new methods, SDFT and adaptive AR modeling is proposed in this paper and the efficiency is demonstrated by comparing the results with the existing methods and shown its superiority on the consumption timing and the ROC analysis.

Background

The enormous amount of genomic and proteomic data that are available in public domain inspires scientists to process this information for the benefit of the mankind. The genomic information is present in the strands of DNA and represented by nucleotide symbols (A, T, C and G).The segments of DNA molecule called gene is responsible for protein synthesis and contains code for protein in exon regions within it. When a particular instruction becomes active in a cell , the corresponding gene is turned on and the DNA is converted to RNA and then to protein by slicing up to exons (protein coding regions of gene).Therefore finding coding regions in a DNA strand involves searching of many nucleotides which constitute the DNA strand. As the DNA molecule contains millions of nucleotide element, the problem of finding the exons in it is really a challenging task. It is a fact that the base sequences in the protein coding regions of DNA molecules exhibit a period-3 pattern because of the non uniform distribution of the codons in it [1] [2]. In recent past many traditional as well as modern signal processing techniques have been applied to process and analyze these data. Tiwari [1] and Anastassiou [3] have used DFT for the coding region prediction. P.P.vaidyanathan et al. have suggested digital filtering [4] [5] to identify the protein coding region. Later on an autoregressive modeling (AR) was developed [6] [7] [8] for detection of coding regions in small DNA sequences. For large sequence DNA, the AR based method requires more computational efforts. But rapidly acquiring the genomic data demands accurate and fast tools to analyze genomic sequences. In this paper we propose an alternate but efficient and cost effective technique for the identification of the protein coding regions exhibiting period-3 behavior. These new methods employ the sliding DFT (SDFT) and adaptive AR modeling which require substantially less computation and yield comparable performance than the conventional approaches.

Methods

Preliminary spectral measure for coding regions

To perform the gene prediction based on period-3 property the total DNA sequence is first converted into four indicator sequences, one for each base. The DNA sequence D(n) is mapped into binary signals $u_A(n)$,$u_C(n)$,$u_G(n)$ and $u_T(n)$,which indicate the presence or absence of these nucleotides at location n. For example the binary signal $u_A(n)$, attributed to nucleotide A takes a value of 1 at $n = n0$ if $D(n0) = A$, else $u_A(n0)$ is 0. Suppose the DNA sequence is represented as

$D(n) = [ATGATCGCAT]$

Then its numerical representation is given by u$_A(n) = [1001000010]$

u$_C(n) = [0000010100]$

u$_G(n) = [0010001000]$

u$_T(n) = [0100100001]$

Thus $u_A(n) + u_C(n) + u_G(n) + u_T(n) = 1$

In the spectral measure method, the DFTs of the four binary indicator sequences are employed to exploit the 3-base periodicity [9]. Let $U_A[k]$, $U_G[k]$, $U_C[k]$ and $U_T[k]$ represent the Discrete Fourier transform (DFT) of the corresponding binary sequences and is given by

$$U_x\left[k\right] = \sum_{n=0}^{N-1} u_x\left[n\right] e^{-\frac{j2\pi nk}{N}} \tag{1}$$

for $x = A$, C, G or T and $k = 0, 1, \cdots, N-1$

Then the spectral content at k is given by

$$S\left[k\right] = \left|U_A\left[k\right]\right|^2 + \left|U_C\left[k\right]\right|^2 + \left|U_G\left[k\right]\right|^2 + \left|U_T\left[k\right]\right|^2 \tag{2}$$

$S\left[k\right]$ acts as a preliminary indicator of a coding region (exon) giving a peak at the $2\pi/3$ frequency. This coding procedure can be used to detect the probable coding region in the DNA sequence. The exon regions can be identified by evaluating $S[N/3]$ over a window of N samples, then sliding the window by one or more samples and recalculating $S[N/3]$. This process is carried out over the entire DNA sequence. Fig.3 shows how spectral content evolves along the length of DNA sequence. It is necessary that the window length N be sufficiently large (typical sizes are a few hundred eg. 351 to a few thousand) so that the periodicity effect dominates the background noise spectrum [10]. The conventional DFT method involves large computations which poses difficulty for online evolution of protein coding regions. The following section deals with the sliding DFT method, a fast approach in DSP literature for spectral estimation of the DNA sequences.

The sliding DFT algorithm based approach

The sliding DFT algorithm is an improved version of the traditional DFT process where the spectral bin output rate is equal to the input data rate on sample by sample basis with improved computational efficiency. The sliding DFT employs the idea of Goertzel algorithm [11] and computes the DFT spectra through implementing an infinite impulse response (IIR) filter. The Z-transform function of the Goertzel filter is given by:

$$H_G(z) = \frac{1 - e^{-j2\pi k/N} z^{-1}}{1 - 2cos(2\pi k/N) z^{-1} + z^{-2}} \tag{3}$$

where N is no. of samples of which the DFT is to be calculated. The sliding DFT (SDFT) algorithm performs an N-point DFT on time samples within a sliding window. The time window is then advanced by one or more samples and a new N-point DFT is calculated. The importance of this process is that each new DFT is efficiently computed directly from the result of the previous DFT. The principle used in the SDFT is the DFT shifting theorem or the circular shift property [11] [12]. It states that if the DFT of a windowed time domain sequence is $X(k)$, then the DFT of that sequence, circularly shifted by one sample, is $X(k)e^{j2\pi k/N}$, where k is the DFT bin of interest. This process is expressed as

$$P(k) = P(k-1)e^{j2\pi k/N} - x(n-N) + x(n) \tag{4}$$

where $x(n)$ is the new sample to be included in the windowed sequence, $x(n-N)$ is the first sample of the sequence which is to be discarded and $P(k)$, $P(k-1)$ are the new and previous spectral components of the windowed time sequence respectively. This can be represented by an IIR filter and the Z-domain transform function of the sliding DFT filter is given by

$$H_{SDFT}(z) = \frac{1 - z^{-N}}{1 - e^{j2\pi k/N} z^{-1}} \tag{5}$$

A better representation of the algorithm with correct phase and magnitude [13] is given in (6)

$$P(k) = j2\pi k/N \, [P(k-1) - x(n-N) + x(n)] \tag{6}$$

The IIR filter implementation of SDFT is shown in Fig.1 .Each of the four binary indicator sequences within the window is passed through this IIR filter and the corresponding DFT output sequence is obtained. The resulting DFTs are represented as $P_A[k]$, $P_C[k]$, $P_G[k]$ and $P_T[k]$.
Then the spectral content is computed as
$S[k] = \sum |P_x[k]|^2$ where x stands for either of the sequences A, C, T or G. The successive progression of $S[k]$ within the sliding window and the plot of $S[N/3]$ exhibit the coding regions in DNA. Thus the SDFT requires only one complex multiplication and two real additions per output sample. Hence the computational complexity of each successive N-point output is $O(N)$ for the sliding DFT compared to $O(N^2)$ for the DFT. As a result there is substantial computational saving in identifying the coding region by the proposed method.

Autoregressive modeling

The Fourier analysis based methods for detecting the protein coding regions in DNA lose their effectiveness in the case of small as well as standard DNA sequences [6]. These methods are constrained by the frequency resolution and spectral leakage effects of the data record. Hence to overcome this problem model based approaches have been evolved which look the spectral analysis problem differently compared to Fourier Transform methods. The autoregressive (AR) model is one such method which has been proposed for detection for the 3-base periodicity in DNA sequence and to predict the protein coding regions [7]. The AR modeling approach has been used successfully in speech processing and radar signal processing [14]. It is a simple and robust method and requires no a priori knowledge of the sequence to be analyzed and also works well with low signal to noise (SNR) ratio. In AR modeling the observed signal $x(n)$ can be modeled as a linear combination of its M past output values $x(n-k)$ and present input $u(n)$ as

$$x(n) = -\sum_{m=1}^{M} a(m)x(n-k) + u(n) \tag{7}$$

where $a(k)$ represents the coefficient of the model to be estimated. This AR process can be viewed as a recursive all-pole digital filter whose transfer function is

$$H(z) = \frac{1}{1 + \sum_{m=1}^{M} a(m)z^{-k}} \tag{8}$$

In this case the number of filter coefficients is equal to the order of the model and is same as the number of filter poles which is to be determined efficiently. The Yule-Walker or Burg methods are usually used to determine these coefficients using the Levinson-Durbin recursive procedure [14]. The spectral estimation for the AR model given in (8) is

$$S_x(k) = \frac{\sigma^2}{\left|1 + \sum_{m=1}^{M} a(m)e^{j2\pi mk/N}\right|^2} \tag{9}$$

where σ^2 is the variance of the input signal. The power spectrum estimated from (9) at frequency $\theta = 2\pi/3$ is used to predict the protein coding region (exon) in the DNA sequence.

Proposed adaptive AR modeling approach

For achieving on line prediction of gene and exon the computational time needs to be reduced. Further the fixed AR method requires all data to be available simultaneously which is not always feasible. With a motive to alleviate these limitations an adaptive AR model based approach is suggested in this section for efficient prediction. The AR process can be viewed as an adaptive prediction error filter (all-zero filter) that adaptively adjusts its coefficients to flatten the spectrum of the signal to be observed. It is a fact that with a proper learning algorithm like LMS and RLS the weight vector of the adaptive prediction error filter converges to optimal AR coefficients [15].

The LMS prediction error filter

A signal x(n) modeled as a M order AR process can be expressed as

$$\hat{x}(n) = \sum_{m=1}^{M} w_m(n)x(n-m) + e(n) \tag{10}$$

where e(n) is the prediction error and $w1, w2, \cdots wM$ are AR coefficients. The LMS prediction error filter illustrated in Fig. 2 is used to adaptively estimate the optimal AR coefficients by minimizing the mean square value of prediction error.

The weight update equation of the filter is given by

$$W(n) = W(n-1) + \mu\bar{X}(n)e(n) \tag{11}$$

where $e(n) = x(n) - \hat{x}(n)$ is the prediction error

$\hat{x}(n) = \bar{X}^T(n)W(n)$is the prediction of $S(n)$,

$\bar{X}(n) = [x(n-1)x(n-2)\cdots x(n-M)]$,

$W(n) = [w_1(n)w_2(n)\cdots w_M(n)]$ and μ is the step size that determines the rate of converge and stability of weights which lies between 0 to 1.

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

$$S_x(k) = \frac{\sigma_e^2}{\left|1 - \sum_{m=1}^{M} w(m)e^{j2\pi mk/N}\right|^2} \tag{12}$$

where σ_e^2 is the variance of the prediction error signal.

Results and Discussion

Simulation results

To validate the performance of the proposed two methods simulation study is carried out on many short and long DNA sequences. We compare the spectra obtained through simulation from the conventional DFT, the sliding DFT algorithm, the fixed and adaptive AR modelings of the DNA sequence of gene F56F11.4a of C-elegans chromosome III. This gene is used as a benchmark problem for different gene detection techniques and known to have five distinct exons, relative to nucleotide position 7021. The results of the simulation obtained using four different methods of this particular gene are shown in Figs 3, 4, 5 and 6. All these plots exhibit identical coding regions. This indicates that the two newly proposed methods are as potential in predicting the protein coding region as provided by standard DFT and fixed AR methods. The computational complexities involved in the two DFT and SDFT methods are obtained for comparison and are shown in Table 1. The power spectra of the same gene for various window sizes and model orders by fixed AR modeling are obtained through simulation study. It is observed that the combination of window size of N=81 and order M=8 provides the best result as shown in Fig.5. Further the spectra of the same gene by adaptive AR modeling for various window sizes and model orders have also been obtained. It is observed that window size of N=21 and order M=5 provides the best as shown in Fig. 6. From Figs 5 and 6 it is clear that both fixed and adaptive AR method provides identical results, but the computation time associated with the adaptive AR method is very less. Table 2 shows a comparison of the computation time in identifying the coding region required by the four different methods. It indicates that the adaptive AR method provides minimum computation time compared to the other three methods.

Evaluation methods

In order to assess and compare the efficiency of all these methods the receiver operating characteristics (ROC) curves are first obtained. It is a representation of the prediction accuracy of the separation of exons and introns in the gene. In this test four terms are used for the evaluation. These are TPF (True Positive Fraction), FPF (False Positive Fraction), FNF (False Negative Fraction) and TNF (True Negative Fraction). The significance of these terms are

TPF: Truly predicted as coding region

FPF: Falsely predicted as coding region

FNF: Falsely predicted as non coding region

TNF: Truly predicted as non coding region

The ROC curve relates the TPF as a function of FPF of an exon and intron separation method for varying threshold values. The closer the ROC curve to a diagonal, the less effective the method at discriminating between exon and intron. More steep the curve towards the vertical axis and then across, the better is the method. The ROC curves for all the four methods are shown in Fig. 7. A more precise way of evaluating the performance is to calculate the area under the ROC curve. The closer the area to 0.5, the poorer is the method and closer to 1.0, the better is the method. The area under the ROC curve for all the four methods is shown in Table 3. It reveals that the adaptive AR method of exon prediction outperforms other three methods as it offers highest area under the curve.

Conclusions

The paper presents two new efficient approach as based on sliding DFT (SDFT) and adaptive AR modeling for identification of coding region exploring the period-3 behavior. The performance of the new methods is shown to be identical to that obtained by the Fourier transform based method. In addition, the proposed method offers substantial computational advantage over the conventional methods. The comparison of CPU time of different methods to achieve the gene prediction region also reveals the computational advantage same of the proposed adaptive AR method followed by the new SDFT method. Thus in general the two novel methods proposed have distinct computational advantage without sacrificing the quality of gene and exon prediction.

References

1. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R: **Prediction of probable genes by Fourier analysis of genomic sequence**. *CABIOS* 1997, **13**:263–270.
2. Fickett JW: **The gene prediction problem: an overview for developers**. *Computers chem.* 1996, **20**:103–118.
3. Anastassiou D: **Genomic signal processing**. *IEEE signal processing magazine* 2001, **18**:8–20.
4. Vaidyanathan PP, Yoon BJ: **Gene and exon prediction using all pass based filters**. In *workshop on genomic signal processing and statistics, Raleigh, NC, USA* 2002.
5. Vaidyanathan PP, Yoon BJ: **Digital filters for gene prediction application**. In *Proc. Asilomar conference in signal, systems and computers, pacific Gove, calit, USA* 2002:306–310.
6. Rao N, Shepherd SJ: **Detection of 3-periodicity for small genomic sequence based on AR techniques**. In *Int. Conf. On comm., IRC. And Sys*, *Volume 2* 2004:1032–1036.
7. Akhatar M, Ambikairajah E, Epps J: **Detection of Period-3 behavior in genomic sequence using singular value decomposition**. In *IEEE Int. Conf. On Emerging Technologies* 2005:13–17.
8. Akhatar M: **Comparison of gene and exon prediction techniques for detection of short coding regions**. *Int. J. of Inf. Tech., Special issue on Bioinformatics and Biomedical Systems* 2005, **11**:26–35.
9. Trevor WF, Carrira A: **A digital signal processing method for gene prediction with improved noise suppression**. *EURASHIP journal on applied signal processing* 2004, **1**:108–114.
10. Vaidyanathan PP: **Genomic and Proteomics: A Signal Processing's Tour**. *IEEE Circuits and Systems Magazine* 2004, :6–29.
11. Jacobson E, Lyons R: **The sliding DFT**. *IEEE signal processing magazine* 2003, :74–80.
12. Jacobson E, Lyons R: **An update to the sliding DFT**. *IEEE signal processing magazine* 2003, :110–111.
13. Springer T: **Sliding FFT compute frequency spectra in real time**. *EDN magazine* 1998, :161–170.
14. Makhoul J: **Linear prediction : A Tutorial review**. *Proceedings of the IEEE* 1975, **63**:562–580.
15. Wu W, Chen P: **Adaptive AR modeling in white Gaussian noise**. *IEEE transaction on signal processing* 1997, **45**:1184–1192.

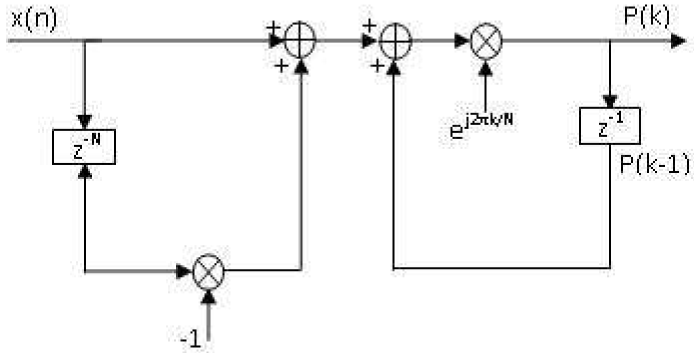

Figure 1: An IIR implementation of the Sliding DFT

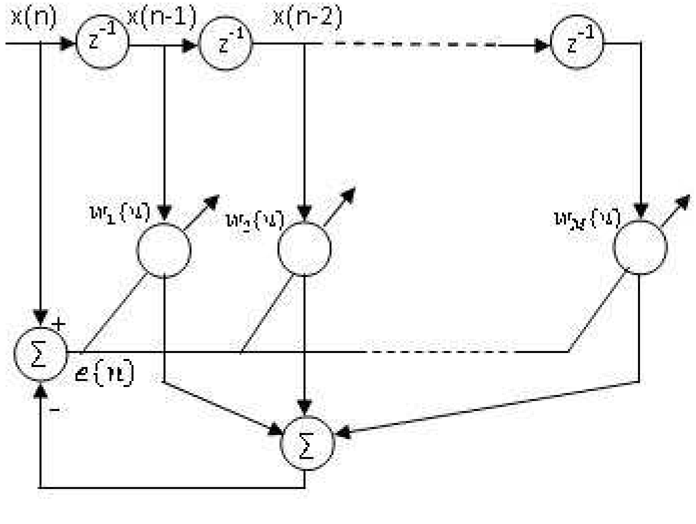

Figure 2: LMS prediction error filter

Method	Complex multiplications	Complex Additions
	(N= window size,	L=length of the DNA sequence)
DFT	L*N^2	L*(N(N-1))
SDFT	L*N	L*(N-1)

Table 1: Comparison of computational complexity of DFT and SDFT

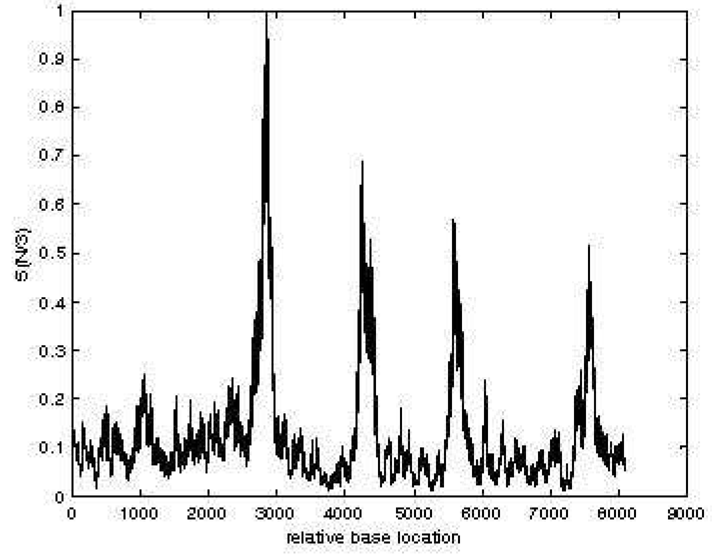

Figure 3: Spectral plot of S[N/3] for gene F56F11.4a in the C-elegans chromosome III using DFT

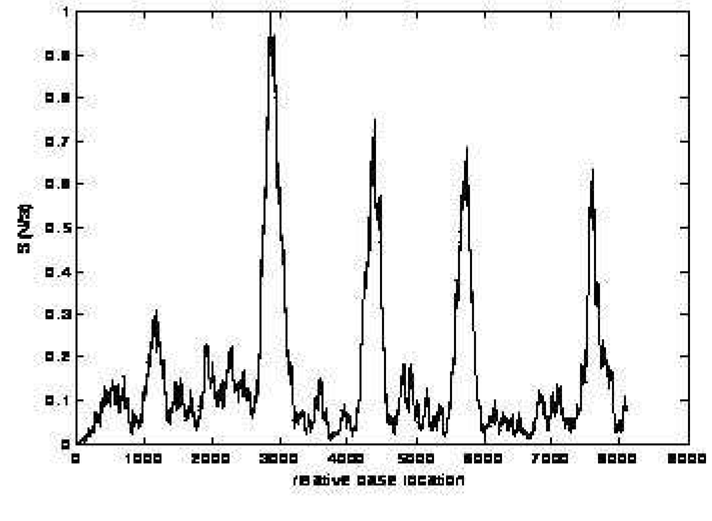

Figure 4: Spectral plot of S[N/3] for gene F56F11.4a in the C-elegans chromosome III using SDFT

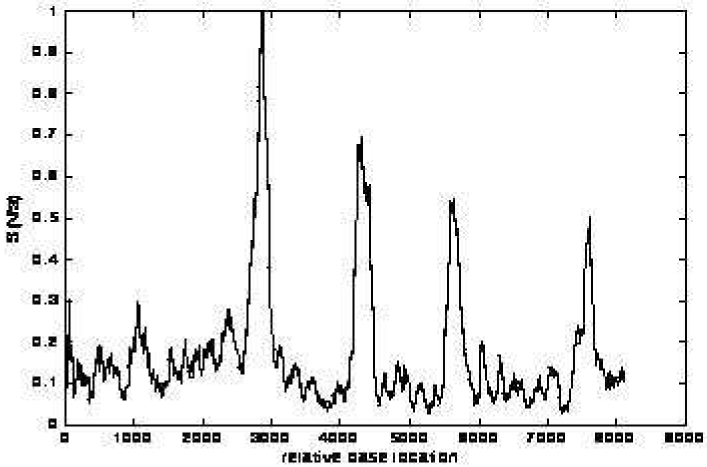

Figure 5: Spectral plot of gene F56F11.4a in the C-elegans chromosome III using fixed AR modeling

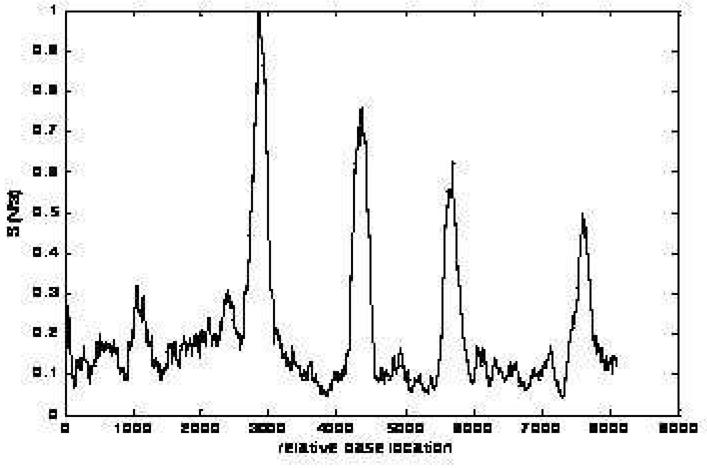

Figure 6: Spectral plot of gene F56F11.4a in the C-elegans chromosome III using adaptive AR modeling

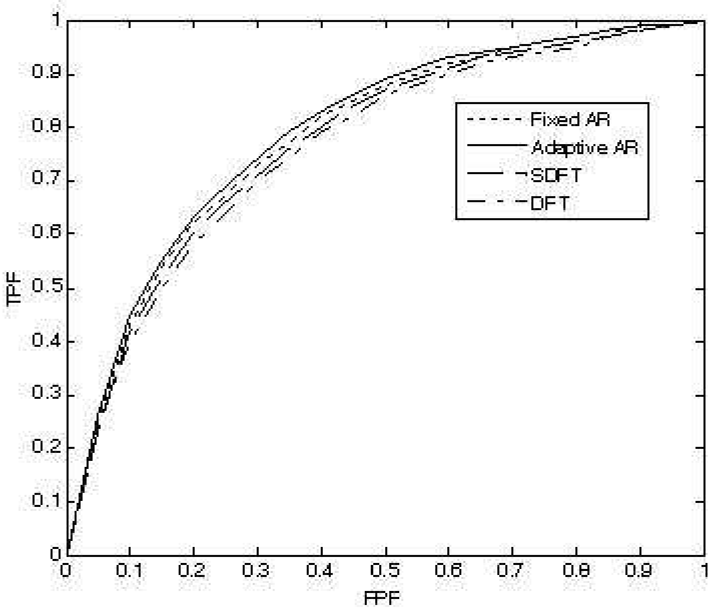

Figure 7: ROC curves of all the predictors

Method	CPU time (sec)
DFT	47
SDFT	14
Fixed AR	36
Adaptive AR	11

Table 2: Comparison of computation time of different methods

Method	Area Under the ROC Curve
DFT	0.7725
SDFT	0.7755
Fixed AR	0.7835
Adaptive AR	0.7875

Table 3: Area under the ROC curve associated with the predictors