Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 402567, 14 pages
http://dx.doi.org/10.1155/2014/402567

Research Article

Identification of Protein Coding Regions in the Eukaryotic DNA Sequences Based on Marple Algorithm and Wavelet Packets Transform

Guangchen Liu and Yihui Luan

School of Mathematics, Shandong University, Jinan, Shandong 250100, China

Correspondence should be addressed to Yihui Luan; yhluan@sdu.edu.cn

Received 11 April 2014; Revised 30 June 2014; Accepted 1 July 2014; Published 15 July 2014

Academic Editor: Caihong Li

The identification of protein coding regions (exons) plays a critical role in eukaryotic gene structure prediction. Many techniques have been introduced for discriminating between the exons and the introns in the eukaryotic DNA sequences, such as the discrete Fourier transform (DFT) based techniques, but these DFT-based methods rapidly lose their effectiveness in the case of short DNA sequences. In this paper, a novel integrated algorithm based on autoregressive spectrum analysis and wavelet packets transform is presented to improve the efficiency and accuracy of the coding regions identification. The experimental results show that the new algorithm outperforms the conventional DFT-based approaches in improving the prediction accuracy of protein coding regions distinctly by testing GENSCAN65, HMR195, and BG570 benchmark datasets.

1. Introduction

Deoxyribonucleicacid (DNA) sequence consists of genic and intergenic regions. Identification of protein coding regions is an elementary but very important problem in bioinformatics because the exonic regions code for amino acids. So learning the primary structure of a protein leads to studying and analyzing the secondary and tertiary structures of a protein in addition to protein function. Once we could clearly know the structure and function of a protein, we can design drugs, cure diseases, improve crop productivity, and synthesize biofuel. In addition, coding regions represent the conserved part of genomes. On the other hand, predicting conserved regions is also important to study evolution and predict phylogenetic trees [1, 2]. Nowadays, the rapid growth of raw genome sequence data requires efficient biological interpretations, but biological experiments for gene identification in DNA sequences are costly to conduct, so there is still a real demand for accurate and fast tools to analyze these sequences, especially to find genes and determine their functions [3, 4].

All living organisms can be divided into two categories according to their fundamental cell structures: prokaryotes and eukaryotes. In prokaryotes, the coding genes, which are in charge of protein synthesis, are long and continuous (that is open reading frames (ORFs)). But in eukaryotes, genes consist of coding segments interrupted by long noncoding segments. These coding segments are termed as exons and noncoding segments as introns (Figure 1). In case of human eukaryotes only 3% of DNA sequence is coding [5, 6], so it is a challenging task to identify the protein coding regions (exons).

Genomic sequence processing has been an active area of research for the past twenty years and has increasingly attracted the attention of many researchers, and a number of methods have been proposed to predict the protein coding regions [2]. These methods can be divided into two groups in comprehensive categorization [7–9]: model-dependent methods and model-independent methods (or filter-based methods). Model-dependent methods are built upon some a priori information usually gathered from database of previously known organisms' genomics, while model-independent methods do not assume such a priori information. Some model-dependent methods such as hidden Markov model (HMM) [10, 11], support vector machine (SVM), and neural network [12] have been successfully used to find splice sites and identify protein coding regions.

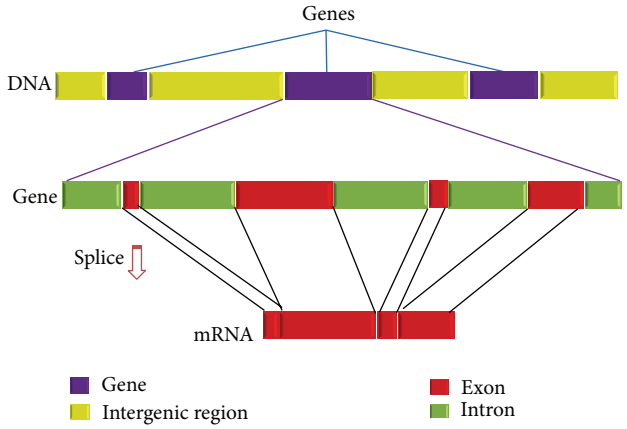

Figure 1: The DNA structure of eukaryotes and the splicing process. This figure shows that eukaryotic DNA consists of genic and intergenic regions, and the exon regions are interrupted by introns in eukaryotic DNA. Generally, the introns are much longer than the exons.

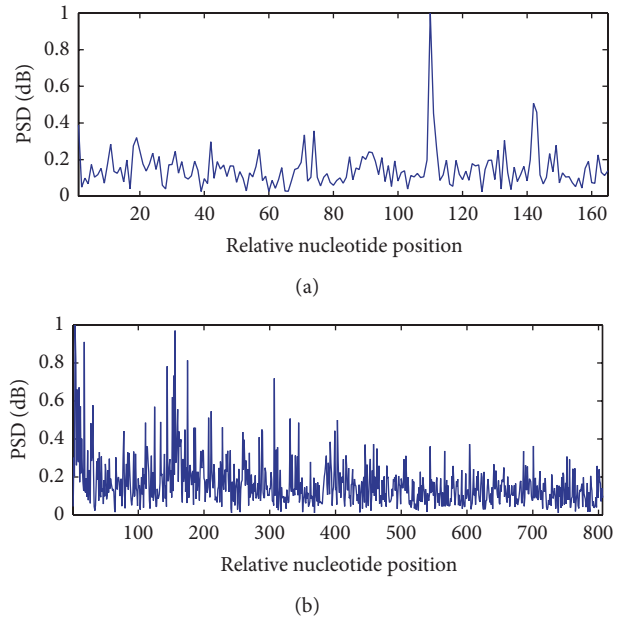

Figure 2: The power spectrum density (PSD) calculated by Voss-DFT of exon and intron from the gene F56F11.4. For the symmetry characteristics, only the first half of PSD is presented. (a) The PSD of an exon with the length of 330 basepair (bp), and the nucleotide position from 2528 to 2857, and the three-base periodicity (TBP) demonstrating peak at frequency $k = N/3$, where N is the length of the exon. (b) The PSD of an intron with the length of 1612 bp, and the nucleotide position from 5644 to 7255, but there is no remarkable peak in the whole region.

Though these model-dependent methods perform more precisely by means of a priori information to train the classifiers, nevertheless, the coding regions may not be represented on the available datasets but exist in the sequenced organism [7, 9]. In such situations, model-independent methods based on digital spectral analysis, which convert sequences text into numeric signal, have been proposed in recent years to detect the coding regions [2]. For most of DNA sequences, one of the principal features is the fact that the dominant signal in coding regions of genomic sequences exhibits a three-base periodicity (TBP) which is evidenced as a sharp peak at frequency $f = 1/3$ in the PSD [13–15], but this behavior is not found in other parts of the DNA (including the introns) (Figure 2). The origin of the TBP in protein coding sequences derives from the triplet nature of the codon, and the reasons for this distinction lie in the unequal usage of codons (codon bias) in coding regions [14, 16]. Meanwhile this phenomenon caused lots of background noise, which leads to more difficulty in finding exons in DNA sequences [7, 17–19]. Using this property, several model-independent techniques, mainly DFT-based methods, have been mentioned in the recent literatures. Tiwari et al. [14] firstly used the DFT to calculate the PSD; then a fixed-length window was used to move on to the numerical sequence to determine the exonic regions. This technique is also named sliding discrete Fourier transform (SDFT). Rao et al. [20] proposed an efficient sliding window strategy based on the SDFT (also named the periodogram method) to identify the accurate location of the protein coding regions, and their algorithm could increase the location accuracy greatly than Tiwari's [14]. Digital filters such as antinotch filter [21] and notch filter [19] with the central frequency of $2\pi/3$ were used to remove the background noise, and then SDFT technique was utilized to find exons.

The aforementioned DFT-based spectrum analysis techniques may be roughly categorized as one kind of nonparametric (also named classical spectrum estimation) methods, that is, the periodogram method. This method has the advantage of possible implementation using the fast Fourier transform (FFT) and has made obvious progress in exons finding area [22]. But these methods rapidly lose effectiveness in the case of short DNA sequences because they lead to many false alarms and false dismissal errors [15]. So parametric spectrum methods such as AR modeling were developed for detection of coding regions in small DNA sequences [15, 23–27]. And AR modeling has been proved to be able to produce stronger PSD and weaker artifacts than DFT-based methods [26–28], especially performing well in finding the small size coding regions.

In this paper, a novel technique based on Marple algorithm of AR PSD and wavelet packets transform is presented to identify the protein coding regions in eukaryotic DNA sequences. This method firstly employs a mapping method to convert the DNA sequences into numerical sequences; then the sequences are passed through a bandpass filter to enhance the TBP characteristics. After that, by taking the numerical sequence as the observed signal of an AR model, the efficient Marple algorithm is utilized to estimate the PSD of the AR model by calculating the parameters of the Yule-Walker equations. Then wavelet packets transform (WPT) technique is employed to reduce the the background noise of the PSD. Finally, similar to the SDFT [14], the PSD at frequency $\theta = 2\pi/3$ is used to identify the protein coding regions after denoising by WPT. We show that the new algorithm yields comparable performance in improving

the exonic identification accuracy than the conventional approaches.

The remainder of the paper is organized as follows. In Section 2, the datasets used are described, together with a brief explanation of the methods that are involved in this study. The principles of the Marple algorithm for AR PSD estimation and the WPT for noise reduction are expressed in detail. The numerical representations of a DNA sequence and the bandpass filter are also detailed in this section, as well as the evaluation criteria at nucleonic level. Section 3 presents the results of the benchmark datasets tests that demonstrate the performance of the proposed algorithm. Also the results are analyzed in this section. Finally, the most significant findings that emerge from this study are summarized in Section 4.

2. Materials and Methods

2.1. Datasets. In this subsection, several widely used benchmark datasets will be described for the purpose of comparing the performance of different algorithms in identifying exonic regions. They are listed in the following paragraphs.

The gene sequence F56F11.4 (Genbank old number AF009962, new number FO081497, http://www.ncbi.nlm.nih.gov/nuccore/FO081497) is on chromosome III of *Caenorhabditis elegans* which is a free living nematode, about 1 mm in length, and lives in temperate soil environment. It has five distinct exons with the nucleotide position in the complete sequence: 7949–8059, 9548–9877, 11134–11397, 12485–12664, and 14275–14625. As former literatures [7, 29], we select one part from the complete sequence from nucleotide positions 7021 to 10580, so it just covers the aforementioned five exons. For convenience, in the following analysis, we all use the position that is relative to the first nucleotide position of the selected part sequence, not to the real position in the complete sequence.

In order to demonstrate the performance of our proposed algorithm, we also apply it on three benchmark datasets HMR195 [30], BG570 [31], and GENSCAN65 [32]. HMR195 consists of 195 mammalian (including human, mouse, and rat) sequences, totally 2649 exons, with exactly one complete either single-exon or multiexon gene. BG570 is a genomic test datasets of 570 single gene vertebrate sequences, totally 948 exons, prepared by Burset and Guigó [31]. Datasets HMR195 and BG570 can be available from http://www.imtech.res.in/raghava/genebench/datasets.html. GENSCAN65 contains 65 selected human genome sequences which comprise 381 exons from 2 bp to 1210 bp [32]. For convenience, all sequences contain exactly one gene which starts with the "ATG" initial codon and end with a stop codon (TAA, TAG, or TGA) [30]. GENSCAN65 dataset can be available from http://www.imtech.res.in/raghava/genebench/datasets/Kulp-Reese/Human/.

2.2. Identification of Exonic Regions Based on Marple Algorithm and Wavelet Packets Transform. In this section, a novel integrated algorithm using Marple algorithm and WPT denoising technique is proposed for the identification of protein coding regions. We divide the integrated algorithm into five steps (Figure 3). The block diagram of our proposed universal integrated algorithm is shown in Figure 3, and the procedure of our algorithm is as follows.

(S1) Convert the DNA sequence into numerical sequence using Code13 mapping method.

(S2) Enhance the TBP characteristics of the numerical sequence using an FIR band pass filter.

(S3) Extract the TBP components using the Marple algorithm with proper model order. Similar to the SDFT [14], firstly, a sliding window with length N is determined, and in the window the Marple algorithm is used to calculate the PSD of the windowed sequence. Then the PSD at frequency $\theta = 2\pi/3$, that is, the PSD at $N/3$, is extracted and then divided by the mean PSD of the windowed sequence, so we obtain a ratio, which is referred to as the signal to noise ratio (SNR). Sliding the window along the sequence one by one (i.e., one position by one time), this successive progression and the plot of SNR exhibit the coding regions in DNA.

(S4) Remove the noise effect of SNR by wavelet packets transform.

(S5) Classify (or predict) the protein coding regions according to the optimal threshold.

The following subsections give the detailed presentation of the aforementioned steps, respectively.

2.2.1. Numerical Representation of DNA Sequences. It is an important foundation to convert the DNA sequences into digital signals because it opens the possibility to employ all kinds of powerful DSP techniques for analyzing of genomic data and reveals features of chromosomes [7]. For example, once the DNA sequence has been converted into digital signal while retaining the biological meaning of the represented information, we can utilize the spectral analysis technique to find the exons. That is, coding regions exhibit the three-base periodicity property in the spectral domain which is less apparent in sequences other than exon sequences and can therefore be used to detect exon sequences and to distinguish exonic regions from intronic regions in genomic sequences [33].

There are a number of representations for nucleotide sequences [2, 33, 34]. Kwan et al. [33] reviewed the former widely used numerical representation methods and developed several novel mapping methods. Totally 17 numerical represented methods were compared for exon finding purpose in [33], and they concluded that the Code13 (named K-Quaternary Code I) method offered an attractive performance. Abo-Zahhad et al. [34] also reviewed the published mapping techniques and broadly classified them into two major groups: fixed mapping techniques and physicochemical property based mapping techniques. The former group include the famous Voss method [17] and the following developed methods, such as the tetrahedron [35], the complex [36], and the integer [37] methods. And the latter group is comprised of the electron-ion interaction potential (EIIP)

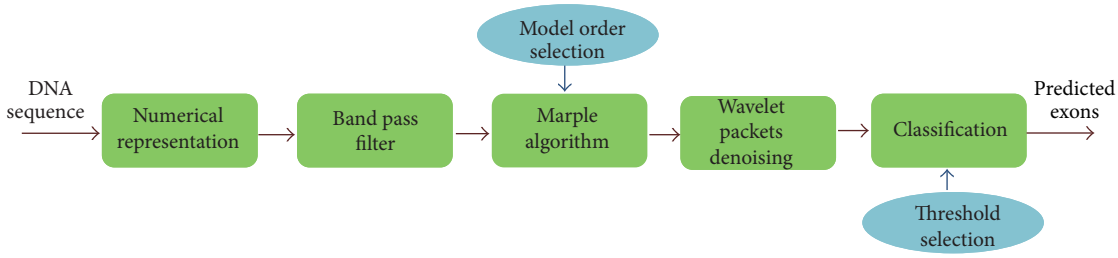

Figure 3: Block diagram of the proposed algorithm.

[38], the paired numeric (PN) [39], the Z-curve [40], the structure profile (SP) [41] representations, and so forth.

In this paper, we use the K-Quaternary Code I (denoted as Code13) technique to convert sequence into numerical signal. In the mean time, for comparison purpose, we select four conventional mapping methods from the aforementioned methods for the following spectral analysis, that is, in shorthand form, the Voss method, the EIIP method, the SP method, and the PN method. The detailed representation methods are described as follows.

Voss Method. Perhaps the earliest and most popular mapping of DNA is the binary or Voss method [17, 39], which represents DNA with four binary indicator sequences $x_A[n], x_C[n], x_G[n]$, and $x_T[n]$. The presense of a nucleotide at a particular base pair position is represented by 1, and the absence of it is represented by 0. For example, the binary representation of DNA sequence S = gctatctatc is given by $x_A[n] = \{0,0,0,1,0,0,0,1,0,0\}$, $x_C[n] = \{0,1,0,0,0,1,0,0,0,1\}$, $x_G[n] = \{1,0,0,0,0,0,0,0,0,0\}$, and $x_T[n] = \{0,0,1,0,1,0,1,0,1,0\}$.

EIIP Method. In EIIP method [38], the electron-ion-interaction potential associated with each nucleotide is used for mapping of the DNA sequence. The EIIP values for the nucleotides are $A = 0.1260$, $C = 0.1340$, $G = 0.0806$, and $T = 0.1335$. The aforementioned sequence S can be converted into one numerical sequence $x[n] = \{0.0806, 0.1340, 0.1335, 0.1260, 0.1335, 0.1340, 0.1335, 0.1260, 0.1335, 0.1340\}$.

SP Method. In the structure profile method [41, 42], the structural information of physical properties of DNA molecule are utilized for mapping nucleotide sequence to numerical sequence. These properties are DNA-bending stiffness, duplex-free energy, duplex disrupt energy, and propeller twist, etc. These structural profiles are calculated according to the conversion tables [42] with step size of one along the DNA sequence, which transforms two nucleotides at a step. So a DNA sequence can be converted into four numerical sequences, but the lengths of the numerical sequences will be one unit shorter than the DNA sequence. Also taking the S sequence as an example, one of whose four profiles, the DNA-bending stiffness profile, is given by $x[n]$=$\{85, 60, 20, 20, 60, 60, 20, 20, 60\}$. The other three numerical sequences can be obtained according to [42].

PN Method. The paired numeric method [39, 42] is based on the statistical evidence that exons are rich in nucleotides C and G, while introns are rich in nucleotides A and T. The PN technique assigns the values +1 and −1 to the presence of the A-T and C-G nucleotides. So the S sequence is mapped into a single sequence $x[n] = \{-1,-1,1,1,1,-1,1,1,1,-1\}$.

Code13 Method. The Code13 method [33, 39, 42] is a kind of 1-sequence complex-value numerical representations, and in this representation, the features of the nucleotides have been retained by translating them into numerical properties. The Code13 method assigns the values 1, −1, $-j$, and j to the presence of the four nucleotides A, C, G, and T, respectively, where j is the imaginary unit ($j^2 = 1$). Then according to the Code13 method, the S sequence is mapped into a single sequence $x[n] = \{-j,-1,j,1,j,-1\,j,1,j,-1\}$.

2.2.2. Emphasizing TBP of the Numerical Sequences by FIR Bandpass Filter. In order to emphasize the three-base property in the protein coding regions, the numerical sequences are passed through an FIR bandpass filter with a Hamming window, whose order is 8 and central frequency is $2\pi/3$. Lack of distortions in FIR filters is one reason for their preferred use over IIR filters in medical applications [22, 29].

2.2.3. Autoregressive Spectrum Estimation Using Marple Algorithm. The spectrum estimation techniques available may be categorized as nonparametric (also named classical spectrum estimation) and parametric. The nonparametric methods include the periodogram, the Bartlett and Welch modified periodogram, and the Blackman-Tukey methods. All these methods have the advantage of possible implementation using the fast Fourier transform (FFT), but with the disadvantage in the case of short data lengths of limited frequency resolution, and the requirement for windowing to reduce the spectral leakage. Parametric methods on the other hand can provide high resolution, applicability to short data lengths, and avoidance of spectral leakage, scalloping loss, spectral smearing, and window biasing effects [22]. Its disadvantage lies in being computationally efficient, and parameter methods mainly include three methods, Yule-Walker autoregressive method, Burg method, and the Marple algorithm.

The idea of the AR spectrum analysis is that the digitized signal is modeled as an AR time series plus a white noise error term. The spectrum is then obtained from the AR model parameters and the variance of the error term. The model parameters are found by solving a set of linear equations

obtained by minimizing the mean squared error term (the white noise power) over all the data.

The process of the AR spectrum analysis using Marple-WPT method is described as follows. Firstly, an important consideration is the choice of the number of terms in the AR model. This is known as its order. If the order is too low the power density estimate will be excessively smoothed, so some peaks may be obscured. If the order is too high, spurious peaks may be introduced. Hence, it is important to determine the appropriate model order for each set of data.

In an AR model of a time series the current value of the series, $x(n)$, is expressed as a linear function of previous values plus an error term, $e(n)$; thus

$$\begin{aligned} x(n) = &-a(1)\,x(n-1) - a(2)\,x(n-2) - \cdots \\ &- a(k)\,x(n-k) - \cdots - a(p)\,x(n-p) + e(n). \end{aligned} \tag{1}$$

This equation incorporates p previous terms and represents a model of order p. It is more compactly written as

$$\begin{aligned} x(n) &= -\sum_{k=1}^{p} a(k)\,x(n-k) + e(n) \\ &= -\sum_{k=1}^{p} a(k)\,z^{-k}x(n) + e(n), \end{aligned} \tag{2}$$

where z^{-k} is the back-shift operator which denotes a delay of k sampling intervals. So (2) can be rewritten as

$$\begin{aligned} x(n) + \sum_{k=1}^{p} a(k)\,z^{-k}x(n) &= \left(1 + \sum_{k=1}^{p} a(k)\,z^{-k}\right)x(n) = e(n), \\ x(n) &= \frac{e(n)}{1 + \sum_{k=1}^{p} a(k)\,z^{-k}}. \end{aligned} \tag{3}$$

Then we obtain AR model

$$\frac{x(n)}{e(n)} = \frac{1}{1 + \sum_{k=1}^{p} a(k)\,z^{-k}} = H(z), \tag{4}$$

where $H(z)$ is interpretable as the z-transform of an all-pole IIR digital filter with coefficients, $a(k)$. This filter is called an AR filter. In (4) the $x(n)$ may be regarded as the output of this filter caused by random inputs, $e(n)$. $e(n)$ represents the error between the value predicted by the model, $\hat{x}(n)$, and the true datum value, $x(n)$. $e(n)$ is usually assumed to have the properties of white noise; that is, it is assumed to have a Gaussian probability density distribution and a uniform power density spectrum. Thus $x(n)$ may be regarded as having been generated by the AR filter from a white noise source.

The power spectrum density, $P_x(f)$, of the AR series $x(n)$ is as follows:

$$P_x(f) = \frac{\sigma_e^2(n)}{\left|1 + \sum_{k=1}^{p} a(k)\,e^{-jk\omega T}\right|^2}. \tag{5}$$

It can be found that the parameters in the right-hand side of (5) and the autoregressive function of $x(n)$, R_{xx}, have the following relationship [22]:

$$\begin{aligned} &\begin{pmatrix} R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(p-1) \\ R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(p-2) \\ \vdots & \vdots & \vdots & \vdots \\ R_{xx}(p-1) & R_{xx}(p-2) & \cdots & R_{xx}(0) \end{pmatrix} \begin{pmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{pmatrix} \\ &= -\begin{pmatrix} R_{xx}(1) \\ R_{xx}(2) \\ \vdots \\ R_{xx}(p) \end{pmatrix}. \end{aligned} \tag{6}$$

The model parameters, $a(k)$, may now be obtained from (6), which are the famous Yule-Walker (YW) equations. If we obtain these parameters, then we can calculate the PSD of $x(n)$.

There are several methods to solve YW equation, such as autocorrelation method (also named the Levinson-Durbin algorithm) [43], the Burg method, and the Marple method. Here we choose the Marple method because it can yield statistically stable spectral estimates of high resolution. This method minimizes the forward and backward prediction errors in the least squares sense.

In the Marple method, the YW equations (6) have the equivalent formulation

$$\begin{aligned} &\begin{pmatrix} C_{xx}(1,1) & C_{xx}(1,2) & \cdots & C_{xx}(1,P) \\ C_{xx}(2,1) & C_{xx}(2,2) & \cdots & C_{xx}(2,p) \\ \vdots & \vdots & \vdots & \vdots \\ C_{xx}(p,1) & C_{xx}(p,2) & \cdots & C_{xx}(P,P) \end{pmatrix} \begin{pmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{pmatrix} \\ &= -\begin{pmatrix} C_{xx}(1,0) \\ C_{xx}(2,0) \\ \vdots \\ C_{xx}(p,0) \end{pmatrix}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} C_{xx}(j,k) = \frac{1}{2(N-p)} \Bigg\{ &\sum_{n=p}^{N} x(n-j)\,x(n-k) \\ &+ \sum_{n=1}^{N-p} x(n+j)\,x(n+k) \Bigg\}. \end{aligned} \tag{8}$$

The $p \times p$ matrix $C_{xx}(j,k)$ is Hermitian and positive semidefinite, and (8) may be solved using the Cholensky decomposition method [44]. So we can obtain the PSD of signal $x(n)$ after we solved the YW equations. And then the spectral estimation for the AR model given from (8) at frequency $\theta = 2\pi/3$ is utilized to predict the exons in the eukaryotic DNA sequences.

Order Selection of AR Model. The order of the AR model depends on the statistical properties of the data, so it should

be carefully chosen for the data to fit well. Models of low order are preferred on the ground that fewer parameters have to be fitted. However, if the order is too small, the resulting spectral estimate will be smoothed and will have poor resolution. On the other hand, if the model order is too large, the spectral estimate may contain spurious peaks and lead to spectral line splitting. Two of the most commonly used order estimation parameters were developed by Akaike. These are the final prediction error, FPE(p) [45], given by

$$\mathrm{FPE}(p) = \frac{N+p}{N-p} E(p) \tag{9}$$

and the Akaike information criterion, AIC(p) [46], which is

$$\mathrm{AIC}(p) = N \ln(E(p)) + 2p, \tag{10}$$

where $E(p)$ is the modeling error, N is the data record length, and p is the order of the model.

Generally, AIC(p) is particularly recommended for short data records, while FPE(p) is recommended for longer data records. A practical approach is to attempt to select p to minimize both FPE(p) and AIC(p). To guarantee a valid output, Lang and McClellan [47] recommended to set the estimation order parameter to be less than or equal to two-thirds the input vector length. Zhao et al. [48] suggested that the best order could be reached between $1 \sim \sqrt{N}$. In this paper, we will make a comprehensive consideration of the aforementioned criterions and choose the best order to improve the prediction accuracy.

2.2.4. Extracting the TBP Components Using the Marple Algorithm. After the mapping and TBP enhancement steps, the next critical step of our algorithm is to extract the TBP components, which can be implemented similarly to the SDFT [14]. As for a numerical sequence $x(n), n = 1, 2, \ldots, N$, firstly, a sliding window with length M is determined; for example, $M = 351$. In the kth 351-length window, the Marple algorithm is used to calculate the PSD $P_k(f)$. Then the PSD $P_k(f)$ at frequency $\theta = 2\pi/3$; that is, the PSD at position $M/3$ [14] is extracted, referred to as $P_k(M/3)$. In order to make a fair comparison of the $\theta = 2\pi/3$ frequency spectrum in different windows, we introduce the following signal to noise ratio (SNR):

$$P_{\mathrm{SNR}}(k) = \frac{P_k(M/3)}{\overline{P_k}}, \tag{11}$$

where $\overline{P_k} = (1/(M-1)) \sum_{k=1}^{M-1} P_k(f)$, $\overline{P_k}$ is the mean of the total PSD of the kth windowed sequence.

Sliding the window along the sequence one by one position, this successive progression and the SNR curve exhibit the coding regions in DNA. It is expected that in the SNR curve, the protein coding regions have high SNR, while the noncoding regions have low SNR. So we can identify those exonic regions by proper threshold; that is, if the SNR curve of a region is above the threshold horizonal line, this region may be the exonic region while the region which is under the threshold horizonal line may be noncoding region.

There are several assistant strategies for the identification algorithm.

(1) The values on SNR curve will be normalized by dividing by their max value, which contributes to the following comparisons.

(2) Different mapping methods and the sliding window technique will make the obtained SNRs have different lengths, so we will use the mirror-symmetric boundary-extension method [49] to overcome this and make the SNRs have the same length as the numerical sequence.

(3) It should be noted that before we use SNR curve and the threshold to determine the exons, the background noise in SNR curve should be reduced. The noise reduction technique by WPT and the optimal threshold selection method are described in the following two subsections in detail.

2.2.5. Noise Reduction Using Wavelet Packets Transform. Wavelet packets transform (WPT) is a generalization of wavelet decomposition that offers a richer signal analysis. In the decomposition of a signal by using discrete wavelet transform (DWT), only the lower frequency band is decomposed, giving a right recursive binary tree structure, where its right lobe represents the lower frequency band. Its left lobe represents the higher frequency band. In the corresponding decomposition by using WPT, the lower, as well as the higher, frequency bands are decomposed giving a balanced binary tree structure [50–52]. That is, a single wavelet packet decomposition gives a lot of bases from which you can look for the best representation with respect to a design objective. This can be done by finding the "best tree" based on an entropy criterion. Such a tree is given in Figure 4 (MATLAB R2011a Wavelet Toolbox).

Denoising is an important application of WPT and its main idea is to reconstruct the useful frequency contents after the decomposition. The WPT denoising procedure of MATLAB toolbox (MATLAB R2011a Wavelet Toolbox) involves four steps.

Decomposition. For a given wavelet, compute the wavelet packet decomposition of signal x at level M.

Computation of the Best Tree. For a given entropy, compute the optimal wavelet packet tree.

Threshold of Wavelet Packet Coefficients. For each packet (except for the approximation), select a threshold and apply threshold to coefficients. The graphical tools from MATLAB toolbox automatically provide an initial threshold based on balancing the amount of compression and retained energy. This threshold is a reasonable first approximation for most cases. However, in general you will have to refine your threshold by trial and error so as to optimize the results to fit your particular analysis and design criteria. The tools facilitate experimentation with different thresholds and make it easy to alter the tradeoff between amount of compression and retained signal energy.

Reconstruction. Compute wavelet packet reconstruction based on the original approximation coefficients at level M and the modified coefficients.

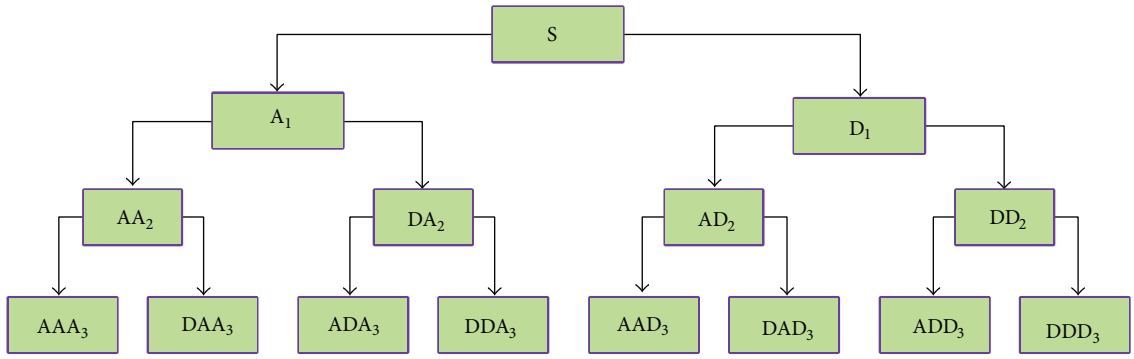

Figure 4: Wavelet packets decomposition tree at level 3.

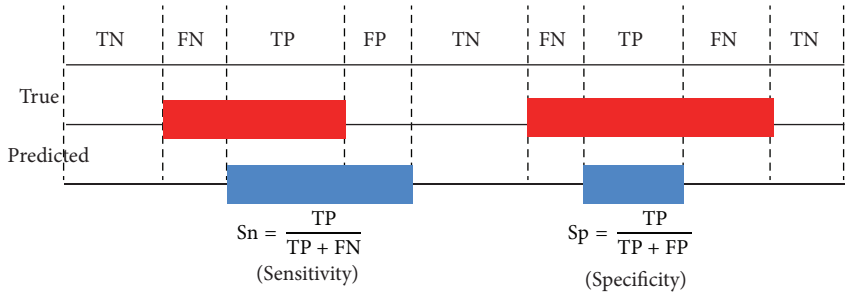

Figure 5: Measures of prediction accuracy at the nucleotide level.

2.2.6. Threshold Selection. Threshold selection plays an important role in discriminating between coding and noncoding regions based on the SNR curve. The proper threshold can help to optimize the accuracy of the identification. Xu et al. [28, 53] developed a novel method based on the bootstrap algorithm and Rao's sliding window strategy [20] to infer organism-specific optimal thresholds for different eukaryotes, and this integrate algorithm has improved the prediction accuracy than the conventional universal threshold based methods. In this paper, we use the threshold selection method developed by Kwan et al. [29, 54]. The mean and standard deviations of the TBP values determined from a training set of exon and intron sequences are used to calculate the threshold level T, which is defined as

$$T = \frac{\text{sd}\, P_{3e} * \text{mean}\, P_{3i} + \text{sd}\, P_{3i} * \text{mean}\, P_{3e}}{\text{sd}\, P_{3e} + \text{sd}\, P_{3i}}, \quad (12)$$

where mean P_{3e} and sd P_{3e} represent the mean and standard deviations of the TBP values obtained from the exon sequences of a training set, respectively, and mean P_{3i} and sd P_{3i} represent, respectively, the mean and standard deviations of the TBP values obtained from the intron sequences of the same training set.

2.2.7. Evaluation Criteria at Nucleotide Level. In these evaluations, results of different methods are compared at the nucleotide level. At this level, we evaluate the accuracy of a prediction on a test sequence by comparing the predicted coding value (coding or noncoding) with the true coding value for each nucleotide along the test sequence [31]. For this purpose, the following measures are employed [55].

Sensitivity, Specificity, and AC. Sensitivity and specificity are probably the most widely used measures for prediction accuracy evaluation. Similar to [31], a figure of the two measures is utilized to explain them (Figure 5). In Figure 5 true positive (TP) is the number of coding nucleotides correctly predicted as coding, false negative (FN) is the number of coding nucleotides predicted as noncoding, true negative (TN) is the number of noncoding nucleotides correctly predicted as noncoding, and false positive (FP) is the number of noncoding nucleotides predicted as coding. Based on the aforementioned four quantities, sensitivity (Sn) and specificity (Sp) are defined as

$$Sn = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad Sp = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (13)$$

That is, Sn gives the measure of the proportion of coding nucleotides that have been correctly predicted as coding, and Sp is the proportion of coding nucleotides that are actually coding.

Neither Sn nor Sp is sufficient by itself because perfect sensitivity of 1 can be obtained if all the nucleotides were predicted as coding, and perfect specificity can be obtained if

all nucleotides were predicted as noncoding [30]. So accuracy defined as

$$\text{accuracy} = \frac{(Sn + Sp)}{2}, \tag{14}$$

is a widely used compound measure which considers both sides of Sn and Sp from the global perspective [28].

Here we introduce several other global measures as the previous researchers [30, 31]. Correlation coefficient (CC) is a single scalar value summarizing both Sn and Sp as a measure of global accuracy, with the definition

$$\begin{aligned} \text{CC} = {} & (\text{TP} \times \text{TN} - \text{FN} \times \text{FP}) \\ & \times ((\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \\ & \times (\text{TN} + \text{FN}))^{-1/2}. \end{aligned} \tag{15}$$

CC appears to be particularly appropriate as a measure of overall prediction accuracy, but CC has many shorcomings [31], as an ideal alternative measure of CC. Approximate correlation (AC) was firstly proposed by Burset and Guigó [31]. Its definition is as follows:

$$\text{AC} = (\text{ACP} - 0.5) * 2, \tag{16}$$

where average conditional probability (ACP) is

$$\begin{aligned} \text{ACP} = {} & \frac{1}{4}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TP}}{\text{TP} + \text{FP}} + \frac{\text{TN}}{\text{TN} + \text{FN}}\right. \\ & \left. + \frac{\text{TN}}{\text{TN} + \text{FP}}\right). \end{aligned} \tag{17}$$

According to Burset and Guigó [31], AC can be used to measure the association between prediction and reality appropriately. AC is not only a measure of gene structure prediction accuracy, but also a measure to optimize when developing gene structure prediction programs. Unlike the CC, it has a probabilistic interpretation, and it can be computed in any circumstances. As ACP is the average of four conditional probabilities [31], it ranges from 0 to 1, so AC ranges from −1 to 1, which can be compared to CC. So AC can be looked upon as approximate measure of CC. And according to [31], if an algorithm has the larger AC value (strictly speaking, the absolute value of AC), that means this algorithm has the better accuracy. So we will calculate the three measures S_n, S_p, and AC in the algorithm evaluation.

Receiver Operating Characteristics (ROC) Curves. The ROC curves were developed in the 1950s as a technique for visualizing, organizing, and selecting classifiers based on their performance [55, 56]. In the exon identification problems, an ROC curve can help to explore the effects on TP and FP as the position of an arbitrary decision threshold is varied. Also, the curve can be characterized as a single number using the area under the ROC curve (AUC). The larger AUC leads to the better performance of the tested technique.

3. Results and Discussion

In this section, the results of the proposed algorithm are compared with those of existing techniques, such as sliding Fourier transform spectrum (SDFT) (referred to as VossDFT) [14], and those improved techniques are based on DFT, such as EIIPDFT [38], PNDFT [39], and SPDFT [41, 42].

The outline of this section is as follows. Firstly, the denoising performance of WPT technique is given by comparing the SNR of a short benchmark data. Then we compare our Code13 mapping method with four widely used mapping methods selected from the aforementioned mapping approaches, that is, Voss, EIIP, SP, and PN mapping methods. It should be noted that in the comparison only the mapping method is different; that means in Figure 3 procedures only the first step is different, the following steps are totally identical. Finally, we compare our proposed algorithm with other existing techniques, and three widely used benchmark datasets will be utilized for comparison purpose. To evaluate and compare the results, the aforementioned measures such as Sn, Sp, AC, ROC, and AUC are calculated.

Firstly, we use the DNA sequence F56F11.4 to test the noise reduction performance of WPT. The SNR curve of sequence F56F11.4 calculated by our algorithm is shown in Figure 6. During the calculation process, we determine the window length as 351 [14]; the order of the AR model is 9 according to the order selection strategy (see Section 2.2.3). The *soft threshold* function is utilized to process the data, entropy is sure criterion, and symlets wavelet is selected as the orthogonal wavelet which will be decomposed into 5 levels according to the signal. It can be seen that the burrs shape noise in original SNR (Figure 6(a)) is distinctly reduced by WPT technique, and the "smoothed" SNR (Figure 6(b)) will contribute to the following exonic identification accuracy.

Secondly, the Code13 and the other four aforementioned mapping methods (Voss, EIIP, SP, and PN) are, respectively, utilized to map the F56F11.4 sequence into five different numerical sequences. Then according to the procedure of our proposed algorithm (Figure 3), the following techniques of the integrated algorithm, including the Marple algorithm and WPT, are used to identify the exons with the same parameters settings in the whole calculate process. So it can be looked upon as five different algorithms whose difference only lies in the mapping methods, and we compare their final identification performance. Similar to the aforementioned WPT denoising section, we determine the following settings: the window length is 351, the order of the AR model is 9, and the soft threshold function, sure entropy criterion, and symlets wavelet with 5-level decomposition are utilized to reduce the noise of the SNR.

The performance measures of five mapping methods for gene sequence F56F11.4 are represented in Table 1. Here, the sensitivities, specificities, and approximate correlation are calculated under the use of optimal threshold according to (14), and the results show that the Code13 method has the largest measures ($Sn = 0.8262, Sp = 0.3011$, and AC = 0.4233). Figure 7 shows the exonic regions identification performance of sequence F56F11.4 using five mapping methods combing with the Marple-WPT technique. As can be seen,

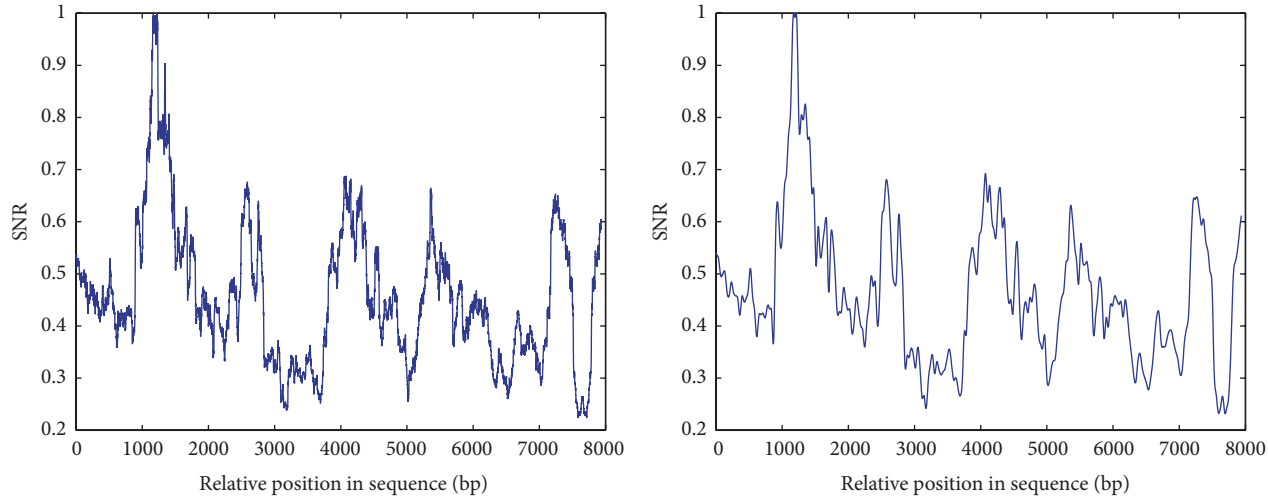

Figure 6: Denoising performance results of WPT for DNA sequence F56F11.4. (a) The output SNR of sequence F56F11.4 based on the proposed algorithm before WPT is utilized. (b) The output SNR based on the proposed algorithm using WPT. Both methods employ the Code13 mapping method and all the other procedures are identical except whether denoising or not.

Table 1: Performance measures of five mapping methods for sequence F56F11.4 based on Marple-WPT technique.

Mapping method	Sn	Sp	AC	Optimal threshold
Voss	0.7810	0.2270	0.3087	0.6338
EIIP	0.4661	0.1402	0.0656	0.3862
SP	0.5079	0.1095	−0.0235	0.6656
PN	0.6411	0.1539	0.1265	0.6690
Code13	**0.8262**	**0.3011**	**0.4233**	0.5062

there are five exonic regions with the relative positions 928–1039, 2528–2857, 4114–4377, 5465–5644, and 7255–7605 in sequence F56F11.4 to be identified (red bold line segments). The Code13 mapping based algorithm (Figure 7(e)) is able to produce distinct SNR peaks for all the exonic regions, and the SNR in noncoding regions is restricted well. So it plays a critical role in the following identification, and the optimal threshold helps to determine the exact start and end position of the predicted exonic regions; the black thin line segments represent the predicted candidate exons. It is worth mentioning that the first exon of sequence F56F11.4 is a typical short sequence, whose length is 112 bp, and it is hard to be identified in many conventional techniques [2]. But in our Code13 based algorithm, it is easy to be identified. However, the first four mapping methods produce poor performance, either miss the true exon (those square shadows in Figures 7(a), 7(b), 7(c), and 7(d)) or give a false forecast (rectangle shadows in Figures 7(c) and 7(d)). So it means that the Voss, EIIP, SP, and PN methods based algorithms cannot distinguish the exonic regions from the noncoding regions precisely. The advantage of Code13 mapping method may lie in the fact that it is a kind of complex number representation, and it reflects the complementary nature of nucleotide *C*-*G* and *A*-*T* pairs relationship [33].

Finally, our proposed algorithm is applied to the three widely used benchmark datasets: GENSCAN65, HMR195, and BG570. For comparison purpose, several conventional exonic identification techniques are employed on the aforementioned datasets in the mean time, and the performance criteria measures such as AC, ROC curves, and AUC are utilized in the comparison process.

Taking HMR195 as an example, this benchmark datasets contains 195 sequences with exactly one complete either single-exon or multiexon gene (including 43 single-exon genes and 152 multiexon genes) [30]. HMR195 has the following characteristics: the ratio of human : mouse : rat sequences is 103 : 82 : 10; the mean length of the sequences in the set is 7096 bp; there are 948 exons in the datasets with the total length 199176 bp; the minimum exon length is 12 bp; the mean length of the exons is 208 bp; and the mean intron length is 678 bp. Figure 8 shows the distribution of exonic regions length in HMR195, and most exonic regions length concentrate about 210 bp.

Five identification techniques are utilized for the aforementioned three datasets, that is, VossDFT [14], EIIPDFT [38], SPDFT [41, 42], PNDFT [39], and our proposed Code13-Marple. The output results are represented in Table 2 and Figure 9. As can be seen, the proposed Code13-Marple method achieves the largest AC values for all the three datasets (that is, 0.2324, 0.2508, and 0.2131), which means that this method outperforms the other four traditional DFT-based methods in general. From Table 2 and Figure 9, it also can be found that the Code13-Marple algorithm has the largest AUC values for the three test datasets, that is, 0.6801, 0.7179, and 0.6522. Also taking HMR195 datasets as an example, our proposed algorithm achieves relative improvements of 27.6%, 28.7%, 16.9%, and 16.8% over the VossDFT, EIIPDFT, SPDFT, and PNDFT techniques, respectively, in terms of the AUC values. That is, our proposed algorithm

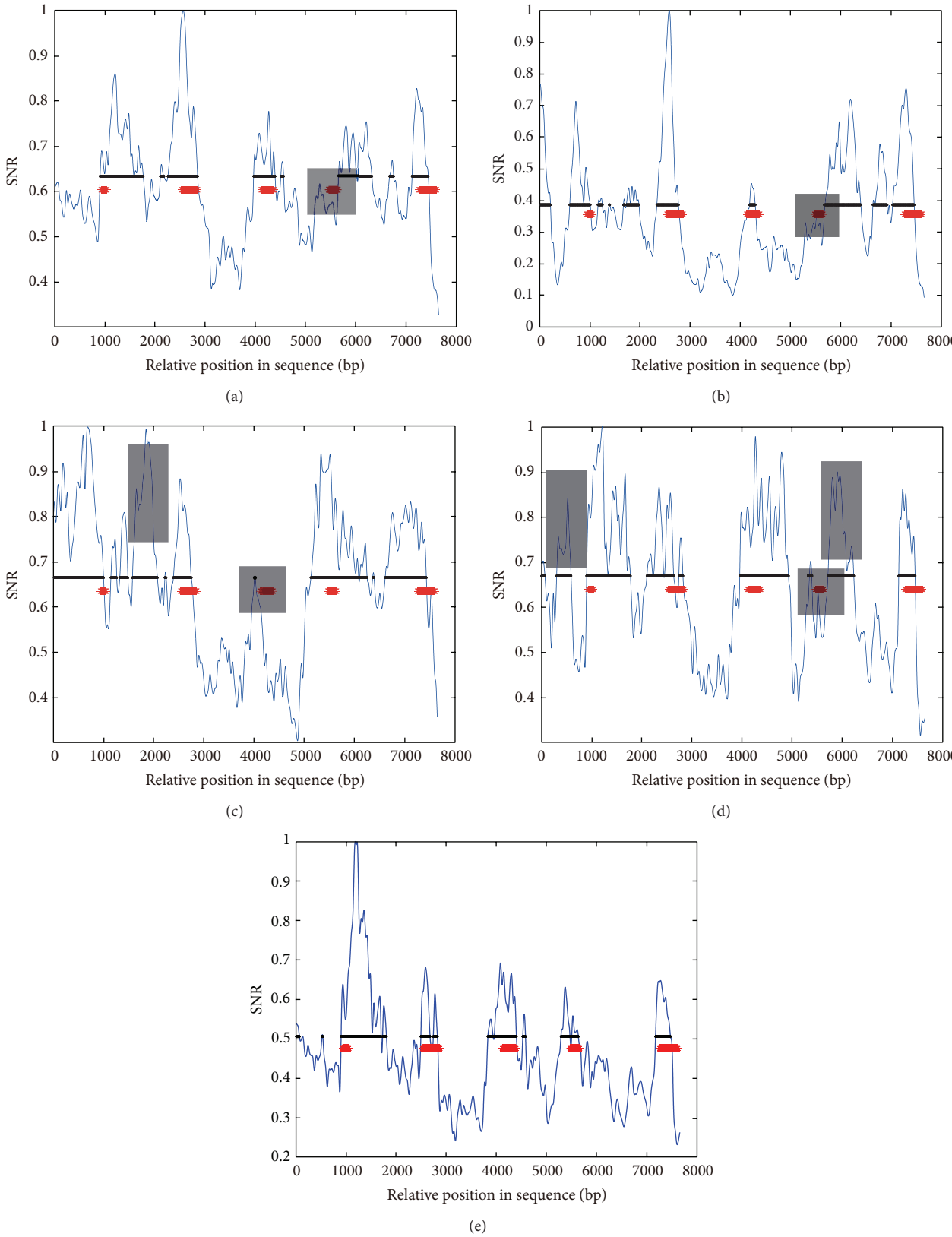

Figure 7: Exonic identification results of gene sequence F56F11.4 using five mapping methods and the proposed algorithm. (a) Voss method, (b) EIIP method, (c) SP method, (d) PN method, and (e) Code13 method. The red bold line segments represent the true exons that must be identified, the black thin line segments represent the predicted candidate exons, and the vertical heights of those line segments represent their optimal thresholds. The square shadow represents the missing true exon; the rectangle shadow represents the false predicted exon.

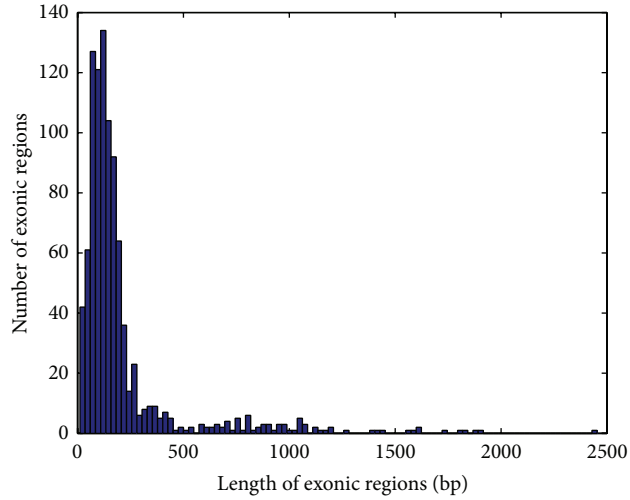

Figure 8: The distribution of exonic regions lengths in HRM195. The horizonal axis represents the exonic regions length, and the vertical axis represents the number of exonic regions.

Table 2: Performance measures of five mapping methods for three benchmark datasets.

Datasets	Measure	VossDFT	EIIPDFT	SPDFT	PNDFT	Code13-Marple
GENSCAN65	AC	0.1533	0.1162	0.1850	0.2010	**0.2324**
	AUC	0.6385	0.5754	0.6372	0.6283	**0.6801**
HMR195	AC	0.1045	0.0572	0.1300	0.1434	**0.2508**
	AUC	0.5626	0.5113	0.5962	0.5971	**0.7179**
BG570	AC	0.1093	0.0599	0.1263	0.1356	**0.2131**
	AUC	0.5329	0.4867	0.5470	0.5391	**0.6522**

achieves more accuracy than the other four methods. The ROC curves of the proposed algorithm in Figure 9 are all distinctly "close" to the top left corner which visually verifies that the Code13-Marple-WPT method is more effective than the other techniques. The reason of the aforementioned output may lie in the fact that those conventional nonparametric methods especially those DFT-based techniques have the advantage of possible implementation using the FFT, but with the disadvantage in the case of short data lengths of limited frequency resolution, and the requirement for windowing to reduce the spectral leakage. Parametric methods on the other hand can provide high resolution, applicability to short data lengths, and avoidance of spectral leakage, scalloping loss, spectral smearing, and window biasing effects [22].

4. Conclusions

In this paper, we propose a new technique based on Marple algorithm and wavelet packets transform with the Code13 numerical mapping approach to improve the accuracy of identification of the protein coding regions in the eukaryotic DNA sequences. The outputs of the test by many benchmark datasets show that the proposed algorithm outperforms some well-known DFT-based methods. There are several reasons attributed to the improvement of the identification accuracy: first, the FIR filters help to enhance the TBP characteristics of the numerical sequences before PSD calculation; second, the Marple algorithm can calculate the PSD more efficiently and accurately than those conventional methods because it can yield statistically stable spectral estimates of high resolution; third, the WPT can reduce the noise in SNR curves, which attributes to the following identification of exonic regions distinctly; finally, those assistant strategies such as threshold selection, normalization of SNR curves, the mirror-symmetric boundary-extension method also can help to improve the final accuracy of the whole algorithm.

In the same time, it should be noted that there are still some shortcomings in our proposed algorithm, such as the order selection of the AR model when using Marple algorithm and the Marple algorithm being a little more time-consuming.

Also there are still two important and challengeable problems which deserve further study. First, how can we obtain the precise exons location information [14, 20]? Second, how can we utilize the algorithm to identify the dual coding genes in eukaryote? The dual coding genes are the phenomenon where there are two overlapped open reading frames (ORFs) in the same direction of a protein coding region (such as three known humans GNAS1, XBP1, and INK4a) [57, 58]. And whether or not we can take the original overlapped sequence as two equal length sequences with partial overlapped signal, and convert it into a blind signal separation problem, is

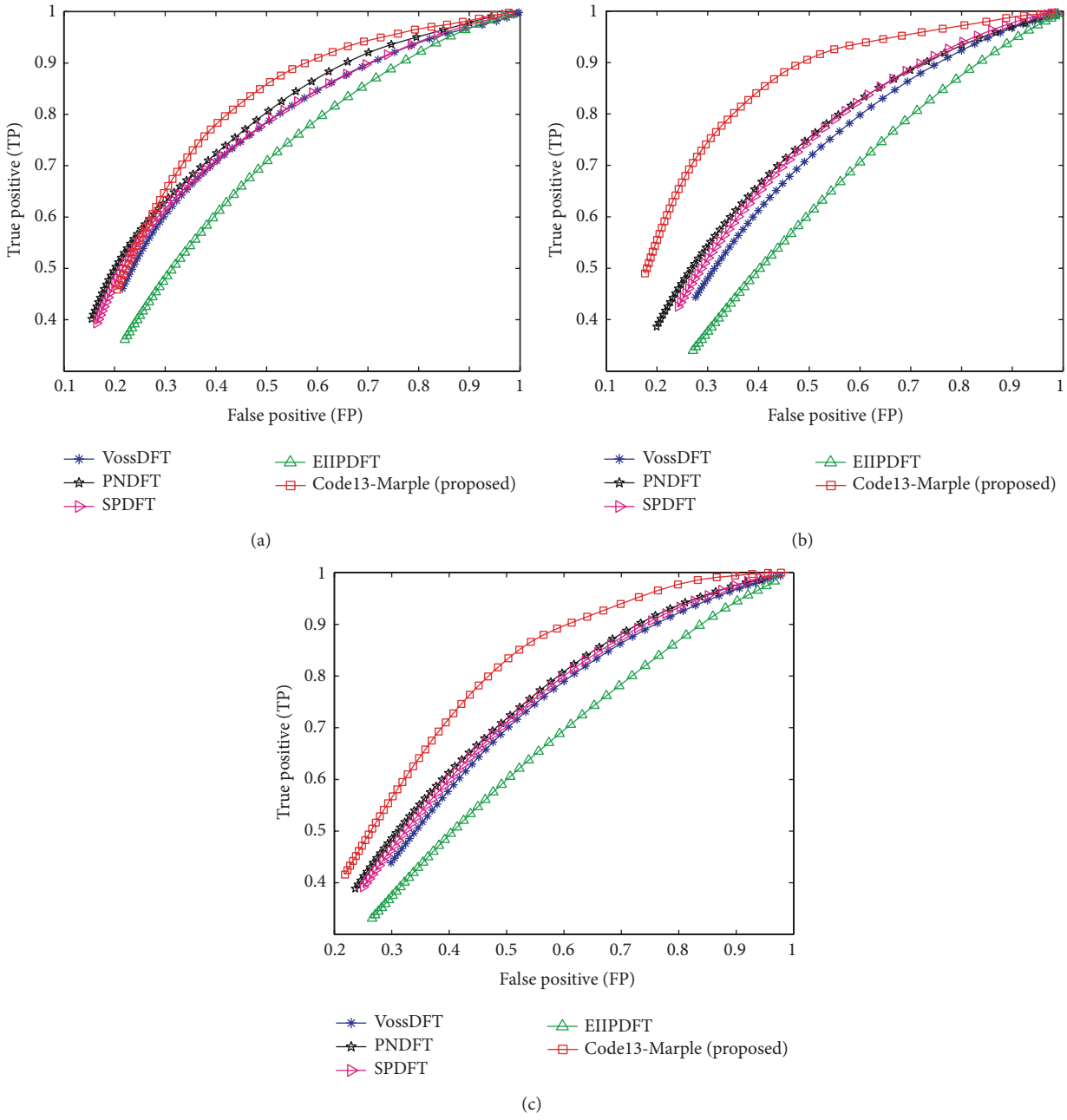

FIGURE 9: ROC curves of different techniques for three benchmark datasets. (a) The ROC curves of five methods (VossDFT, EIIPDFT, SPDFT, PNDFT, and Code13-Marple) for GENSCAN65 datasets. (b) The ROC curves of five methods for HMR195 datasets. (c) The ROC curves of five methods for BG570 datasets.

unknown and deserves further study in the future. So it is our next target to overcome the aforementioned tasks for more effective and accuracy algorithm and to help the related biologists to identify the DNA structure more clearly.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank Professor Fengzhu Sun from the University of Southern California deeply for the interest in the project and useful discussion about the coding region discriminating criteria. The authors also thank all the anonymous reviewers for their valuable suggestions and support. This work is supported by the Natural Science Foundation of China Grants 11371227 and 10921101 and Graduate Independent Innovation Foundation of Shandong University (GIIFSDU) (yzc12098).

References

[1] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12086–12094, 2009.

[2] S. A. Marhon and S. C. Kremer, "Gene prediction based on DNA spectral analysis: a literature review," *Journal of Computational Biology*, vol. 18, no. 4, pp. 639–676, 2011.

[3] C. Mathé, M. Sagot, T. Schiex, and P. Rouzé, "Current methods of gene prediction, their strengths and weaknesses," *Nucleic Acids Research*, vol. 30, no. 19, pp. 4103–4117, 2002.

[4] N. Y. Song and H. Yan, "Short exon detection in DNA sequences based on multifeature spectral analysis," *Eurasip Journal on Advances in Signal Processing*, vol. 2011, Article ID 780794, 8 pages, 2011.

[5] S. Maji and D. Garg, "Progress in gene prediction: principles and challenges," *Current Bioinformatics*, vol. 8, no. 2, pp. 226–243, 2013.

[6] N. Goel, S. Singh, and T. C. Aseri, "A review of soft coputing techniques for gene prediction," *ISRN Genomics*, vol. 2013, Article ID 191206, 8 pages, 2013.

[7] H. Saberkari, M. Shamsi, H. Heravi, and M. H. Sedaaghi, "A novel fast algorithm for exon prediction in eukaryotic genes using linear predictive coding model and Goertzlel algorithm based on the Z-curve," *Journal of Medical Signals and Sensors*, vol. 3, pp. 139–149, 2013.

[8] J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar Jr., "Identification of protein coding regions using the modified Gabor-wavelet transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 198–206, 2008.

[9] R. Guigo, "DNA compositon, codon usage and exon prediction," in *Genetic Databases*, Academic Press, 1999.

[10] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in DNA with a hidden Markov model," *Journal of Computational Biology*, vol. 4, no. 2, pp. 127–141, 1997.

[11] S. Agoes, "A Hidden Markov Model for identificatino of exons in DNA of genes Plasmodium falciparum," *International Journal of Electrical & Computer Sciences*, vol. 11, pp. 33–36, 2011.

[12] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, 2001.

[13] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, no. 17, pp. 5303–5318, 1982.

[14] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Computer Applications in the Biosciences*, vol. 13, no. 3, pp. 263–270, 1997.

[15] R. Nini and S. J. Shepherd, "Detection of 3-periodicity for small genomic sequences based on AR technique," in *Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS'04)*, vol. 2, pp. 1032–1036, June 2004.

[16] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, "Periodicity in DNA coding sequences: implications in gene evolution," *Journal of Theoretical Biology*, vol. 151, no. 3, pp. 323–331, 1991.

[17] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992.

[18] C. A. Chatzidimitriou-Dreismann and D. Larhammar, "Long-range correlations in DNA," *Nature*, vol. 361, no. 6409, pp. 212–213, 1993.

[19] H. Saberkari, M. Shamsi, M. Sedaaghi, and F. Golabi, "Prediction of protein coding regions in DNA sequences using signal processing methods," in *Proceedings of the IEEE Symposium on Industrial Electronics and Applications (ISIEA '12)*, pp. 355–360, Bandung, Indonesia, September 2012.

[20] N. Rao, X. Lei, J. Guo, H. Huang, and Z. Ren, "An efficient sliding window strategy for accurate location of eukaryotic protein coding regions," *Computers in Biology and Medicine*, vol. 39, no. 4, pp. 392–395, 2009.

[21] P. P. Vaidyanathan and B. J. Yoon, "Gene and exon prediction using allpass-based filters," in *Proceedings of the IEEE International Workshop on Genoimc Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, USA, 2002.

[22] E. Ifeachor and B. Jervis, *Digital Signal Processing: A Practical Approach*, Prentice-Hall, 2nd edition, 2002.

[23] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," *Eurasip Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 13–28, 2004.

[24] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," in *Proceeding of the IEEE 2005 International Conference on Emerging Technologies (ICET '05)*, pp. 13–17, September 2005.

[25] M. Akhatar, "Comparison of gene and exon prediction techniques for detection of short coding regions," *International Journal of Information Technology*, vol. 11, pp. 26–35, 2005.

[26] H. Yan and T. D. Pham, "Spectral estimation techniques for DNA sequence and microarray data analysis," *Current Bioinformatics*, vol. 2, no. 2, pp. 145–156, 2007.

[27] M. K. Choogn and H. Yan, "Multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves," *Bioinformation*, vol. 2, pp. 273–278, 2008.

[28] S. Xu, N. Rao, X. Chen, and B. Zhou, "Inferring an organism-specific optimal threshold for predicting protein coding regions in eukaryotes based on a bootstrapping algorithm," *Biotechnology Letters*, vol. 33, no. 5, pp. 889–896, 2011.

[29] O. Abbasi, A. Rostami, and G. Karimian, "Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform," *BMC Bioinformatics*, vol. 12, article 430, 2011.

[30] S. Rogic, A. K. Mackworth, and F. B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome Research*, vol. 11, pp. 817–832, 2001.

[31] M. Burset and R. Guigó, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, no. 3, pp. 353–367, 1996.

[32] C. Burge, *Identification of genes in human genomic DNA, [Ph.D. distertation]*, Stanford University, Stanford, Calif, USA, 1997.

[33] H. K. Kwan, B. Y. M. Kwan, and J. Y. Y. Kwan, "Novel methodologies for spectral classification of exon and intron sequences," *Eurasip Journal on Advances in Signal Processing*, vol. 2012, no. 1, article 50, 2012.

[34] M. Abo-Zahhad, M. A. Ahmed, and S. A. Abd-Elrahamn, "Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques," *International Journal of Information Technology and Computer Science*, vol. 4, no. 8, pp. 22–36, 2012.

[35] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *Journal of Theoretical Biology*, vol. 118, no. 3, pp. 295–300, 1986.

[36] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[37] P. D. Cristea, "Genetic signal representation and analysis," in *International Conference on Biomedical Optics*, vol. 4623 of *Proceedings of SPIE*, pp. 77–84, 2002.

[38] S. N. Achuthsanar and S. S. Pillai, "A coding measure scheme empoying electron-ion interaction pseudo potential (EIIP)," *Bioinformatics*, vol. 1, pp. 197–202, 2006.

[39] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in *Proceedings of the 5th IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '07)*, Tuusula, Finland, June 2007.

[40] R. Zhang and C. Zhang, "Identification of replication origins in archaeal genomes based on the *Z*-curve method," *Archaea*, vol. 1, no. 5, pp. 335–346, 2005.

[41] K. Florquin, Y. Saeys, S. Degroeve, P. Rouzé, and Y. van de Peer, "Large-scale structural analysis of the core promoter in mammalian and plant genomes," *Nucleic Acids Research*, vol. 33, no. 13, pp. 4255–4264, 2005.

[42] W. F. Zhang and H. Yan, "Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences," *Pattern Recognition*, vol. 45, no. 3, pp. 947–955, 2012.

[43] S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[44] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1974.

[45] H. Akaike, "Fitting autoregressive models for prediction," *Annals of the Institute of Statistical Mathematics*, vol. 21, pp. 243–247, 1969.

[46] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[47] S. W. Lang and J. H. McClellan, "Frequency estimation with maximum entropy spectral estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 716–724, 1980.

[48] L. C. Zhao, J. J. Ma, S. Q. Fan, and Z. Z. Si, "Research on AR model in vibration analysis of rolling bearings," *Chinese Mechanical Engineering*, vol. 15, no. 3, pp. 210–213, 2004.

[49] L. T. Guan, "Wavelet interpolation and decomposition in a finite interval with boundary conditions," *Chinese Journal of Engineering Mathematics*, vol. 12, no. 3, pp. 1–9, 1995.

[50] X. Wang, C. Liu, F. Bi, X. Bi, and K. Shao, "Fault diagnosis of diesel engine based on adaptive wavelet packets and EEMD-fractal dimension," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 581–597, 2013.

[51] C. M. Vong and P. K. Wong, "Engine ignition signal diagnosis with Wavelet Packet Transform and Multi-class Least Squares Support Vector Machines," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8563–8570, 2011.

[52] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, 3rd edition, 2009.

[53] S. Xu, *Research on the thresholds selection based on the bootstrap algorithm in gene-prediction [M.S. thesis]*, University of Electronic Science and Technology of China, Chengdu, China, 2008.

[54] J. Y. Y. Kwan, B. Y. M. Kwan, and H. K. Kwan, "Spectral analysis of numerical exon and intron sequences," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '10)*, pp. 876–877, Hongkong, China, December 2010.

[55] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal processing in sequence analysis: advances in eukaryotic gene prediction," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 310–321, 2008.

[56] T. Fawcett, *ROC Graphs: Notes and Practical Considerations for Researchers*, HP Laboratories, Palo Alto, Calif, USA, 2003, http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf.

[57] Z.-F. Li, C.-G. Zhang, Z.-Y. Shen, and X.-Y. Hang, "Dual coding genes in eukaryote," *Progress in Biochemistry and Biophysics*, vol. 36, no. 5, pp. 536–540, 2009.

[58] W. Y. Chung, S. Wadhawan, R. Szklarczyk, S. K. Pond, and A. Nekrutenko, "A first look at ARFome: dual-coding genes in mammalian genomes," *PLoS Computational Biology*, vol. 3, article e91, no. 5, 2007.