EURASIP Journal on Applied Signal Processing 2004:1, 13–28

Autoregressive Modeling and Feature Analysis of DNA Sequences

Niranjan Chakravarthy
Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, USA
Email: niranjan.chakravarthy@asu.edu

A. Spanias
Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, USA
Email: spanias@asu.edu

L. D. Iasemidis
Harrington Department of Bioengineering, Arizona State University, Tempe, AZ 85287-9709, USA
Email: leon.iasemidis@asu.edu

K. Tsakalis
Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, USA
Email: tsakalis@asu.edu

Received 28 February 2003; Revised 15 September 2003

A parametric signal processing approach for DNA sequence analysis based on autoregressive (AR) modeling is presented. AR model residual errors and AR model parameters are used as features. The AR residual error analysis indicates a high specificity of coding DNA sequences, while AR feature-based analysis helps distinguish between coding and noncoding DNA sequences. An AR model-based string searching algorithm is also proposed. The effect of several types of numerical mapping rules in the proposed method is demonstrated.

Keywords and phrases: DNA, autoregressive modeling, feature analysis.

1. INTRODUCTION

The complete understanding of cell functionalities depends primarily on the various cell activities carried out by proteins. Information for the formation and activity of these proteins is coded in the deoxyribonucleic acid (DNA) sequences. For detection purposes, the vast amount of genomic data makes it necessary to define models for DNA segments such as the protein coding regions. Such models can also facilitate our understanding of the stored information and could provide a basis for the functional analysis of the DNA. Since the DNA is a discrete sequence, it can be interpreted as a discrete categorical or symbolic sequence and hence, digital signal processing (DSP) techniques could be used for DNA sequence analysis. The DNA sequence analysis problem can be considered as analogous to some forms of speech recognition problems. That is, coding and noncoding regions in DNA need to be identified from long nucleotide sequences, a process that bears some similarities to the problem of identifying phonemes from long sequences of speech signal samples. Currently proposed DSP techniques include the study of the spectral characteristics [1, 2, 3, 4] and the correlation structure [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18] of DNA sequences. The measurement of spectra in most cases has been characterized by nonparametric Fourier transform techniques [1]. In some of the most common cases, the presence of a spectral peak [1] was used to characterize protein-coding regions in the DNA. On the other hand, correlations have been often characterized on the basis of the extent of power-law (long-range) behavior and the persistence of the power-law correlation sequence [6, 8]. Attempts have been also made to parameterize these correlations in terms of the scale of the power law [6].

In this paper, we propose the use of parametric spectral methods for the analysis of DNA sequences. Parametric spectral analysis techniques have been widely used to study time series of speech, seismic, and other types of signals. Specifically, we investigate the use of autoregressive (AR) spectral

estimation tools for DNA sequence analysis. AR models effectively capture spectral peaks and model the correlation in sequences [19]. After the model fit, the AR model parameters, and AR related signals such as the prediction residual, can be used as features of the DNA sequences. The studies that we carried on AR models include the following. First, we explored the use of linear prediction residuals to compare coding and noncoding regions as well as distinguish between different genes. Different numerical mapping rules for the representation of nucleotides were considered. Second, we used the AR parameters as DNA sequence features.

The paper is organized as follows. A few basic biological properties of the DNA are described in Section 2. An overview of DNA sequence analysis techniques based on correlation functions and DSP-based methods is presented in Section 3. The motivation for the use of parametric spectral analysis methods for DNA analysis and its various implementation aspects are presented in Section 4. Results from the application of AR model-based analysis to DNA sequences are presented in Section 5. A discussion of the results and possible extensions to these techniques are given in Section 6.

2. DNA STRUCTURE AND FUNCTION

DNA is the basic information storehouse in living cells. Various cell activities are carried out by proteins which are produced based on information stored in genes. DNA is a polymer formed from 4 basic subunits or nucleotides, namely, adenine (A), cytosine (C), thymine (T), and guanine (G). A single DNA strand is formed by the covalent bonds between the sugar phosphate groups of the nucleotides. Two DNA strands are then weakly bonded by hydrogen bonds between the nucleotides. Since the nucleotide A forms such a bond only with T, and G only with C, the two DNA strands are complementary to each other and each of them is used as a template during cell division to transfer information. Usually, two complementary DNA strands form a double helix. The synthesis of proteins is governed by certain regions in the DNA called protein *coding regions* or genes. The 64 possible nucleotide triplets ((nucleotide alphabet size)$^{\text{word length}} = 4^3$), called *codons*, are mapped into 20 amino acids that bond together to form proteins. Certain codons known as start and stop codons indicate the beginning and end of a gene. The DNA also consists of regions that store information for regulatory functions. In advanced organisms, the protein coding regions are not generally continuous and are separated into several smaller subregions called exons. The regions between the exons are known as introns. During the protein coding process, these introns are eliminated and the exons are spliced together. The splicing can be carried out in a number of different ways depending on the cell function. Splicing thus also determines the type of protein synthesis and hence genes can be used for the production of a variety of proteins. The central dogma (Figure 1) in cellular biology describes the information transfer from the DNA to the ribonucleic acid (RNA) and the production of proteins. The formation of proteins takes place in two stages, namely, transcription and translation. During transcription, the genes in the DNA sequence are used as templates to form the pre-messenger RNA (pre-mRNA). The pre-mRNA is a polymer formed from 4 basic subunits, namely, A, C, G, and uracil (U). Next, the exons in the pre-mRNA are spliced together to form a polymer of only coding regions known as the mRNA. The mRNA along with the transfer RNA (tRNA) controls protein formation. The complete process is controlled and catalyzed by a number of enzymes. Almost all cells in a living system have the same DNA structure and information content. The gene expression depends on the cell requirements. Microarray technology basically captures the amount of expression of various genes. The structure and organization of the DNA and various cell functions are explained in [20].

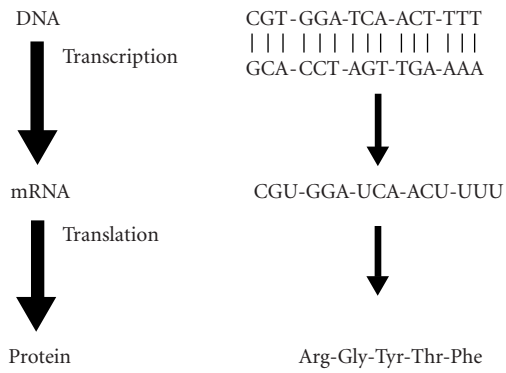

Figure 1: Central dogma; the information transfer from DNA to proteins.

One of the relevant problems in bioinformatics is to accurately identify the protein coding regions and thus predict the protein that will be generated using the information in these segments. In addition, some effort is expended in understanding the role of noncoding regions. It is therefore of central interest to analyze and characterize various DNA regions such as coding and noncoding sequences.

3. REVIEW OF METHODS FOR DNA SEQUENCE ANALYSIS

A primary objective of DNA sequence analysis is to automatically interpret DNA sequences and provide the location and function of protein coding regions. Methods to locate genes, and various coding measures are described in [21]. The gene identification problem is challenging especially in eukaryotic DNA sequences in which the coding regions are separated into several exons. An overview of standard techniques for gene identification is provided in [22]. Computational techniques for gene identification are classified into template methods and lookup methods. Template methods attempt to model prototype objects or sequences and identify genes based on these models. On the other hand, lookup methods use exactly known gene sequences and search for similar segments in a database. Computational techniques, to accomplish the above, include identification measures like Fourier spectra and sequence similarity measures. An overview of the

standard coding measures and their accuracy in identifying genes is also given in [22]. A discussion on the regulation of gene expression, techniques to integrate various gene models, for example, hidden Markov models (HMM), and methods for efficient computation are presented in [22] as well.

3.1. Correlations in DNA sequences

Correlation functions have been widely used to study the statistical properties of DNA sequences. The autocorrelation of a stationary and ergodic numerical sequence x at lag m is defined as

$$\begin{aligned} r_{xx}(m) &= E[x(n+m)x(n)] \\ &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n+m)x(n), \end{aligned} \tag{1}$$

where $E[\cdot]$ is the statistical expectation operator and N is the length of the window over which the averaging is performed. A typical statistically well-behaved estimator for the autocorrelation is

$$\hat{r}_b(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n+|m|)x(n). \tag{2}$$

The power spectrum of a signal is the Fourier transform of its correlation [19]. To use (2) in DNA analysis, one has to assign numerical values to the nucleotides A, T, C, and G. One of the early analyses of the correlation structure in the DNA was done in [6]. Binary indicator sequences are used therein to calculate correlations in the DNA sequence. The power spectra of the sequences are shown to have a power-law behavior. The spectra are reported to change according to the evolutionary categories of the DNA sequences analyzed. Similar analysis is also presented in [11], wherein a simple model, called expansion-modification model, is considered to exhibit correlations similar to those present in the DNA. Results are therein presented based on three correlation measures, that is, the mutual information function, the power spectrum to calculate the correlations, and a cumulative approach (similar to a DNA walk). Various issues of the DNA correlation structure and its interpretation are also discussed.

The calculation and relation between correlation functions and mutual information of symbol sequences are explained in [5]. Correlation functions and mutual information function differ in quantifying statistical dependencies. While correlations measure only the linear dependencies in sequences, the mutual information function detects other statistical dependencies (e.g., nonlinear) in the signal as well. The correlation measurements depend on the assignment of numbers to the symbols in the sequence, whereas the mutual information is independent of such coordinate transformations. The binary mapping rules used in [7] carry certain biological interpretations and are used in the calculation of the autocorrelation and the other related statistical dependencies. A study on the statistical correlations in the DNA sequence is presented in [8], in which possible errors in estimating correlations from short DNA sequences is also described. The direct measure of correlations from long sequences is advocated to be better than measures obtained through detrended fluctuation analysis (DFA) [10], indirect autocorrelation computation from the power spectra, and correlation estimates from the mutual information function [11]. The DFA technique removes heterogeneities in the DNA sequence, but since it has been reported that important details of the correlation structure in the DNA may be due to these heterogeneities [23], the use of the DFA technique is questioned. The autocorrelation function is considered to be useful in measuring the compositional heterogeneity. A series of studies on the use of correlation in DNA analysis is also given in [9, 14, 15, 16, 17, 18]. Other methods for DNA analysis include DNA walk [24] and Markov chains of various orders.

Observed correlation properties have also been interpreted in terms of the underlying biology [11, 12, 13, 18]. One of the important characteristics of protein coding segments in DNA sequences is the presence of persistent correlations with a pronounced period of three. It is shown in [12] that these correlations arise due to the nonuniform usage of codons in the coding regions. This nonuniformity is considered to exist due to a number of factors including the many-to-one mapping of codons to amino acids, the use of certain amino acids for protein formation, the preferential coding of codons into amino acids, and the correlations between the G + C contents in the third codon positions with G + C contents in the surrounding DNA. These factors may cause the concentrations of nucleotides in the three codon positions to be different. Such a positional asymmetry is believed to be the cause of the pronounced period-three pattern in the coding segment correlations and mutual information. The pronounced periodicity mentioned in [12] has also been used to differentiate coding and noncoding DNA segments [25]. Covariance matrix decay is used for analysis of correlation functions in [13]. The observations of long-range correlations and the various periodicities in the observed correlations are related to biological facts in genomes.

The characterization of coding and noncoding regions based on the mutual information function is described in [25]. That paper basically explores the existence of phylogenetic origin-free statistical features in coding and noncoding regions. The mutual information function decays to zero for noncoding DNA, whereas it oscillates for coding DNA with a period of three. Gene identification based on the mutual information function is reported to perform better than traditional techniques which require training on datasets [26]. A number of other information theory measures have also been used for coding segment characterization [5, 18, 23, 27, 28, 29, 30, 31]. A measure for sequence complexity is presented in [23]. The sequence compositional complexity is based on an entropic segmentation method to divide a sequence into homogenous segments. The complexity measure is compared for coding and noncoding segments and is related to the correlation structure. An entropic segmentation method is also used in finding borders between coding and noncoding regions [27]. A 12-letter alphabet or mapping rule is used, which takes into account the

differential base composition at each codon position. This is used to find different compositional domains for coding and noncoding regions. General statistical properties of coding regions are used in the segmentation, and this method is reported to be highly accurate in identifying borders. Another information theory tool which has been reported to be useful in the analysis of DNA sequences is given in [28]. This is the Jensen-Shannon divergence which quantifies the difference between different statistical distributions. A description of statistical properties of the divergence measure is followed by the application to the analysis of DNA sequences. The segmentation method based on the divergence measure is reported to segment a nonstationary sequence into stationary subsequences, and is also applied to DNA. Finally, a good overview on information theory and applications to molecular biology can be found in [32].

3.2. DSP techniques for DNA sequence analysis

The string of nucleotides in the DNA sequence is a categorical or symbolic sequence. Each of the nucleotides is assigned a numerical value, in order to apply DSP methods. Examples of such numerical assignment techniques are the binary indicator sequences [6] or the assignment of the integers 1, 2, 3, and 4 to A, C, G, and T, respectively [33]. The numerical sequences thus obtained are analyzed using DSP methods. Tiwari et al. [1] identify coding regions in DNA sequences by computing the Fourier spectra of a moving window across the sequence. The value of the spectrum at $f = 1/3$, is used to clarify the DNA regions as either coding or noncoding. The relative strength of the periodicity is used as the coding measure (ratio of the spectral value at $f = 1/3$ to the average spectrum). The effectiveness of the GeneScan method in identifying coding regions is also discussed. The method is robust to sequencing errors resulting from frameshift errors; the computations are simple and training is not required, which is an additional advantage. Anastassiou [2] extends on the ideas from [1, 3] and provides a method to differentiate coding and noncoding regions based on weighted spectra. Two numerical assignment schemes, namely, binary and complex number assignments are used for analysis in [2]. A procedure to compute the protein sequence from the coding regions, based on the principles of finite impulse response filters and quantization, is also described. Methods to calculate DNA spectrograms, and the use of power spectra to identify coding regions, are given. The paper also describes the method for the identification of reading frames and summarizes the uses of DSP-based techniques in DNA sequence analysis. Analysis of chromosome genomic signals has also been carried out using a complex numerical representation of nucleotides [34]. Therein, a model of the structure of the chromosome has been presented through techniques such as phase analysis, two- and three-dimensional sequence path analysis, and statistical analysis. The signal processing of symbolic sequences has also been addressed in [35, 36]. In [35], binary indicator sequences are used for DNA sequence analysis. For any mapping rule, a symbolic sequence is mapped to a numerical sequence by assigning a weight to each symbol. This mapping can be represented as a matrix multiplication. The subsequent linear transformation of the numerical sequence can also be represented by a matrix multiplication operation. Since linear transformations are performed, the weights can be optimized to obtain a required property in the transformed signal. These operations are explained in the case of discrete Fourier transforms (DFTs). The computation of linear transforms for symbolic signals is also explained in [36]. Spectral and wavelet analyses of symbolic sequences are explained and applied to DNA sequences, and results are presented for "pseudo DNA" sequences and *E. Coli* DNA.

Concepts from digital IIR filtering were used in [4] to detect coding regions. This paper uses antinotch IIR filters to identify these regions. This is achieved by designing a filter which has a sharp frequency response peak at $2\pi/3$. On passing the nucleotide sequence through this filter, if the sequence is from a coding region, the output will have a pronounced frequency peak at $2\pi/3$. The authors explain various tradeoffs in the design of the IIR filter and efficient design procedures. They conclude with examples where the output of the antinotch filter has a more discernible spectral peak at $2\pi/3$ when coding sequences are analyzed.

Two DSP-based approaches to genome sequences analysis are explained in [24]. The methods are the three-dimensional DNA walks and Gauss wavelet-based analysis, and Huffman-based encoding technique. The three-dimensional DNA walk is used as a tool to visualize changes in nucleotide composition, base pair patterns, and evolution along the DNA sequence. The proposed DNA walk model is reported to provide similar results as those obtained from a purine-pyrimidine walk, in terms of long-range correlations. Gauss wavelet analysis is then used to analyze the fractal structure of the three-dimensional DNA walk. With the use of Huffman coding, the transformation of the DNA sequence into an encoded domain can help visualize the sequences from a new perspective.

The spectral analysis of a categorical time series is explained in [37, 38]. In [37], the statistical theory for analyzing a categorical time series in the frequency domain is discussed, and the methodology that is developed is applied to DNA sequences. A discussion on the application of the spectral envelope methodology to a number of sequences, including the DNA, is given in [38]. Various spectral peaks in the sequence can be observed in the spectral envelope that is obtained through this technique. Techniques based on time-frequency and wavelet analysis have also been used to analyze DNA and protein sequences [18, 39, 40, 41].

3.3. Numerical mapping of nucleotides

Numerical mapping can be broadly classified into two types, namely, fixed mapping as in [1, 2, 4, 5, 6, 7, 8, 13, 16, 17, 24, 33] and a mapping based on some optimality criterion as in [36, 37]. Fixed mappings include binary [8], integer [33], and complex representations [2]. In this work, we use a real-number mapping rule based on the complement property of the complex mapping in [2]. The real-number representation is A $= -1.5$; T $= 1.5$; C $= 0.5$; and G $= -0.5$.

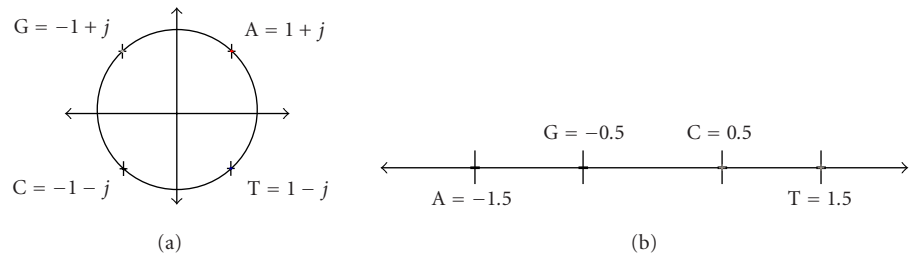

Figure 2: A constellation diagram for (a) complex-number representation and (b) real-number representations.

The complement of a sequence of nucleotides can be obtained by changing the sign of the equivalent number sequence and reversing the sequence. For example, CTGAA: 0.5; 1.5; −0.5; −1.5; −1.5 → Change Sign and Reverse Sequence → 1.5; 1.5; 0.5; −1.5; −0.5: TTCAG. In the computation of correlations, real representations are preferred over complex representations. Furthermore, it is interesting to note that the complex, real, and integer representations can also be viewed as constellation diagrams, which are widely used in digital communications. Figure 2 shows the constellation diagram for the complex and real representations. The complex constellation is similar to that of the quadrature phase shift keying (QPSK) scheme, and the real representation is similar to the pulse amplitude modulation (PAM) scheme. The constellation diagram helps visualize the DNA sequence in the context of digital communications, where a symbol mapping is followed by transmission of information. Analysis of DNA sequences using digital communications techniques could reveal certain aspects of the DNA like error-correcting capability. An information theory perspective of information transmission in the DNA, namely, the central dogma, is explained in [32].

4. AR MODEL-BASED DNA SEQUENCE ANALYSIS

The aforementioned DNA sequence analysis techniques can be divided into two main categories. In the first category, correlations within coding and noncoding sequences are characterized and used thereafter. In the second category, the Fourier transform of sequences is used to observe spectral characteristics that could distinguish between coding and noncoding DNA regions. The typical spectral signature found in a coding region is a spectral peak [1], and AR spectral estimators are effective in modeling spectral peaks of short sequences [19]. AR spectral parameters can also reflect the underlying difference in the correlation structure between coding and noncoding regions. Since correlations have been related to biological properties of the DNA, AR models could also be used as models of biological functions. Hence, it is a logical extension to use AR spectral estimators to analyze DNA sequences.

4.1. AR modeling

The AR modeling of DNA sequences can be performed using linear prediction techniques. In the linear prediction analysis, a sample in a numerical sequence is approximated by a linear combination of either preceding or future sequence values [42]. The forward linear prediction operation is given by

$$e(n) = x(n) - a_1 x(n-1) - a_2 x(n-2) - \cdots - a_p x(n-p), \tag{3}$$

where x is the numerical sequence, n is the current sample index, $a_1, a_2, \ldots, a_p$ are the linear prediction parameters, and $e(n)$ is the linear prediction error. Equation (3) represents forward linear prediction since the current sample is predicted by a linear combination of previous samples. Similarly, in backward linear prediction, a sample is predicted as a linear combination of future samples. The linear prediction coefficients are calculated by minimizing the mean squared error. The linear prediction polynomial is given by

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}. \tag{4}$$

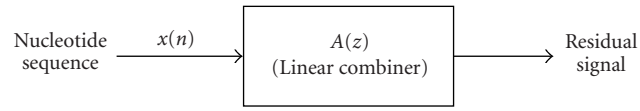

Figure 3: AR process and linear prediction; $A(z)$ is the filter polynomial.

Figure 3 depicts the DNA linear prediction in the context of AR processes.

The output of the linear combiner is known as the residual signal. In speech processing, linear prediction has been used for efficient modeling with a considerable level of success [43]. The AR Yule-Walker and Burg algorithms are widely used to compute the AR model parameters. The involved autocorrelation matrix values are typically calculated using the biased estimate in (2). Issues related to the AR modeling of DNA sequences are discussed in Section 4.2.

4.2. Proposed AR model-based DNA sequence analysis

The AR modeling of a DNA sequence is done by first mapping the sequence into the numerical domain and then calculating the AR parameters of the resulting numerical sequence. Since the numerical mapping of the DNA affects

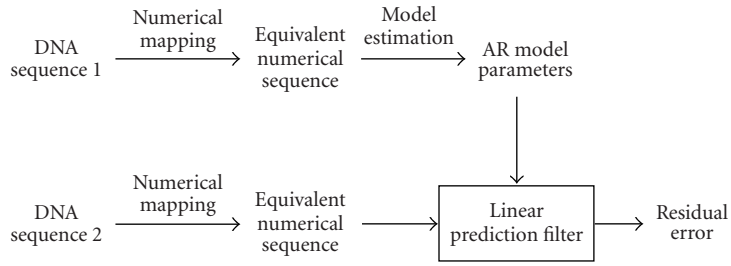

Figure 4: Block diagram of AR model-based residual signal analysis of DNA segments.

the correlation function [5], the AR parameters, which are derived from the correlation values, also depend on the numerical assignment. In this paper, the real, integer, and binary mapping rules [8] have been used for analysis. Another important issue pertains to the application of AR modeling to DNA sequences. As mentioned in Section 4.1, the calculation of AR parameters from the linear prediction model involves minimizing the error between the current signal sample and a linear combination of past samples. This definition pertains to causal AR modeling. In the case of DNA sequences, there appears to be no constraint to consider only a causal AR model, since the nucleotides in a spatial series need not be constrained to depend on the ones positioned before them only. However, the protein coding information is stored in nucleotide triplets and certain codons signal the start and stop of these gene regions. The start/stop codons and the transcription of the nucleotide triplets implicitly confer directionality to the nucleotide sequences in the genes. Hence, a causal AR model appears to be more appropriate for modeling gene sequences. The fact that the polymerase enzyme which is responsible for reading the information from the genes physically reads this DNA information from the start to the stop codons augurs our assumption. However, it needs to be noted that no such directionality apparently exists in noncoding regions and it would thus be of considerable interest to analyze both coding and noncoding DNA regions with causal versus noncausal models, respectively.

AR models of DNA sequences were used to perform two basic kinds of analyses. In the first analysis, the residual error variance of DNA sequences was used as a measure to indicate the "goodness" of the AR fit. In other words, AR models of various DNA segments were compared based on their AR residual signal. That is, suppose that signals $s_1(n)$ and $s_2(n)$ are modeled using respective AR models. When $s_1(n)$ is input to the linear predictor defined by the parameters of the AR model of $s_2(n)$, the residual signal error would be lower if $s_1(n)$ and $s_2(n)$ are described by similar AR models than if described by different AR models. The residual signal can thus be used as a measure of similarity between two signals (e.g., two DNA regions). Furthermore, it is evident that the residual error (a one-dimensional measure) alone is not sufficient to parameterize multidimensional signals, that is, different signals may yield similar residual error values. Thus, the inadequacy of the residual error was one of the motivations to use AR model parameters as sequence features. For example, if the parameters $a_1, a_2, \ldots, a_p$ are obtained by AR analysis of a gene segment, the vector $[1, a_1, a_2, \ldots, a_p]^{\mathrm{T}}$ is used as the segment feature. This is similar to the analysis of speech signals, where the AR model parameters or their derivatives, such as cepstral parameters, are used as feature vectors. Furthermore, by representing DNA sequences of different lengths with AR models of equal order, their comparison becomes possible by many simple measures such as Euclidean distance and vector correlations. Subsequently, AR features of coding and noncoding DNA sequences were analyzed using techniques such as feature space distribution analysis. Finally, we did not use the AR spectrum to distinguish between coding and noncoding features. This is due to the fact that working with high-order AR models, spurious spectral peaks were observed.

4.3. Analyzed DNA sequences

The analyses presented herein were performed on the *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Streptococcus agalactiae* genomes. The *S. cerevisiae* genome has 16 chromosomes and its complete length is approximately 12 million bp. *C. elegans* and *C. cerevisiae* are eukaryotes, while *S. agalactiae* is a prokaryotic organism.

Prokaryotes are single-celled organisms while eukaryotes can be single- or multicelled. Major differences between prokaryotic and eukaryotic genomes are that the genome size of prokaryotes is typically less than that of eukaryotes, and that prokaryotic DNA has a higher percentage of genetic information content in contiguous gene segments than eukaryotic DNA. Furthermore, the number of repetitive sequences in eukaryote DNA sequences is larger than the number of repeats in prokaryote DNA. The above-mentioned genomes can be obtained from the National Center for Biotechnology Information (NCBI) public database.

5. RESULTS

5.1. Residual error analysis

We will first discuss the AR residual error-based DNA analysis. Results only from the analysis of *S. cerevisiae* chromosome 4 DNA sequence are presented herein. The binary SW mapping rule [8] and the real-number mapping rule were used. The analysis' block diagram is shown in Figure 4. AR models of coding and noncoding DNA regions were compared based on their AR residual errors as follows.

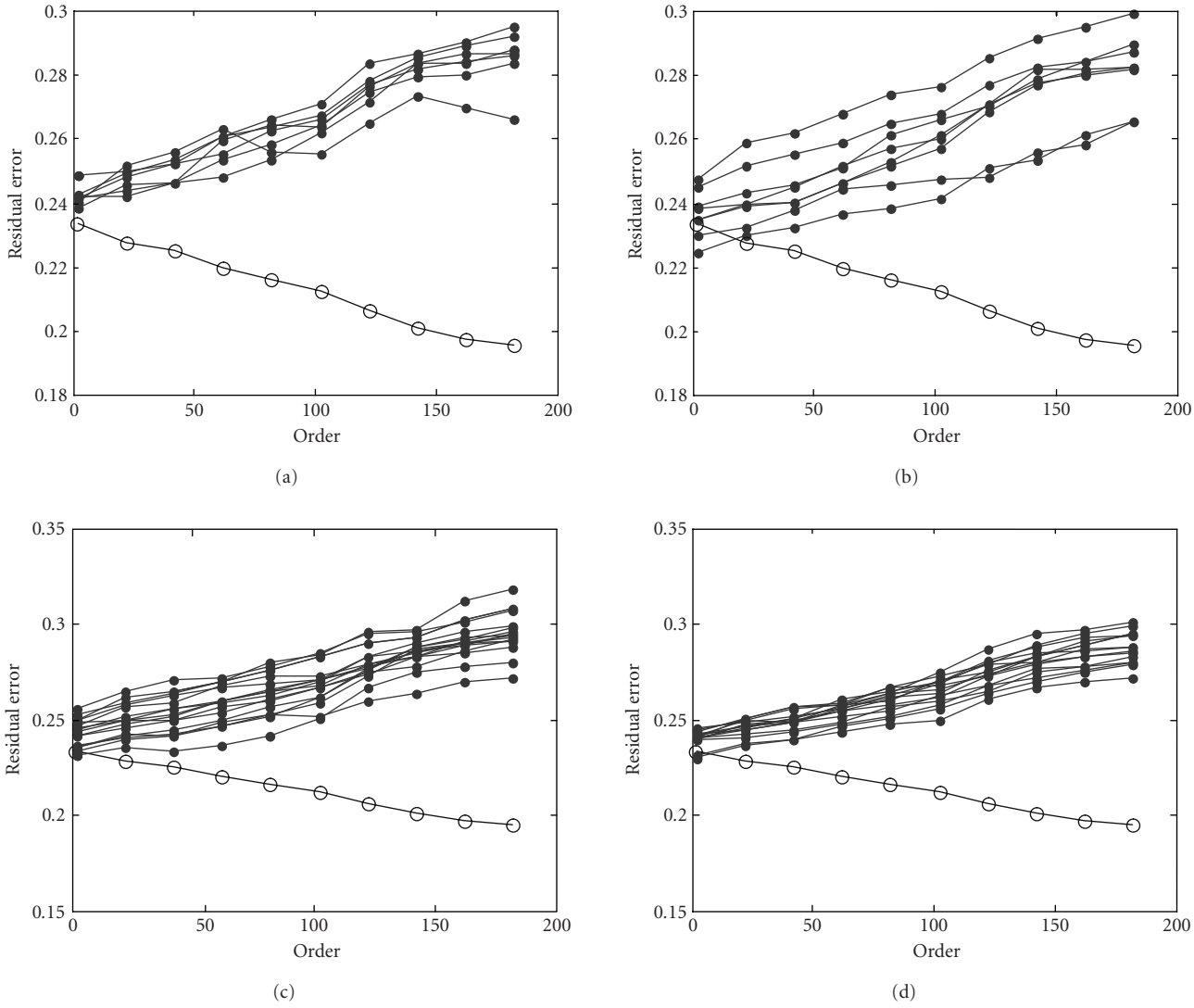

Figure 5: AR model of gene 1 of *S. cerevisiae* is used to perform residual signal analysis on its other genes using binary mapping. Residual signal variance versus AR model for gene 1 (-o-) and other genes (-•-) from chromosome 4, (a) error in gene 1 and genes 3–9; (b) error in gene 1 and genes 11–18; (c) error in gene 1 and genes 20–35; and (d) error in gene 1 and genes 36–50. Genes of length less than 150 bp were not considered since they cannot be modeled using high-order AR models.

First, the AR models were computed for each gene. Then, these AR model parameters were used to perform linear prediction and obtain the residual signal variances when applied to other genes. Genes of shorter length for which higher-order AR models could not be computed were not considered. The residual signal variances from 47 genes obtained with the AR model of gene 1 are shown in Figure 5. It can be noted that with increasing AR model order, the residual signal variance in gene 1 decreases. This is in conformance with the well-known fact from statistical signal processing that when a signal is modeled using AR models of increasing order, the residual signal error for that signal decreases monotonically [19]. On the other hand, it is interesting to note that for the other gene sequences, the residual error variance increases with increasing AR model order (see Figure 5). A similar result was observed when the real mapping rule was used (see Figure 6). This observation implies that with increasing model order, the similarity between the AR models of different genes decreases due to the increased specificity of the AR models to genes. The specificity could be due to the absence of redundancy between the analyzed genes and emphasizes the idea that, since different genes typically code for different amino acid sequences, they may not contain a lot of similar or redundant information.

Next, noncoding segments were compared with coding segments. Gene 1 in chromosome 4 of *S. cerevisiae* was modeled using an AR model, and the model parameters were used to compute the residual error variances of 50 noncoding

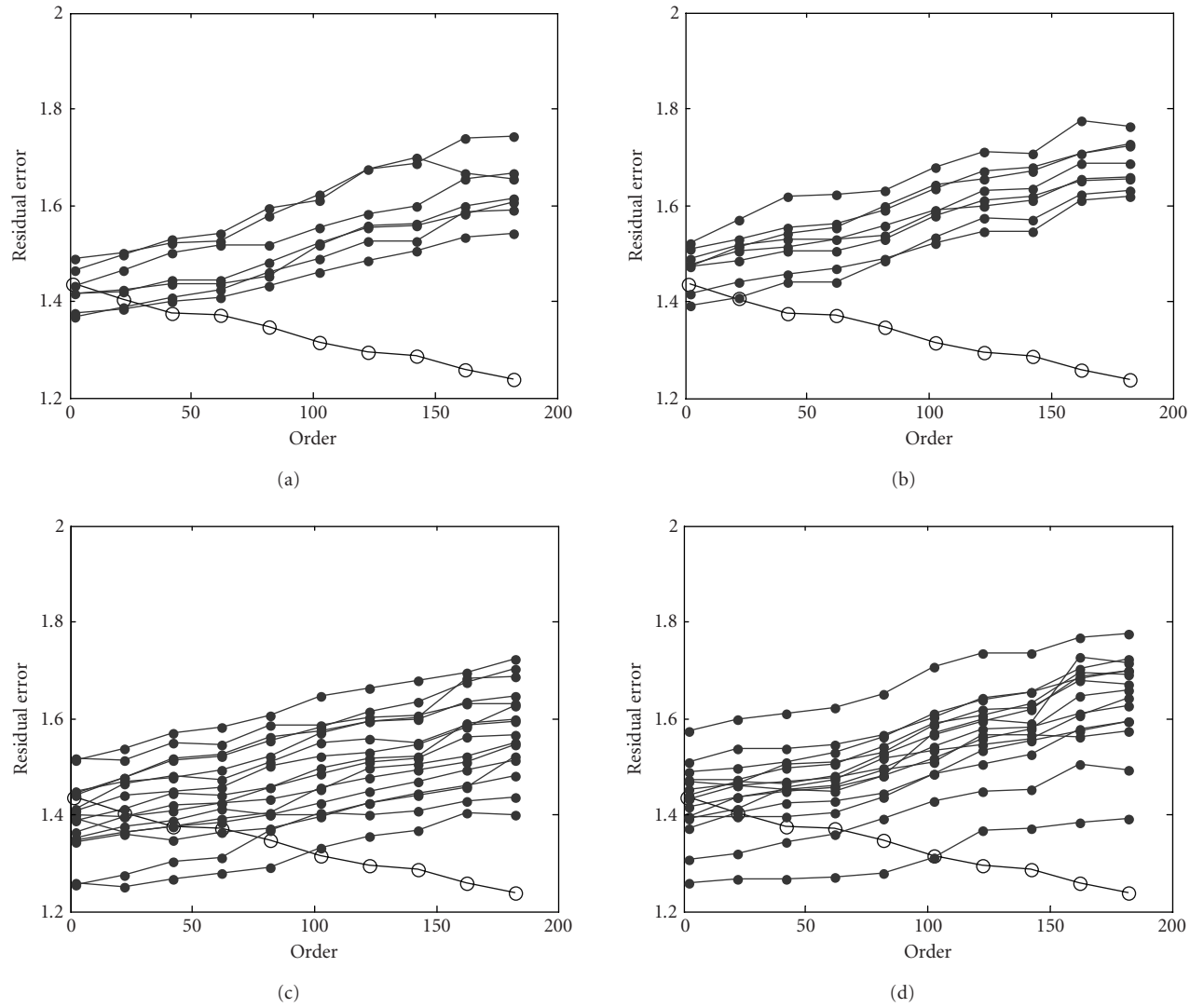

Figure 6: AR model of gene 1 of of *S. cerevisiae* is used to perform residual signal analysis on its other genes using real-number mapping. Residual signal variance versus AR model for gene 1 (-o-) and other genes (-•-) from chromosome 4, (a) error in gene 1 and genes 3–9; (b) error in gene 1 and genes 11–18; (c) error in gene 1 and genes 20–35; and (d) error in gene 1 and genes 36–50.

segments. Similarly, gene 17 was modeled using an AR model and the model parameters were used to compute the residual error variances of 50 noncoding segments. The residual error variances of 50 noncoding segments when the AR model from gene 1 and gene 17 was applied are depicted in Figures 7 and 8, respectively. It can be observed that the residual signal variance values for a few noncoding sequences are smaller than the ones for gene 1, for the full range of model orders. This implies the existence of similarities between coding and noncoding segments. Similar observations were also obtained when real mapping was applied.

It is evident from the above observations that the classification of an analyzed sequence to either a coding or noncoding region based on the residual signal alone is difficult as different regions may have similar residual errors for a range of AR model orders. The above results also show that when AR models are used to parameterize DNA segments based on the residual error, higher-order models may be required to model the characteristics and capture their differences.

5.2. AR feature-based analysis

One of the important problems in DNA sequence analysis is identifying regions with similar nucleotide compositions. This is then typically applied in studies such as identifying conserved regions across different organisms. A number of algorithms, such as BLAST, have been developed to perform string searches and template matching. These string searching tools are typically based on dynamic programming concepts, wherein the actual template or query string is compared with segments of a long DNA sequence. In this paper,

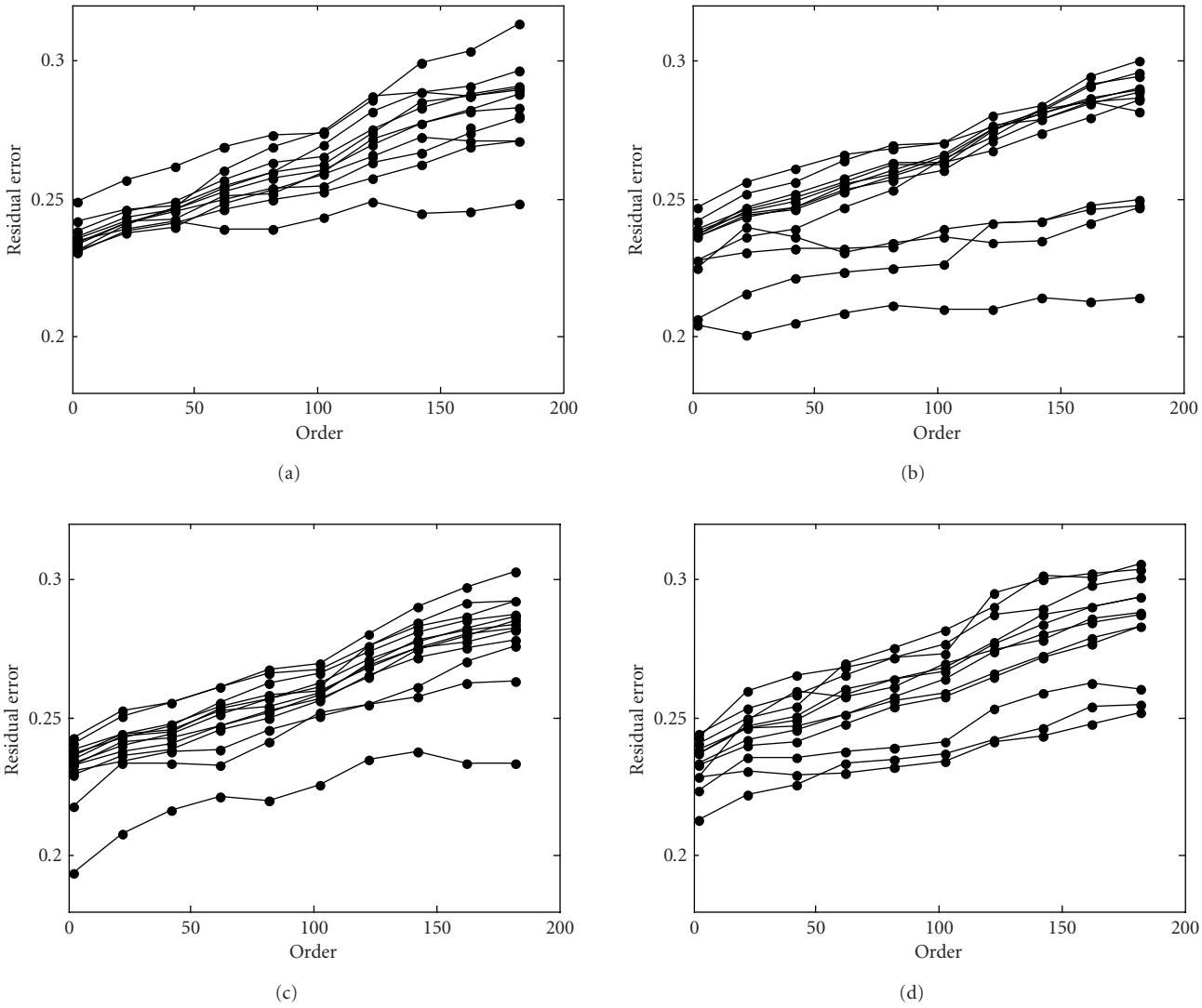

Figure 7: AR model of gene 1 is used for linear prediction on 50 noncoding segments using binary mapping. (a) Error in noncoding segments 1–12; (b) error in noncoding segments 13–25; (c) error in noncoding segments 26–38; and (d) error in noncoding segments 39–50.

the AR model parameters of the template nucleotide sequence are used as features to identify similar segments in a long DNA sequence. AR models capture the global spectral characteristics of the modeled sequences. Thus, the identification is based on similar spectral characteristics (AR) rather than one-to-one nucleotide matching (dynamic programming techniques).

The analysis was performed on a segment of the *S. cerevisiae* genome using binary, real-number, and integer mapping. The template matching procedure was performed as follows. First, a segment of nucleotides of length L was chosen as the template. The AR model of this template was estimated for various orders, and the model parameters were used as template features. Second, the AR features were calculated over the whole DNA sequence from overlapping moving windows of the same length L as the template. Third, the feature vectors obtained from each moving window were compared with the template feature vector by computing the Euclidean distance between them.

It was observed that using the real mapping, similar segments to either the template, its reversed sequence, its complementary sequence, or its reversed complementary sequence are detected. One such example is presented in Table 1, wherein the template and its complement were identified. Using integer mapping, the DNA locations where similar features were found are cited in Table 2. In this case, the features of the template sequence alone was detected. Using binary SW mapping, although the actual template occurred only once in the complete sequence, other segments also yielded the same features (see Table 3). Here the template and the matched sequences differ in the actual nucleotide but on a closer look, they have a similar sequence of strong and weak

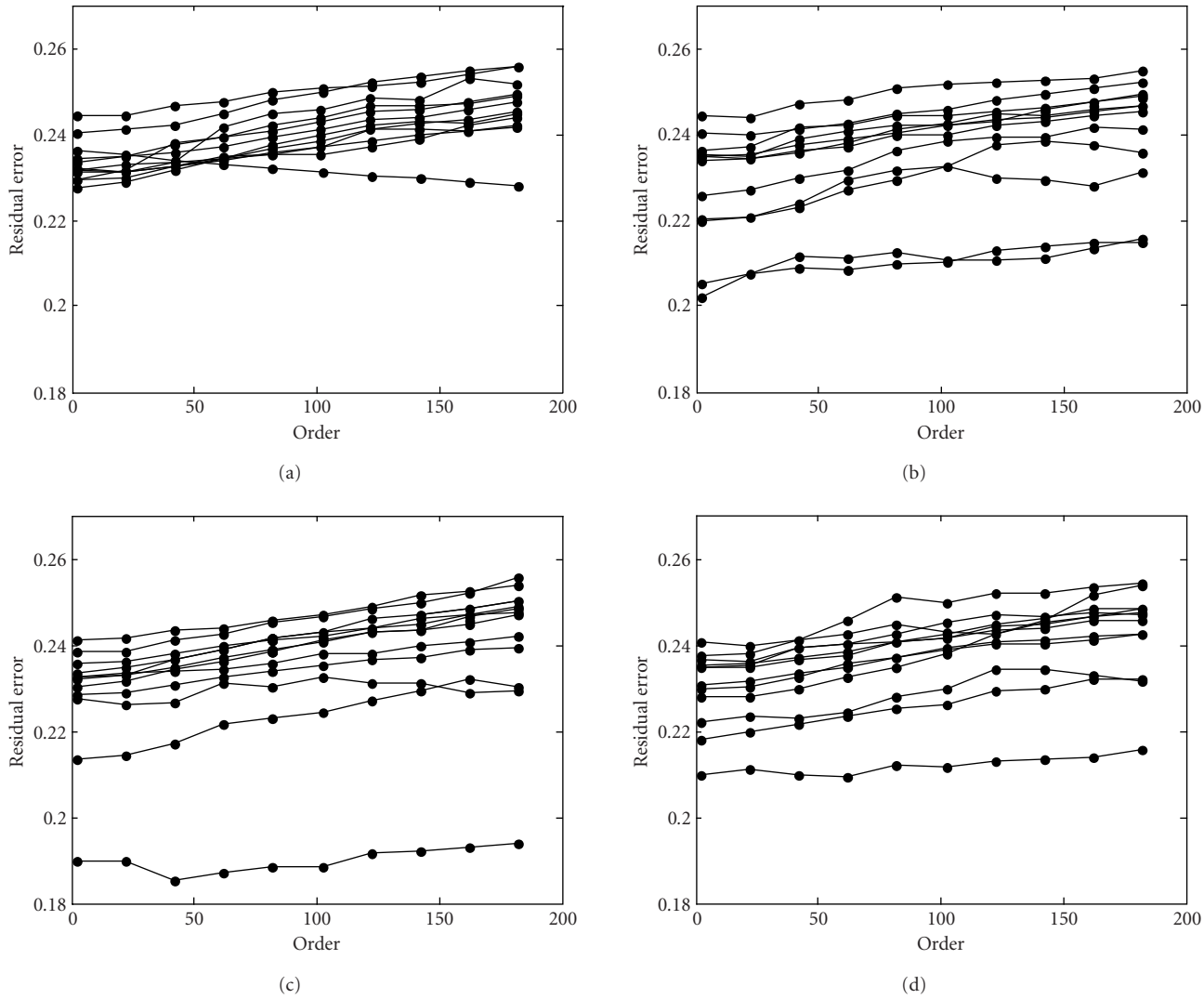

Figure 8: AR model of gene 17 is used for linear prediction of 50 noncoding segments using binary mapping. (a) Error in noncoding sequences 1–12; (b) error in noncoding sequences 13–25; (c) error in noncoding sequences 26–38; and (d) error in noncoding sequences 39–50.

hydrogen bonds. Analysis with the binary RY mapping rule [8] yielded similar results, that is, segments with a similar sequence of purines and pyrimidines as the one in the template.

In the aforementioned analysis, the mapping rule used played an important role in identifying matches. The real- and integer-number mapping rules yielded different string matches. This is due to the inherent complementary property of the real mapping rule and the noncomplementary property of the integer mapping rule. The difference is further elucidated through the following exercise. Say, for example, the occurrences of the template 5′-TACGTGC-3′ need to be found in a long DNA string. The corresponding numerical sequence obtained through real mapping would be 5′-1.5, −1.5, 0.5, −0.5, 1.5, −0.5, 0.5-3′. The following numerical sequences will have the same AR parameters as the above template:

(i) 5′- −1.5, 1.5, −0.5, 0.5, −1.5, 0.5, −0.5-3′ = 5′-ATGCACG-3′: (reversed complement of the template);

(ii) 5′-0.5, −0.5, 1.5, −0.5, 0.5, −1.5, 1.5-3′ = 5′-CGTGCAT-3′: (reversed template);

(iii) 5′- −0.5, 0.5, −1.5, 0.5, −0.5, 1.5, −1.5-3′ = 5′-GCACGTA-3′: (complement of the template).

This is due to the fact that (a) the sign-reversed numerical sequence and the actual numerical sequence have the same linear dependence and hence the same AR parameters, and (b) minimizing the forward or the backward linear prediction error would theoretically yield the same AR model. This is observed with the Burg algorithm AR estimation, wherein

Table 1: Detection of repeats of DNA segments via AR modeling. Real mapping rule and second-order AR model features are used; the template is 8 bp long. There are 5 repeats in the whole sequence. Identification of complementary and reversed sequences is obtained as well.

Position with the same features	DNA segment
210–217 (template)	CTCACATT
5174–5181	CTCACATT
12572–12579	CTCACATT
19278–19285	AATGTGAG
29624–29631	CTCACATT
36387–36394	AATGTGAG
55805–55812	AATGTGAG
63106–63113	CTCACATT

Table 2: Detection of repeats of DNA segments via AR modeling. Integer mapping rule and second-order AR model features are used; the template is 8 bp long. There are 5 repeats in the whole sequence. The template is exactly identified.

Position with the same features	DNA segment
210–217 (template)	CTCACATT
5174–5181	CTCACATT
12572–12579	CTCACATT
29624–29631	CTCACATT
63106–63113	CTCACATT

Table 3: Detection of repeats of DNA segments via AR modeling. Binary SW mapping rule and fourth-order AR model features are used; the template is 14 bp long and it has one occurrence in the whole sequence. Identification of DNA with similar sequences of strong and weak hydrogen bonds is obtained. Nucleotides C and G (mapped to one), A and T (mapped to zero) are highlighted differently.

Position with the same features	DNA segment
210–221 (template)	C T C A C ATTA CCC TA
7424–7435	C T C T G AAAT GCC AT
9283–9294	G A C T G ATAA GGG TT
80726–80737	C A G T G ATAT CGG TA

both the forward and backward linear prediction errors are minimized together. In the case of the integer mapping rule ($A = 1, C = 2, G = 3, T = 4$), the corresponding numerical sequence of the template is $5'$-$4, 1, 2, 3, 4, 3, 2$-$3'$. The reversed sequence, namely, $2, 3, 4, 3, 2, 1, 4$, has the same AR model parameters as the template (by minimizing the forward and reverse prediction errors). On the other hand, the sequence corresponding to the complement of the template may not have the same AR model. Hence, using the integer mapping rule, the exact template and its reversed sequence are matched.

The features of the nucleotide segments are also affected by the use of the binary mapping rule. This is explained through the following example. The sequence $5'$-TGACAAGC-$3'$ is mapped to $5'$-$0, 1, 0, 1, 0, 0, 1, 1$-$3'$ using the binary SW mapping rule. The above numerical sequence also corresponds to $5'$-ACACATGG-$3'$, and a number of other nucleotide combinations. The AR model parameters of all these combinations are the same, and hence, it is possible to identify sequences with certain similar chemical properties like similar sequences of strong and weak hydrogen bonds.

The above observations are of great interest because they show that identification of regions with similar biological/chemical properties may be possible using AR feature-based template matching under different mapping rules. For example, the ability to identify a template and its complement can help in identifying genes in complementary strands as well, which may not be possible in a single "run" using traditional string searching tools. The AR model string search method can be used as an analytical tool to reveal additional information about the interrelations between different DNA sequences. The knowledge acquired by this analysis could be used in knowledge or rule-based methods. Two DNA signals with similar AR spectra are more related in a global manner than in a one-to-one nucleotide basis. In this sense, the above method can provide clues about similarities between apparently nonidentical DNA sequences that could then be used in the identification of the underlying biochemical mechanisms of such similarities. The results of AR model-based analysis are related to fast Fourier transform (FFT)-based methods. The pros and cons have to do with the well-known advantages and disadvantages of using parametric versus nonparametric signal processing methods (e.g., ability to analyze short versus long segments, computational speed, etc).

The above algorithm was also applied to gene searches in a long string of DNA. It was observed that the distance between the feature vectors is zero at the exact location of the gene even with an AR model of an order as low as 2. The distance between the gene sequence AR feature vector and the moving window AR feature vector is plotted for various feature dimensions (AR model orders) in Figure 9. It was also observed that the average distance between the gene feature vector and features of the moving windows increased with AR model order. It can be typically expected that the average distance between vectors tends to increase with increasing dimension. Nevertheless, in conjunction with our previous observations from the residual signal-based analysis, it appears that the increasing average distance of the gene features with the AR model orders may mainly be due to the greater specificity of the AR modeling to the presence of genes. To further investigate the above observations, a study of the distributions of coding and noncoding AR features was undertaken.

The complete *S. cerevisiae* genome with all coding and noncoding sequences was considered. We mapped the DNA segments into the numerical domain using the binary SW mapping rule. Then, the AR model parameters of all segments were calculated and used as the DNA segment features.

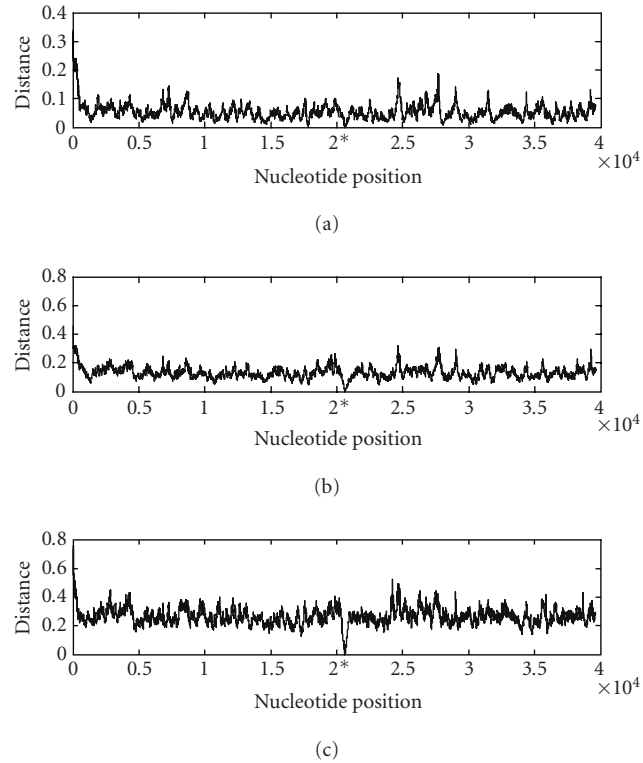

Figure 9: The distance between the feature vector of a gene sequence (position denoted by $*$) and the corresponding features within a moving window segments over the analyzed DNA sequence from *S. cerevisiae* for AR model orders (a) 10, (b) 25, and (c) 50 (real mapping used). It can be noticed that the average distance between the gene feature and the features of the moving windows increases with AR model order, and it is minimal (zero) at the position of the gene.

The analysis was also performed using the real mapping rule. For a particular AR model of order p, the centroid of all coding region feature vectors was calculated, and the Euclidean distance of the feature vectors from the centroid was computed. The distances were similarly computed for noncoding region features from their centroid as well. The distribution density of these distance measures was obtained. The process was repeated for increasing model orders. The distributions from the coding region and noncoding regions were then compared using the Kolmogorov-Smirnov test [44]. Figure 10 shows the distribution densities for *S. cerevisiae* coding and noncoding regions for AR model orders 15 and 35, using binary SW mapping. The distribution densities obtained by using real-number mapping are depicted in Figure 11. Both coding and noncoding features are concentrated near their respective centroids. The noncoding features appear to be more concentrated around their centroid than the coding features.

The p values from the Kolmogorov-Smirnov test of the distributions of the coding and noncoding features using binary SW and real-number mapping, are shown in Figure 12. It is observed that the threshold $p = 0.05$ used in the hypothesis testing is achieved with an AR model order of 21 for the binary SW mapping and only 16 for the real mapping. Thus, it appears that such distance distributions can be used to further classify a DNA segment as coding or noncoding. It also appears that the real mapping is more effective than the binary SW mapping in this analysis.

6. CONCLUSION

A brief survey of the research on the analysis of DNA sequences from a signal processing perspective was presented. The use of nonparametric classical DSP tools like Fourier transforms and time-frequency analysis have been effective in studying DNA sequences of coding and noncoding regions. The use of parametric spectral analysis to capture certain spectral characteristics of such DNA regions was herein introduced. We applied the AR spectral analysis tools to analyze DNA sequences.

The analyses were of two basic types. First, the AR model parameters of the analyzed DNA segments were used to perform linear prediction analysis. The residual error was subsequently used to compare the analyzed segments. An observation of particular interest was that the AR model was very specific to the coding DNA sequences. This specificity increased with increasing model orders. Though the residual error analysis methodology could be used to compare AR models of different DNA segments, it was found not to be adequate for the characterization of these sequences. The AR

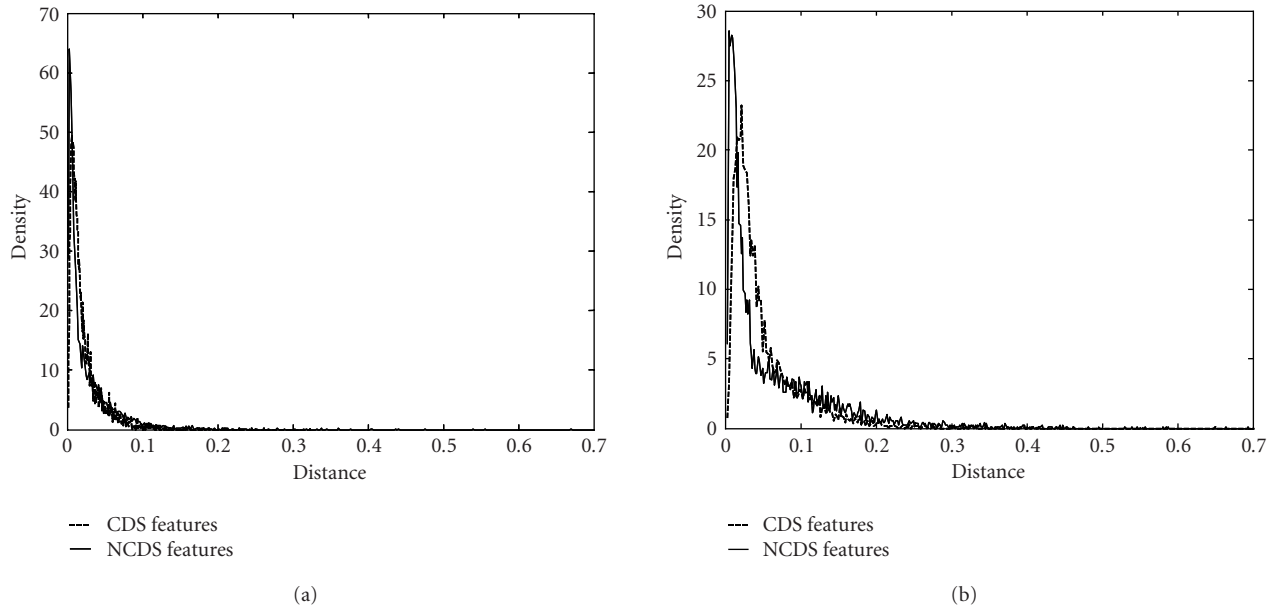

Figure 10: Distribution density of distances of coding segment (CDS) AR feature vectors and noncoding segment (NCDS) AR feature vectors from their respective centroids for AR model orders (a) 15 and (b) 35 (binary SW mapping used).

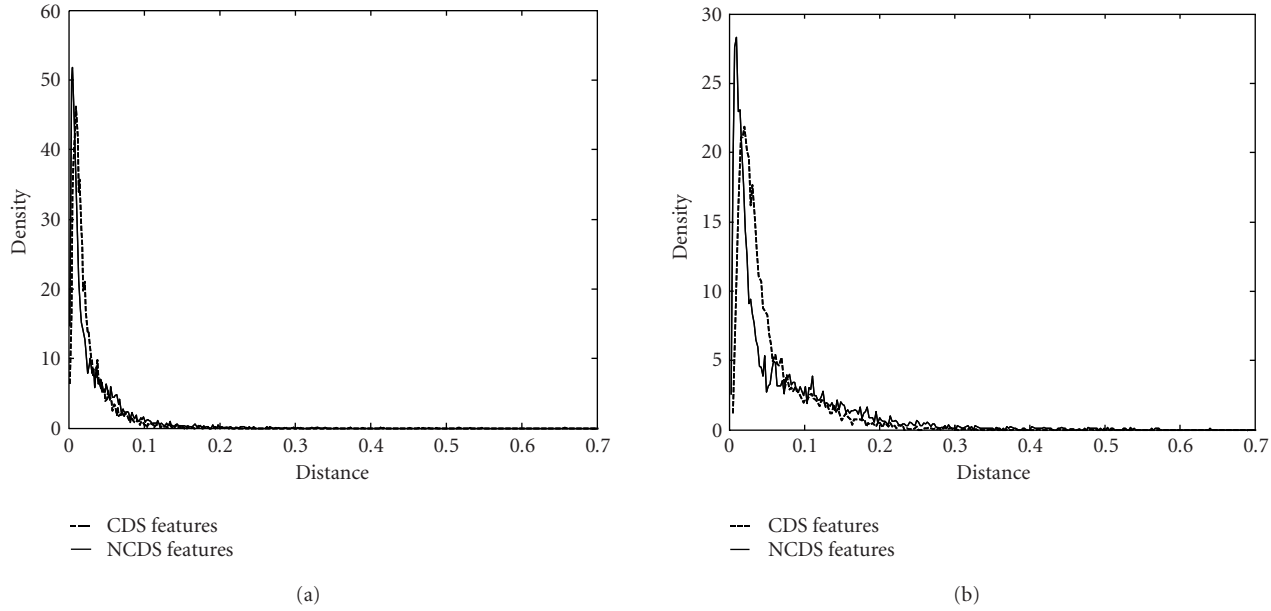

Figure 11: Distribution density of distances of coding segment (CDS) AR feature vectors and noncoding segment (NCDS) AR feature vectors from their respective centroids for AR model orders (a) 15 and (b) 35 (real mapping used).

model parameters themselves were then used as features for DNA string searches.

Depending on the type of the numerical mapping rule used, the AR feature-based string searching technique was highly effective in identifying all repeats of the query string, along with the locations of its complementary sequence. It was also possible to locate regions with similar chemical structures, for example, sequences of similar strong and weak hydrogen bonds. Thus different mapping rules can be used depending on the objective of the analysis. For example, the use of SW or RY mapping rules was necessary to locate regions of similar strong-weak hydrogen bonds or purine-pyrimidine structure. It was observed that modeling with a low-order AR model and working in the generated feature space was sufficient to locate the occurrence of complete genes in a long DNA sequence. Further analysis of the

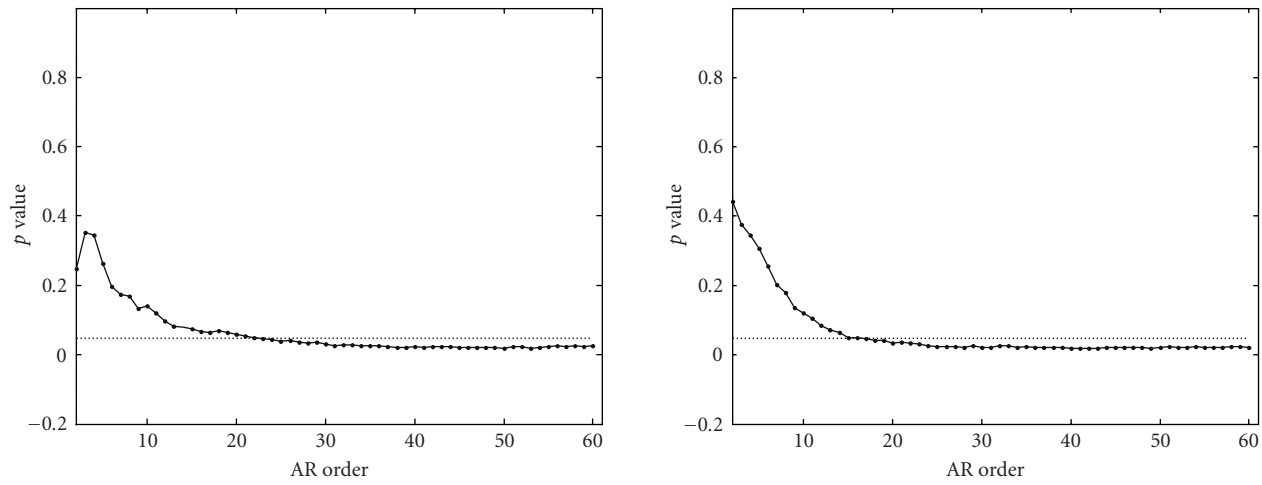

FIGURE 12: p values obtained from the Kolmogorov-Smirnov test, comparing the distribution of coding and noncoding AR features for (a) binary SW mapping and (b) real mapping. The 5% threshold used in the hypothesis testing is also plotted as a dotted horizontal line.

distribution of the coding and noncoding AR features revealed that these distributions differed significantly for high-dimension AR features. It would be of great interest to further investigate the biological implications of differences in the distributions of coding and noncoding region AR features.

The proposed analytical scheme can also be used for the analysis of other biochemical molecules, in addition to DNA, such as amino acid sequences. Further, like in speech recognition, AR features and their derivatives, such as cepstral features, could also be incorporated in an HMM-based gene-finding tool. Analysis of more genomic sequences along the lines proposed herein is underway.

ACKNOWLEDGMENTS

This work is partially supported by the National Institutes of Health through a Bioengineering Research Partnership Grant NS39687 to Dr. L. D. Iasemidis. Portions of the educational components of this work have been supported by the National Science Foundation Grant NSF0089075 to Dr. A. Spanias.

REFERENCES

[1] 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.

[2] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–10, 2001.

[3] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *Journal of Theoretical Biology*, vol. 118, pp. 295–300, 1986.

[4] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using allpass-based filters," in *Proc. Workshop on Genomic Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, USA, October 2002.

[5] H. Herzel and I. Grosse, "Measuring correlations in symbol sequences," *Physica A*, vol. 216, no. 4, pp. 518–542, 1995.

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

[7] S. V. Buldyrev, A. L. Goldberger, S. Havlin, et al., "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis," *Phys. Rev. E*, vol. 51, no. 5, pp. 5084–5091, 1995.

[8] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, "Study of statistical correlations in DNA sequences," *Gene*, vol. 300, no. 1-2, pp. 105–115, 2002.

[9] O. Weiss and H. Herzel, "Correlations in protein sequences and property codes," *Journal of Theoretical Biology*, vol. 190, no. 4, pp. 341–353, 1998.

[10] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Phys. Rev. E*, vol. 49, no. 2, pp. 1685–1689, 1994.

[11] W. Li, T. Marr, and K. Kaneko, "Understanding long-range correlations in DNA sequences," *Physica D*, vol. 75, no. 1–3, pp. 392–416, 1994.

[12] H. Herzel and I. Grosse, "Correlations in DNA sequences: The role of protein coding segments," *Phys. Rev. E*, vol. 55, no. 1, pp. 800–810, 1997.

[13] H. Herzel, E. N. Trifonov, O. Weiss, and I. Grosse, "Interpreting correlations in biosequences," *Physica A*, vol. 249, no. 1–4, pp. 449–459, 1998.

[14] W. Li, "The study of correlation structures of DNA sequences: a critical review," *Computers & Chemistry*, vol. 21, no. 4, pp. 257–272, 1997.

[15] L. Luo, W. Lee, L. Jia, F. Ji, and L. Tsai, "Statistical correlation of nucleotides in a DNA sequence," *Phys. Rev. E*, vol. 58, no. 1, pp. 861–871, 1998.

[16] D. Holste, I. Grosse, and H. Herzel, "Statistical analysis of the DNA sequence of human chromosome 22," *Phys. Rev. E*, vol. 64, no. 4, pp. 1–9, 2001.

[17] A. K. Mohanty and A. V. S. S. Narayana Rao, "Long range correlations in DNA sequences," preprint, 2002, http://arXiv.org/abs/physics/0202075.

[18] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, "Long-range correlations in genomic

DNA: a signature of the nucleosomal structure," *Phys. Rev. Lett.*, vol. 86, no. 11, pp. 2471–2474, 2001.
[19] L. S. Marple, *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1987.
[20] B. Alberts, D. Bray, A. Johnson, et al., *Essential Cell Biology*, Garland Publishing, NY, USA, 1998.
[21] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, no. 17, pp. 5303–5318, 1982.
[22] J. W. Fickett, "The gene identification problem: an overview for developers," *Computers & Chemistry*, vol. 20, no. 1, pp. 103–118, 1996.
[23] R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, "Sequence compositional complexity of DNA through an entropic segmentation method," *Phys. Rev. Lett.*, vol. 80, no. 6, pp. 1344–1347, 1998.
[24] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "New approaches to genome sequence analysis based on digital signal processing," in *Proc. Workshop on Genomic Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, USA, October 2002.
[25] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, "Species independence of mutual information in coding and noncoding DNA," *Phys. Rev. E*, vol. 61, no. 5, pp. 5624–5629, 2000.
[26] J. W. Fickett and C. S. Tung, "Assessment of protein coding measures," *Nucleic Acids Research*, vol. 20, no. 24, pp. 6441–6450, 1992.
[27] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, "Finding borders between coding and noncoding DNA regions by an entropic segmentation method," *Phys. Rev. Lett.*, vol. 85, no. 6, pp. 1342–1345, 2000.
[28] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. L. Oliver, and H. E. Stanley, "Analysis of symbolic sequences using the Jensen-Shannon divergence," *Phys. Rev. E*, vol. 65, pp. 041905-1–041905-16, 2002.
[29] M. Crochemore and R. Vérin, "Zones of low entropy in genomic sequences," *Computers & Chemistry*, vol. 23, no. 3-4, pp. 275–282, 1999.
[30] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "A coding theory framework for genetic sequence analysis," in *Proc. Workshop on Genomic Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, USA, October 2002.
[31] H. P. Yockey, "An application of information theory to the central dogma and the sequence hypothesis," *Journal of Theoretical Biology*, vol. 46, pp. 369–406, 1974.
[32] H. P. Yockey, *Information Theory and Molecular Biology*, Cambridge University Press, Cambridge, UK, 1992.
[33] 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, pp. 323–331, 1991.
[34] P. D. Cristea, "Analysis of chromosome genomic signals," in *Proc. 7th International Symposium on Signal Processing and Its Applications (ISSPA '03)*, vol. 2, pp. 49–52, Paris, France, July 2003.
[35] D. H. Johnson and W. Wang, "Symbolic signal processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '99)*, pp. 1361–1364, Phoenix, Ariz, USA, March 1999.
[36] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 50, no. 3, pp. 628–634, 2002.
[37] D. S. Stoffer, D. E. Tyler, and A. J. McDougall, "Spectral analysis for categorical time series: Scaling and the spectral envelope," *Biometrika*, vol. 80, no. 3, pp. 611–622, 1993.
[38] D. S. Stoffer, D. E. Tyler, and D. A. Wendt, "The spectral envelope and its applications," *Statistical Science*, vol. 15, no. 3, pp. 224–253, 2000.
[39] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Phys. Rev. Lett.*, vol. 74, no. 16, pp. 3293–3296, 1995.
[40] K. Bloch and G. R. Arce, "Time-frequency analysis of protein sequence data," in *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP '01)*, Baltimore, Md, USA, June 2001.
[41] J. Song, T. Ware, and S.-L. Liu, "Test of origin site (oriC) and terminus (terC) of replication by wavelet analysis in bacteria," in *Proc. Workshop on Genomic Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, USA, October 2002.
[42] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
[43] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, NY, USA, 1976.
[44] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC Press, Boca Raton, Fla, USA, 2nd edition, 2000.

Niranjan Chakravarthy received the B.E. degree in electronics and communication engineering from the Government College of Engineering, Tamil Nadu, India, in 2001 and the M.S. degree in electrical engineering from the Arizona State University in 2003. He is currently working towards the Ph.D. degree with the Department of Electrical Engineering at Arizona State University. He is a Research Assistant at the Digital Signal Processing and Brain Dynamics Laboratories and currently pursues research on the prediction and control of epileptic seizures and genomic signal processing. His research interests include digital signal processing, time series modeling, and systems theory and applications to physical and biological systems.

A. Spanias is a Professor of electrical engineering at Fulton School of Engineering, Arizona State University. His research interests are in adaptive signal processing and speech processing. He received the 2003 Teaching Award from the IEEE Phoenix Section for the development of J-DSP. He is a member of the IEEE-CAS Society DSP Technical Committee and has served as a Member in the Technical Committee on Statistical Signal and Array Processing of the IEEE Signal Processing Society (SPS). He has served as an Associate Editor of the IEEE Transactions on Signal Processing, General Cochair of the 1999 International Conference on Acoustics Speech and Signal Processing (Phoenix), IEEE Signal Processing Vice President for Conferences, and Chair of the Conference Board. He served as a Member in the IEEE Signal Processing Executive Committee and as an Associate Editor of IEEE Signal Processing Letters. He is currently serving as a Member in the IEEE SPS Publications Board, and Member-at-Large of the IEEE SPS Conference Board. He has been Chair of the Phoenix IEEE Communications and Signal Processing Chapter, and is a Member in Eta Kappa Nu and Sigma Xi. Andreas Spanias is corecipient of the 2002 IEEE Donald G. FinkPaper Award, and was recently elected as a Fellow of the IEEE. He is appointed as 2004 Distinguished Lecturer of the IEEE SPS.

L. D. Iasemidis received the Diploma in electrical and electronics engineering from the National Technical University of Athens in 1982, M.S. in Physics, M.S. and Ph.D. in biomedical engineering from the University of Michigan, Ann Arbor, Mich in 1985, 1986, and 1991, respectively. Dr. Iasemidis is currently an Associate Professor of Bioengineering at the Arizona State University, Tempe, Ariz, and Director and Founder of the ASU Brain Dynamics Laboratory. Dr. Iasemidis is recognized as an expert in dynamics of epileptic seizures, and his research and publications have stimulated an international interest in the prediction and control of epileptic seizures, and understanding of the mechanisms of epileptogenesis. He is currently on the Editorial Board of Epilepsia and IEEE Transactions on Biomedical Engineering, and is a Reviewer of NIH. He has reviewed articles for more than 10 scientific journals. His research interests are in the areas of biomedical and genomic signal processing, complex systems theory and nonlinear dynamics, neurophysiology, monitoring and analysis of the electrical and magnetic activity of the brain in epilepsy and other brain dynamical disorders, intervention and control of the CNS, neuroplasticity, rehabilitation, and neuroprosthesis. Dr. Iasemidis' research has been funded by NIH, VA, DARPA and the Whitaker Foundation.

K. Tsakalis received his Ph.D. degree in electrical engineering from the University of Southern California. He is currently a Professor of electrical engineering at Arizona State University. His interests are in robust adaptive control, time varying systems, applications of control, identification, and optimization in semiconductor manufacturing problems, and, more recently, the application of adaptive systems theory on the prediction and control of epileptic seizures.