Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction

Mahmood Akhtar, Julien Epps, *Member, IEEE*, and Eliathamby Ambikairajah, *Member, IEEE*

***Abstract*—Genomic sequence processing has been an active area of research for the past two decades and has increasingly attracted the attention of digital signal processing researchers in recent years. A challenging open problem in deoxyribonucleic acid (DNA) sequence analysis is maximizing the prediction accuracy of eukaryotic gene locations and thereby protein coding regions. In this paper, DNA symbolic-to-numeric representations are presented and compared with existing techniques in terms of relative accuracy for the gene and exon prediction problem. Novel signal processing-based gene and exon prediction methods are then evaluated together with existing approaches at a nucleotide level using the Burset/Guigo1996, HMR195, and GENSCAN standard genomic datasets. A new technique for the recognition of acceptor splice sites is then proposed, which combines signal processing-based gene and exon prediction methods with an existing data-driven statistical method. By comparison with the acceptor splice site detection method used in the gene-finding program GENSCAN, the proposed DSP-statistical hybrid technique reveals a consistent reduction in false positives at different levels of sensitivity, averaging a 43% reduction when evaluated on the GENSCAN test set.**

***Index Terms*—Autoregressive processes, correlation, deoxyribonucleic acid (DNA), discrete cosine transforms (DCTs), discrete Fourier transforms (DFTs), Gaussian mixture models.**

I. Introduction

DEOXYRIBONUCLEIC acid (DNA), the material of heredity in most living organisms, consists of genic and intergenic regions, as shown in Fig. 1. In eukaryotes, genes are further divided into relatively small protein coding segments known as *exons*, interrupted by noncoding spacers known as *introns*. In eukaryotes such as human, the intergenic and intronic regions often make up more than 95% of their genomes. *Codons* (i.e., triplets of possible four types of DNA nucleotides A, C, G, and T) in exons encode 20 amino acids and 3 terminator signals, known as *stop codons* (i.e., *TAA*, *TAG*, and *TGA*). Initial exons of the genes begin with a *start codon* "*ATG*." Looking from the $5'$ end of DNA (*upstream*) to its $3'$ end (*downstream*), the exon-to-intron border is known as the *donor splice site* and consists of a consensus dinucleotide "*GT*" as the first two nucleotides of the intron, whereas the intron-to-exon border is known as the *acceptor splice site*, which consists of a consensus dinucleotide "*AG*" as the last two nucleotides of the intron. The accurate identification of genomic protein coding regions, along with the recognition of other signals and/or regions (shown in Fig. 1) would result in an ideal gene finding and annotation system.

Despite the existence of various data-driven gene finding programs, such as AUGUSTUS [1], FGENES [2], geneid [3], GeneMark.hmm [4], Genie [5], GENSCAN [6], HMMgene [7], Morgan [8], and MZEF [9], the accuracy of gene prediction is still limited. Previous investigations of computational gene finding programs [10]–[13] reveal that these data-driven approaches seem to rely more on compositional statistics of the sequences (e.g., $G+C$ content) than the genomic signals (e.g., promoters, acceptor/donor sites, start/stop codons) involved in the translation process from DNA to protein, and are heavily dependent on the statistics of the sequences they learn from and are, thus, not equally suitable for all types of sequences. Furthermore, the accuracy is dependent on the length and position of the exons [14], [15]. High prediction accuracy can often be attributed to friendly training and test sequences, in whose formation certain rules were followed, such as including sequences consisting of one complete gene with consensus intronic dinucleotides "*GT*" and "*AG*," respectively, for their donor and acceptor splice sites, excluding those containing alternatively spliced genes and having any in-frame stop codons.

However, it has been observed that gene prediction accuracy can be substantially increased by combining different methods [16], [17]. The discrete nature of the DNA information, being discrete in both "time" and "amplitude," invites investigation by digital signal processing (DSP) techniques. The conversion of DNA nucleotide symbols into discrete numerical values enables novel and useful DSP-based applications for the solution of different sequence analysis related problems such as gene finding and annotation, and such applications have been overviewed by previous authors [18], [19]. The present role of DSP applications in this area is summarized in Fig. 2. In order to apply DSP methods, the DNA sequences are first converted into suitable numeric values. DSP-based methods for periodicity detection are then applied to the numerical sequences to obtain 1-D or multidimensional features. The resultant features are then passed on for back-end processing to classify between protein coding and noncoding regions. An empirically derived decision threshold can be used for 1-D classification,

Manuscript received September 11, 2007; revised March 8, 2008. This work was supported in part by the National University of Sciences and Technology (NUST), Pakistan, and in part by the University of New South Wales (UNSW), Australia, under a UNSW Faculty Research Grant 2007 for genomic signal processing research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ioan Tabus.

M. Akhtar is with the National University of Sciences and Technology, Rawalpindi Cantt, Pakistan, and also with the University of New South Wales, Sydney, NSW 2052, Australia (e-mail: mahmood@unsw.edu.au).

J. Epps and E. Ambikairajah are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: j.epps@unsw.edu.au, ambi@ee.unsw.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2008.923854

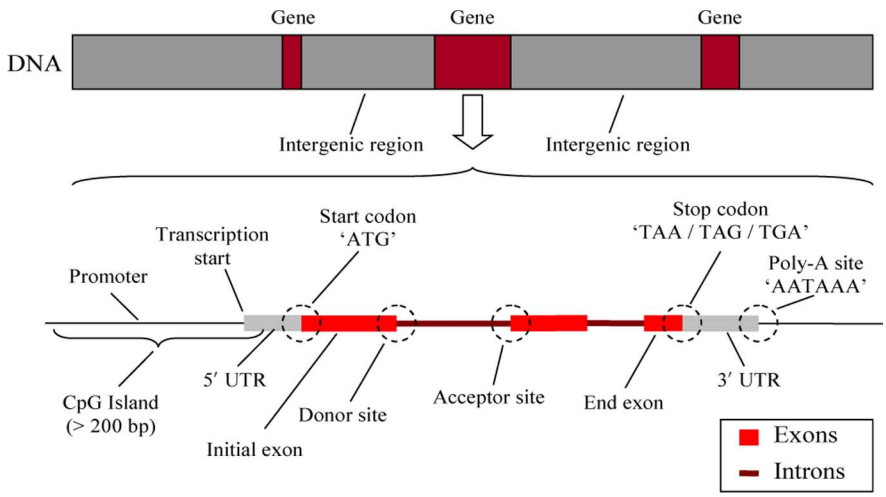

Fig. 1. Eukaryotic DNA consists of genic and intergenic regions. Donor and acceptor sites (i.e., $5'$ and $3'$ ends of all introns) are used to splice exons on both sides of an intron in a process known as *splicing*.

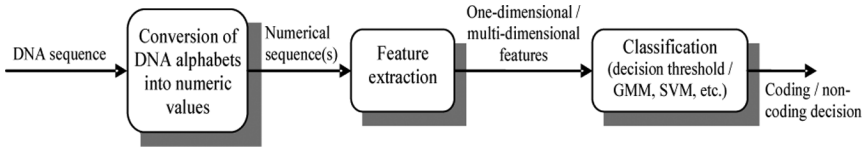

Fig. 2. DSP techniques are applied to DNA sequences following conversion into numerical signals, to extract features. The classification results can be used to provide accurate detection of exonic end-points (acceptor/donor splice sites).

whereas multidimensional classification can be achieved using well-known pattern recognition tools such as Gaussian mixture models (GMMs) or support vector machines (SVMs).

To our knowledge, despite the signal processing research activity in this area, no comparisons with well-established existing data-driven methods for eukaryotic gene prediction (e.g., GENSCAN) are available. We address this shortcoming herein, in addition to proposing newly developed DNA symbolic-to-numeric mappings and gene prediction features. It is our belief that a system combining improved DSP techniques with existing data-driven methods could offer a level of gene prediction accuracy higher than that offered by existing data-driven methods.

This paper is organized as follows. Section II reviews existing methods for DNA numerical representation and DSP-based methods for gene and exon feature extraction from the DNA sequence. In Section III, newly developed DNA representations and gene prediction features are discussed. Selected existing statistical approaches and a proposed DSP-statistical combination for acceptor splice site detection are presented in Section IV. Existing and newly proposed DNA representations, DSP-based gene and exon prediction features, and acceptor splice site detection methods are then compared using standard datasets and evaluation measures, as explained in Section V. Evaluation results and discussion are given in Section VI.

II. DNA Numerical Representation and Feature Extraction

A. DNA Numerical Representations

In recent years, a number of schemes have been introduced to map DNA nucleotides into numerical values. Some possible desirable properties of a DNA numerical representation include: 1) each nucleotide has equal "weight" (e.g., magnitude), since there is no biological evidence to suggest that one is more "important" than another; 2) distances between all pairs of nucleotides should be equal, since there is no biological evidence to suggest that any pair is "closer" than another; 3) representations should be compact, in particular, redundancy should be minimized; and 4) representations should allow access to a range of mathematical analysis tools.

The binary or Voss representation [20] is currently the most popular scheme, which maps the nucleotides A, C, G, and T into the four binary indicator sequences $x_A[n]$, $x_C[n]$, $x_G[n]$, and $x_T[n]$ showing the presence (e.g., 1) or absence (e.g., 0) of the respective nucleotides. Both the Z-curve [21] and tetrahedron [22] methods reduce the number of indicator sequences from four to three in a manner symmetric to all four components. The Voss and tetrahedron representations have been shown to be equivalent representations for the purpose of power spectra computation [28]. The complex representation [18], [23] reflects some of the complementary features of the nucleotides in its mathematical properties. As an alternative to the typical complex representation of DNA nucleotides, certain complex weights for each of the four bases can also be calculated and employed with the binary indicator sequences [18]. In the quaternion representation of DNA symbols [24], pure quaternions are assigned to each symbol. It has been conjectured that the quaternion approach can improve DNA pattern detection in the spectral domain through use of the quaternionic Fourier transform [24]. In the EIIP (electron-ion interaction potential) method [25], the electron-ion interaction potential (related to the quasi-valence number) associated with

each nucleotide is used to map DNA character strings into numerical sequences. The EIIP is just one example of a real number representation. Another can be obtained by attaching digits $\{0,1,2,3\}$ to the four nucleotides: $T=0, C=1, A=2$, and $G=3$ [23]. However, this structure implies that purines (A or G) are in some respect "greater than" pyrimidines (C or T). Similarly, the representation $A=0$, $C=1$, $T=2$, and $G=3$ suggests that $T>A$ and $G>C$. This representation is an example of a Galois field assignment, upon which symbolic Galois field operations are possible [26]. Another real-number representation maps $A=1.5$, $T=-1.5$, $C=0.5$, and $G=-0.5$, similar to the complementary property of the complex method. These assignments of real numbers to each of the four DNA characters do not necessarily reflect the structure present in the original DNA sequences. Alternatively, to calculate weights representing the actual participation of each symbol in the detected pattern, a linear transform and optimization can be performed on the DNA sequences [29]. The internucleotide distance method [27] replaces each DNA nucleotide with an integer representing the distance between the current nucleotide and the next similar nucleotide.

Each of the existing DNA representations offer different properties, and map the DNA sequences into between one and four numerical sequences. Many existing methods, such as Voss [20], Z-curve [21], and tetrahedron [22], map the DNA sequence to three or four numerical sequences, potentially introducing redundancy in the representation. The assignment of arbitrary numbers to each of the four DNA characters in EIIP [25] and other real number representations [23], [26] does not necessarily reflect the structure present in the original DNA sequence. Representations such as quaternions [24] are limited to specific mathematical analysis tools. For example, a discrete quaternion Fourier transform (DQFT) [41] based spectral analysis is required to detect certain DNA patterns. Furthermore, existing DNA representations do not fully exploit the structural differences of protein coding and noncoding regions to facilitate digital signal processing based gene and exon prediction features. These issues are addressed in the DNA representations proposed in Section III-A.

B. DSP-Based Features for Gene and Exon Prediction

Periodicities of 3, 10.5, 200, and 400 have been reported in genomic sequences [30]. In exons, the occurrence of identical nucleotides in identical codon positions is the basis for a periodicity of three interpretation in these regions [31]. The period-3 behavior of exons has been widely used to identify these regions using DSP-based methods, following conversion to numerical sequences.

The discrete Fourier transform (DFT), the most commonly used method for spectrum analysis of a finite-length numerical sequence $x[n]$ of length N, is defined as [36]

$$X[k]=\sum_{n=0}^{N-1} x[n]e^{-j(2\pi nk/N)}, \quad 0\le k\le N-1. \tag{1}$$

Equation (1) can be used to calculate DFTs for numerical sequences representing DNA sequence portions, for example each of the four binary indicator sequences (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$). The periodicity of 3 in exon regions of a DNA sequence suggests that the DFT coefficient corresponding to $k=N/3$ (where N is chosen to be a multiple of 3) in each DFT sequence should be large [18]. Note that the calculation of the DFT at a single frequency ($k=N/3$) is sufficient, so that the Goertzel algorithm [37], which reduces the cost of single point DFT computation by almost a factor of two, can be employed. Various DFT based spectral measures exploiting the period-3 behaviour of exons for the identification of these regions have been proposed. The spectral content (SC) measure [32] combines the individual DFTs (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$) to obtain a total Fourier magnitude spectrum of the DNA sequence, as follows:

$$S[k]=\sum_m |X_m[k]|^2, \quad m=\{A,C,G,T\}. \tag{2}$$

The GeneScan program [32], based on the SC measure, computes the signal-to-noise ratio of the peak at $k=N/3$ as $P=S[N/3]/\hat{S}$, where $\hat{S}$ represents a longer-term average of the spectral content defined in (2). Regions having $P\ge 4$ are assumed to be protein coding (exon) regions. The optimized SC measure [18] assigns complex weights a, c, g, and t to each of the four DFTs $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$ in (2). These weights are calculated using an optimization technique applied to the known genes of a given organism. However, one can also apply complex conjugate pairs $t=a^*$ and $g=c^*$. The spectral rotation (SR) measure [33] rotates four DFT vectors $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$ clockwise, each by an angle equivalent to the average phase angle value in coding regions, to make all of them "point" in the same direction. The SR measure also divides each term by the corresponding phase angle deviations to give more weight to exonic distributions. The feature

$$SR[k]=\left|\sum_m \frac{e^{-j\mu_m}}{\sigma_m}X_m[k]\right|^2, \quad m=\{A,C,G,T\} \tag{3}$$

where μ_m and σ_m are the means and standard deviations of the phase angle value in coding regions, has been used for the detection of exons, and was shown to give better performance than the SC (2) measure at a 10% false positive gene detection rate [33]. Note that all DFT-based techniques suffer from spectral leakage, due to the finite-length analysis window, which introduces small contributions from signal frequencies other than those at the frequency $(2\pi k/N)$.

Autoregressive (AR) methods provide an alternative, more compact characterization of the signal spectrum. Particular advantages of the AR technique are that it requires relatively few base pairs to calculate the AR model (which is convenient if the exon regions are short and/or closely spaced), and that it provides a compact estimate of the signal spectrum. It has been shown that AR spectral estimation using the Burg algorithm and improved covariance analysis performs better than the DFT for the detection of period-3 behaviour in short genomic sequences [34]. However, the selection of the model order is crucial in this approach, since choosing p too low or too high will result in unnecessarily smoothed or spurious modeling of spectral peaks, respectively. Note also that AR models cannot reasonably be applied to binary indicator sequences, since these could not have

resulted from an AR process. A possible solution to the problem is bandpass filtering of numerical sequences before their application to AR modeling.

In order to reduce the spectral leakage present in DFT-based exon prediction, a larger window size is required, which implies longer computation time and also compromises the base-domain resolution. The infinite impulse response (IIR) antinotch (AN) filter approach in [35] attempts to address these problems. The magnitude response of the antinotch filter has a sharp peak at $\theta = 2\pi/3$, which preserves only period-3 components. The individual outputs of the four binary indicator sequences can be combined in a sum-of-squares manner. It has been shown in [35] that digital filter based period-3 detection results are comparable to those of the DFT-based SC measure given by (2).

The autocorrelation function (ACF) is a measure of how well a signal matches a time-shifted version of itself, as a function of the time shift. Practically, the ACF will produce a peak if significant correlation exists at $k = 3$. Besides period-3 detection, DNA sequences have been widely analyzed for other correlations in [20], [31], [38], [39]. Li [40] gives a critical review of the study of correlation structures.

The identification of protein coding regions is difficult mainly due to the noncontiguous and noncontinuous nature of eukaryotic genes. Despite the existence of many DSP-based approaches and also data-driven approaches, the accuracy of exon prediction is still limited and needs to be improved. Moreover, the existing approaches only rely on the identification of period-3 property of exons, and do not fully exploit efficient DNA representations and other complementary features required to separate protein coding and noncoding nucleotides. These problems have been addressed in new DSP-based gene and exon prediction features, proposed in Section III-B.

III. New DNA Representations and DSP-Based Features for Gene and Exon Prediction

A. DNA Numerical Representations

1) Paired Numeric: The paired numeric representation for gene and exon prediction [42] exploits one of the differential properties of exons and introns, according to which introns are rich in nucleotides "A" and "T" whereas exons are rich in nucleotides "C" and "G" [43]. Furthermore, the DFT phase angle histogram distributions for coding and noncoding regions of human datasets have been shown to give smaller and nearly equal values of angular mean for distributions of nucleotides "C" and "G" than those of "A" and "T" [42]. To fully exploit both of these properties, these nucleotides $(A\text{-}T, C\text{-}G)$ can be paired in a complementary manner and values of $+1$ and -1 can be used to denote $A\text{-}T$ and $C\text{-}G$ nucleotide pairs, respectively. A similar approach was used by Datta and Asif [44]; however, no motivation for the "$A\text{-}T$" and "$C\text{-}G$" pairing was given. "$A\text{-}T$" and "$C\text{-}G$" are complementary pairs, joining opposite strands of double helix DNA through hydrogen bonds. However, this is not the reason for their pairing here, as only one strand is used for the computational analysis of DNA. This representation incorporates a very useful DNA structural property, in addition to reducing complexity.

2) Frequency of Nucleotide Occurrence: It has been shown in [42] that the four DNA nucleotides differ considerably in their occurrence in exonic regions, and that the fractional occurrence of any particular nucleotide is reasonably consistent across the Burset/Guigo1996 [10], HMR195 [11], and GENSCAN learning [45] datasets considered therein. It has been further observed that the frequency of nucleotide occurrence in exons is a key parameter for any DNA representation to be used for the detection of these regions. According to the frequency of nucleotide mapping [42], nucleotides are represented by their fractional occurrences in exons of a training database.

B. DSP-Based Features for Gene and Exon Prediction

1) Paired and Weighted Spectral Rotation (PWSR) Measure: The PWSR measure [46] incorporates a statistical property of eukaryotic sequences, according to which introns are rich in the nucleotides "A" and "T" whereas exons are rich in nucleotides "C" and "G." This information leads to an alternative property to the well-known period-3 behavior of exons. In this method, the DNA sequences are first converted into two binary indicators, $x_{A\text{-}T}[n]$ and $x_{C\text{-}G}[n]$. Using training data from DNA sequences of the same organism, the means μ_m and standard deviations σ_m of the distributions of DFT phase angle averaged over coding regions, i.e., one phase angle value per coding region, are calculated. Weights w_m based on the frequency of occurrence of nucleotides "A or T" and "C or G" in coding regions of the training data are also calculated. The expression given in (4) can then be used as a feature, along one direction of the DNA sequence

$$\mathrm{PWSR}_l[k] = \left| \frac{e^{-j\mu_{A\text{-}T}}}{\sigma_{A\text{-}T}} \cdot w_{A\text{-}T} \cdot X_{A\text{-}T}[k] + \frac{e^{-j\mu_{C\text{-}G}}}{\sigma_{C\text{-}G}} \cdot w_{C\text{-}G} \cdot X_{C\text{-}G}[k] \right|^2 \quad (4)$$

where l denotes the forward (F) and reverse (R) directions of DNA sequence, μ_m and $\sigma_m (m = A\text{-}T, C\text{-}G)$ are the mean and standard deviation values obtained from distributions of the DFT phase angle averaged over coding regions of the training data, w_m are frequency of occurrence weights from training data, and $X_m[k]$ are the sliding DFT windows of two indicator sequences. The PWSR is calculated in both directions of the DNA sequence, and combined as

$$\mathrm{PWSR}[k] = \mathrm{PWSR}_F[k] + \mathrm{PWSR}_R[k]. \quad (5)$$

Note that due to paired indicators, a DFT in the reverse direction of the same DNA strand is equivalent to a DFT on its complementary strand.

2) Paired Spectral Content (PSC) Measure: The PWSR measure is a data-driven frequency domain method for gene and exon prediction, which requires training from DNA sequences of the same organism. A more general method, known as the paired spectral content (PSC) measure [47], first converts the DNA sequence into single numerical sequences using the paired-numeric representation, as discussed in Section III-A-I, then combines forward and backward DFTs on the same DNA sequence

$$PSC[k] = |X_F[k]|^2 + |X_R[k]|^2 \quad (6)$$

where $X_F[k]$ and $X_R[k]$ are DFTs of the indicator sequence $x[n]$ in the forward and reverse directions. Contrary to the SR and PWSR measures, the PSC measure can be applied to the sequences taken from any organism, i.e., PSC is not an organism-specific measure.

3) "Time-Domain" Algorithms: In the following approaches [48], DNA sequences are first converted into Voss indicator sequences, which are passed through a second-order resonant filter with a center frequency of $2\pi/3$ (similar to [35]) before being input to either algorithm. This prefiltering helps to remove spectral components at $2\pi k/3$, $k \in \Im$, $k \neq 1$, which arise from the application of correlation-based approaches to a binary indicator sequence at a base-domain lag of 3.

Average magnitude difference function (AMDF)—the average magnitude difference function (AMDF) has long been used in speech processing, and is defined for a discrete signal $x[n]$ as a function of the period k as [49]

$$\mathrm{AMDF}\,[k] = \frac{1}{N}\sum_{n=1}^{N} |x\,[n] - x\,[n-k]| \tag{7}$$

where N is the window length. The AMDF, with $k = 3$, is an efficient time-domain algorithm for gene and exon prediction [48]. Practically, the AMDF will produce a deep null if significant correlation exists at period $k = 3$.

Time domain periodogram (TDP)—the time domain periodogram (TDP) is an algorithm used for periodicity detection in sunspots and pitch detection for speech processing [50]. According to this algorithm, the N-point data are first arranged in a matrix, with N/k rows containing subsequences of length equal to the period (k) being tested, where N is the window length. The columns of the matrix are summed to obtain the TDP vector of size k, as follows:

$$\mathrm{TDP}_{\mathrm{vector}}[k] = \left\{ \sum_{n=1}^{N/k} x[kn-(k-1)], \sum_{n=1}^{N/k} x[kn-(k-2)], \ldots \sum_{n=1}^{N/k} x[kn] \right\}. \tag{8}$$

The final estimate of the degree of periodicity at period k is derived as follows:

$$\mathrm{TDP}[k] = \max(\mathrm{TDP}_{\mathrm{vector}}[k]) - \min(\mathrm{TDP}_{\mathrm{vector}}[k]). \tag{9}$$

It has been shown in [50] that for large N, $TDP[k]$ has a very sharp peak if correlation exists at period k, enabling accurate detection of periodicity.

4) Singular Value Decomposition (SVD): The singular value decomposition (SVD) can be applied to a rectangular data matrix A, decomposing it into matrices U, S, and V [51], as follows:

$$A = U\,S\,V^T \tag{10}$$

where $U^TU = I$, and $V^TV = I$, i.e., U and V are orthogonal. The singular values of S are square roots of the eigenvalues from AA^T or A^TA, where here A comprises the frames of numeric DNA sequence values organized into a $k \times p$ rectangular matrix, where we choose $k = 3$. A linear combination of the largest singular values of all frames obtained using all four binary indicator sequences can then be used for the coding/non-coding decision. SVD-based period-3 detection has also been enhanced using bandpass filtering of the individual binary indicator sequences, emphasizing the period-3 behavior [52].

5) Time-Frequency Hybrid (TFH) Measure: The time-frequency hybrid (TFH) measure [46] combines magnitude and phase-based features, acknowledging earlier results by Kotlar and Lavner [33] showing that additionally considering the DFT phase angle is more informative than the magnitude alone. Features from the time-domain AMDF method and frequency-domain PWSR measure are normalized to the range [0, 1] and then summed to produce a TFH feature.

6) Multidimensional Features: For all existing methods and new methods discussed in Sections III-B1–V, features are combined, typically using a sum-of-squares approach as in (2), to produce a 1-D feature for comparison with some predetermined threshold. Since a sum-of-squares approach will not necessarily produce optimal feature fusion, multidimensional features have been recently proposed in [53] and [54]. One such scheme uses the PWSR and AMDF features from the TFH measure in [46], and transforms each separately using the DCT, to decorrelate them with energy localized to the first few coefficients. A 5-D feature, comprising one transformed PWSR coefficient and all four transformed AMDF coefficients, is then formed. In another such scheme, the linear predictor coefficients are treated as multidimensional AR-based features, modeling coding and non-coding regions in terms of their spectral characteristics within the given window length. The optimized AR based feature, with model order 12 and window size 180 bp is then concatenated with the 5-D TFH feature, as shown in [54]. The resultant higher dimensional feature set is then used for training and testing of the multidimensional feature based classification system.

IV. Acceptor Splice Site Detection Methods

The accurate prediction of eukaryotic protein coding regions requires methods for the detection of their end-points. The intron-exon border is known as the acceptor splice site (or $3'$ end of the intron) and consists of a consensus dinucleotide "*AG*" as the last two nucleotides of the intron. Due to the common occurrences of this dinucleotide at locations other than acceptor sites throughout a gene sequence, detection is very difficult. In order to apply data-driven and other methods herein, the candidate acceptor site sequences were extracted as windows of 140 nucleotides around each consensus dinucleotide "*AG*," similarly to [55]. The nucleotide positions were then labeled relative to the consensus dinucleotide "*AG*," which was assumed to occupy the positions -2 and -1. From the $5'$ end to the $3'$ end of a genomic sequence, these labels would be: $-70, -69, \ldots, -2, -1, +1, +2, \ldots, +69, +70$. In the case of a true acceptor splice site, the first 70 positions represent intronic nucleotides, while the last 70 labels can be treated as exonic nucleotides, as shown in Fig. 3.

A. Existing Methods

Weight matrix method (WMM)—this method assumes that the probabilities of the nucleotides at each position are independent of each other [56]. According to [57], the probabilities of

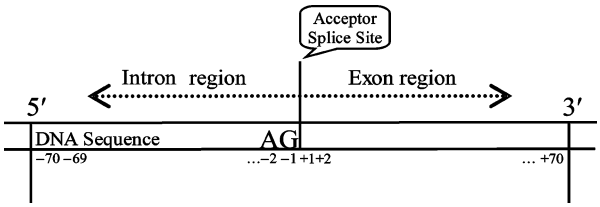

Fig. 3. Looking from the $5'$ end of DNA to its $3'$ end, the acceptor splice site is essentially an intron-to-exon border and consists of a consensus dinucleotide "*AG*" as the last two nucleotides of the intron.

generating a signal S of length N under positive and negative WMMs of the pyrimidine-rich acceptor region $[-20,+3]$ are

$$\mathrm{WMM}\{S\} = \prod_{k=1}^{N} p_{s_k}^{k} \tag{11}$$

where p_m^k is the probability of generating nucleotide m at position k of the signal, which can be estimated from the positional frequencies of nucleotides in the training sets of the true and false acceptor site sequences. Therefore, the positive and negative probabilistic models correspond to learning using true and false acceptor site sequences, respectively. The log of the ratio of the WMM generated under a positive model to the WMM generated under a negative model can be used as a score to discriminate true acceptor splice sites from false.

Weight array method (WAM)—the WAM explores and captures the dependencies between adjacent positions, in contrast to the WMM, which considers each position independently [58], [59]. In [57], probabilities of generating a signal S of length N under positive and negative Weight Array Models of the pyrimidine-rich acceptor region $[-20,+3]$ are computed as

$$\mathrm{WAM}\{S\} = p_{s_1}^{1} \prod_{k=2}^{N} p_{s_{k-1},s_k}^{k-1,k} \tag{12}$$

where $p_{m,n}^{k-1,k}$ is the conditional probability of generating nucleotide n at position k, given nucleotide m at position $k-1$. This quantity can be calculated from the ratio of the frequency of dinucleotides m and n at positions $k-1$ and k, to the frequency of the nucleotide m at position $k-1$.

Windowed weight array method (WWAM)—the WWAM is a second-order WAM model, in which nucleotides of the branch point region $[-38,-21]$ are generated conditional on the nucleotides of the previous two positions [45]. In order to have enough data to model these second-order conditional probabilities, data from a window of adjacent signal positions are pooled. Here, the second-order conditional probability at position k is calculated as the average of the conditional probabilities at positions $k-2$, $k-1$, k, $k+1$, and $k+2$. The WWAM has been combined with the WAM over the region $[-21,+3]$ to compute signal ratio scores for acceptor site recognition in GENSCAN [6], [45].

B. Proposed DSP-Statistical Hybrid Approach

Since the exon region starts from the next nucleotide to the consensus dinucleotide "*AG*" of the true acceptor sites in the $3'$ direction, the detection of the presence or absence of period-3 behavior, as determined by signal processing-based methods, in this region of the candidate acceptor sites can be used to discriminate the true sites from their false counterparts. For this purpose, we employ the AMDF method in conjunction with the "paired numeric" DNA symbolic-to-numeric mapping scheme using the forward-backward window attribute, similar to [42]. These methods were selected based on experimental work from Sections VI-A and B. Furthermore, due to the possibility of a very small exonic region in candidate acceptor sites (e.g., 70 bp) a larger window is inadvisable. With a smaller window (69 base pairs for the AMDF), a score "S" for each candidate acceptor site based on the ratio of the sum of period-3 features (denoted here as $P3$) in putative coding regions to that of putative noncoding regions

$$S = \log_2 \left(\frac{\sum_{\frac{L_c}{U_c}} P3}{\sum_{\frac{L_{nc}}{U_{nc}}} P3} \right) \tag{13}$$

is proposed to discriminate the true and false acceptor sites, where U and L are, respectively, the upper and lower indices for the period-3 summations in the putative coding (denoted c) and noncoding (denoted nc) regions. DSP-based methods are attractive because they mostly do not require any training on the genomic data before use, unlike the WMM, WAM, and WWAM approaches, and also because they are derived from different information from these approaches. Hence, we also combine the DSP-based method with WAM to improve the discrimination power of acceptor splice site detection. Empirically, we found the WAM model of the region $[-26,+28]$, and DSP-based method over the region $[-20,-1]\cup[+1,+20]$ to be optimum for the recognition of human acceptor splice sites.

V. Evaluation

In this section, the evaluations of different DNA representations, feature extraction methods, and acceptor splice site detection methods (reviewed in Sections II–IV) are described, using standard eukaryotic datasets.

A. Data Sets

Table I summarizes the Burset/Guigo1996 [10], HMR195 [11], and GENSCAN [45] standard datasets, used herein. During the original formation of these datasets, certain common rules were followed. For example, each sequence consists of one complete gene starting with an "ATG" codon and ending with one of the three possible stop codons (TAA, TAG, or TGA). The protein coding genes do not have any in-frame stop codons. Moreover, the multiexon genes have "GT" and "AG" consensus dinucleotides for the donor and acceptor splice sites, respectively.

The data sets referred to, respectively, as the GENSCAN learning and test sets comprise the 188 multiexon gene sequences listed in [45, Appendix A] and 64 available multiexon gene sequences listed in [45, Appendix B]. The number of true/false acceptor site sequences of GENSCAN learning and test sets were 1031/156107 and 317/44301, respectively, when extracted from windows of 140 nucleotides around each consensus dinucleotide "AG."

TABLE I
SUMMARY OF DATASETS

Dataset	Organisms	# gene sequence	# bp	# exon	Coding density (%)
Burset/Guigo1996 [10]	Vertebrate	570	2,892,149	2649	15.37
HMR195 [11]	Mammalian	195	1,383,720	948	14
GENSCAN Learning Set [45]	Human	380	2,581,000	1492	16
GENSCAN Test Set [45]	Human	65	591,886	381	10.2

B. System Configurations

For the comparison of DNA symbolic to numeric mappings, evaluated on the exon detection problem, the GENSCAN datasets (learning and test sets) were used for training (where needed) and testing of DNA representations. The GENSCAN test set was mapped into all DNA representations, and the DFT-based SC measure was then applied for gene and exon prediction in each case. A constant length rectangular window ($N = 351$ [18]) was used for all types of DFT calculations. For the quaternion representation, the discrete quaternion Fourier transform (DQFT) [41] was employed to calculate the SC measure.

The second evaluation compared the various DSP-based exon detection methods discussed in Sections II and III. The Burset/Guigo1996, HMR195, and GENSCAN datasets were all used. Note that the SR, PWSR, and TFH methods are organism-specific and can only be trained and tested on datasets consisting of gene sequences taken from one particular organism, such as GENSCAN in our case. In 1-D feature extraction, a rectangular window of constant size $N = 351$ (consistent with previous work [18], [32], [33]) was again used in DFT-based methods. The AMDF, TDP, and SVD methods were prefiltered with a bandpass filter tuned at $2\pi/3$, to emphasize the period-3 component and de-emphasize all other components. In AR model implementation, a model order of 40 and frame size of 135 were used, as were found suitable in the preliminary work of [52]. A frame size of 81 was used for the SVD method, similar to [52]. Empirically, we found a frame length of 117 suitable for AMDF and TDP. A frame size of 117 was used for ACF, consistent with the frame sizes for the other time-domain algorithms. In multidimensional feature extraction, a constant window size of $N = 351$ was used for DFT-related features and frame size of 117 was used for AMDF calculations. For AR modeling, a model order of 12 and window length of 180 were used, as determined in [54].

For exon prediction using multidimensional features, two GMMs were trained, based on protein coding and noncoding features of the GENSCAN learning set, respectively, from which likelihood estimates were extracted as features during testing on the GENSCAN test set, as explained in [54]. Empirically, we found 32 mixtures optimal for training the GMM parameters, and a diagonal covariance matrix was used.

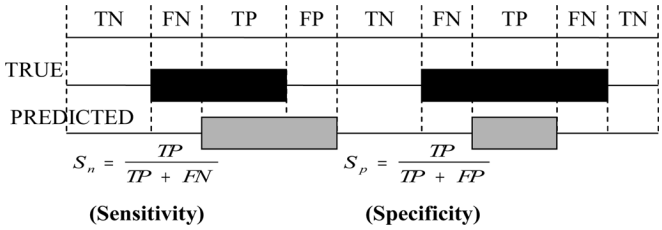

Fig. 4. Nucleotide level measures of prediction accuracy.

In the evaluation of acceptor splice site detection, true and false acceptor sites from the GENSCAN datasets (learning and test sets) were used for training and testing of WMM, WAM, WWAM, and the proposed DSP-based method.

Note that in all cases we do not actually perform the classification to derive an exon/intron decision. Instead, we take advantage of the fact that this is a 2-class classification problem and give results across a range of different threshold settings/decision rules.

C. Evaluation Measures

In these evaluations, results are compared at the nucleotide level, contrary to existing comparisons at exon level or gene level, e.g., [33]. In exon-level detection, the feature value for one point (i.e., nucleotide) in an exon being greater than a decision threshold is sufficient for the detection of that particular exon. The following measures were employed.

Sensitivity and Specificity—The prediction accuracy measures of sensitivity, specificity (similar to [10]) can be explained with the aid of Fig. 4, where 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. The sensitivity (S_n) gives the measure of the proportion of coding nucleotides that have been correctly predicted as coding. The specificity (S_p) is the proportion of predicted coding nucleotides that are actually from the coding region.

Receiver operating characteristic (ROC) curves—The receiver operating characteristic (ROC) curves were developed in the 1950s as a technique for visualizing, organizing and selecting classifiers based on their performance [60]. In the exon-intron separation problem, an ROC curve explores the effects on *TP* and *FP* as the position of an arbitrary decision threshold is varied. The curve can be characterized as a single number using the area under the ROC curve (AUC), with larger areas indicating more accurate detection methods.

False positive (and specificity) versus sensitivity—in this measure, the percentage of false positives and percentage specificity are calculated at different levels of percentage sensitivity. A threshold output feature value *Th* at a particular level of percentage sensitivity s is the minimum value for which $s\%$ of the exonic nucleotides have feature values greater than *Th* [45].

Detection of exonic nucleotides for $p\%$ false positive—the percentage of exonic nucleotides detected for $p\%$ false positives (where $p = 10$, 20, and 30) can also be calculated, generating curves when the decision threshold is varied. False positives are

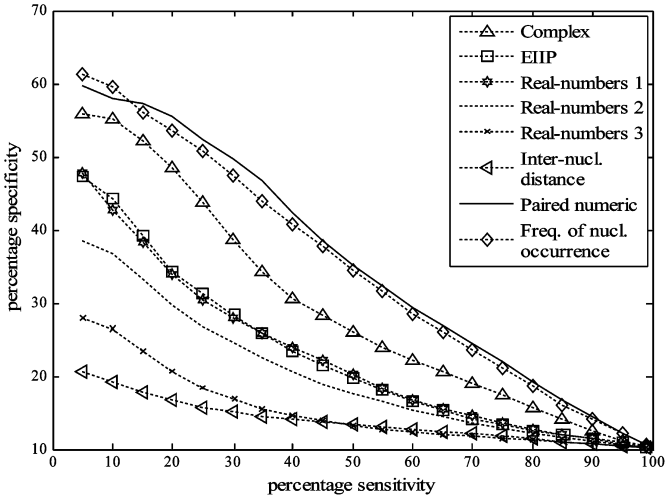

Fig. 5. Comparison of single-sequence DNA representations for the exon detection problem, evaluated on the GENSCAN test set.

TABLE II
SUMMARY OF RESULTS FOR SC-BASED EXON PREDICTION USING GENSCAN TEST SET

DNA symbolic-to-numeric method	Area under ROC curve (AUC) for DFT-based SC measure			
	Single Sequence		Three Sequence (Forward)	Four Sequence (Forward)
	Forward DFT	Forward & backward DFTs		
Voss	—	—	—	0.7778
Z-curve	—	—	0.7777	—
Tetrahedron	—	—	**0.7778**	—
Complex	0.7397	0.7756	—	0.7778
Pure quaternion	0.7751	0.7450	—	0.7778
Complex quaternion	0.7751	0.7449	—	0.7778
EIIP	0.6757	0.6951	—	0.7778
Real numbers 1	0.6782	0.6988	0.7481	—
Real numbers 2	0.6543	0.6787	0.7617	—
Real numbers 3	0.5940	0.6069	—	
Inter-nucl. Distance	0.5923	0.5892	—	0.6948
Paired numeric	**0.7916**	**0.8114**	—	—
Freq. of nucl. occurrence	0.7857	0.8068	—	0.7747

an important problem due to the fact that intronic and intergenic nucleotides make up more than 95% of the eukaryotic genome.

VI. RESULTS AND DISCUSSION

A. DNA Representation Results

All DNA representations were compared for the purpose of identifying protein coding regions, using the DFT based SC measure to characterize period-3 behavior in the exonic regions for the GENSCAN test set. Specificity versus sensitivity and AUC results for single-sequence representations are given in Fig. 5 and Table II, respectively. Note that "real-numbers 1" refers to $T = 0, C = 1, A = 2, G = 3$; "real-numbers 2" is $A = 0, G = 1, C = 2, T = 3$; and "real-numbers 3" is $A = 1.5$, $T = -1.5, C = 0.5, G = -0.5$. These results show that the recently proposed "paired numeric" is the most accurate representation for this application, due to the fact that the approach exploits a key statistical property (according to which introns are rich in nucleotides "A" and "T" whereas exons are rich in nucleotides "C" and "G") for discriminating between structures of the genomic protein coding and noncoding regions. The frequency of nucleotide occurrence method also gives promising results, due to the fact that fractional occurrences of the four nucleotides in protein coding regions is different to those of the noncoding regions. It is perhaps surprising that the paired numeric and frequency of nucleotide occurrence representations provide such a marked improvement over other real number representations, suggesting that real number mappings need to be selected very carefully for a given application. It is also perhaps surprising that the paired numeric and frequency of nucleotide occurrence representations, which exhibit few of the conceptually desirable properties of DNA representations mentioned in Section II-A, are the most successful.

Table II shows that the Z-curve and Tetrahedron schemes are approximately equal, and give improved gene and exon prediction than real number approaches for a complexity equivalent to three sequences. Compared with the paired numeric and frequency of nucleotide representations, however, their higher dimension does not produce gains in detection accuracy.

The four-sequence representations: Voss, complex, quaternion, and EIIP, all give equivalent DFT-based gene and exon prediction accuracy, as seen in Table II. This result was expected as they are all variations on the same representation. Furthermore, their performance is equal to the three-sequence tetrahedron representation, which is also a variation of this representation. It has also recently been shown in [61] that the three-sequence Z-curve method is theoretically equivalent to the Voss approach. By comparison with four-sequence schemes, the recently proposed paired numeric and frequency of nucleotide occurrence methods reveal improved DFT-based gene and exon prediction with 75% less downstream processing. We conjecture that further improvements in gene and exon prediction can be achieved by incorporating more DNA structural properties in existing or new DNA symbolic-to-numeric representations.

Finally, we note that small improvements in detection accuracy can be gained through the use of forward and backward sliding window DFTs, at the cost of increased complexity.

B. Period-3 Exon Detection Results

The results and discussion presented in this section benchmark the period-3 methods for gene and exon prediction in eukaryotes on a large scale, using three standard genomic sequence datasets. ROC plots using Burset/Guigo1996, HMR195 and GENSCAN test set are shown, respectively, in Figs. 6–8. Table III summarizes all period-3 detection results giving AUC values, and exonic nucleotide detection rates for $p\%$ false positive using all three datasets. It can be observed that for the vertebrate dataset, the recently proposed AMDF and TDP outperform other measures, giving consistently improved exonic detection. Time-domain methods are attractive because they are computationally efficient and perform better for an identification of short and/or closely spaced coding regions, using smaller window lengths [48]. Furthermore, the recently proposed frequency-domain DFT based PSC measure improves on the SC measure, with 50% less DFT processing. Interestingly, the ACF, AR, and AN filter methods give very poor

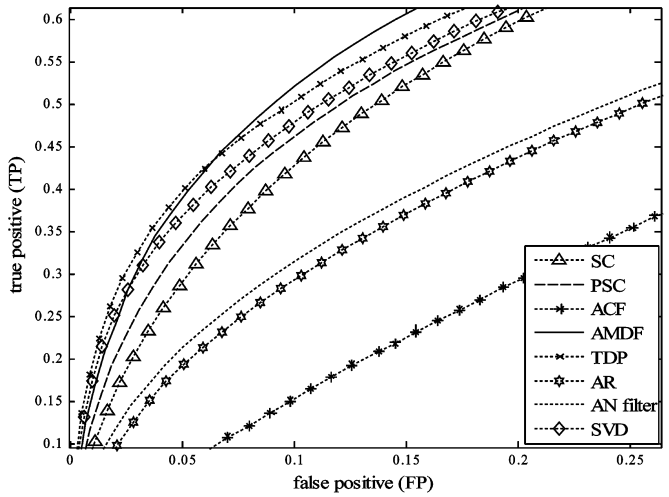

Fig. 6. ROC curves for period-3 detection, using the Burset/Guigo1996 dataset.

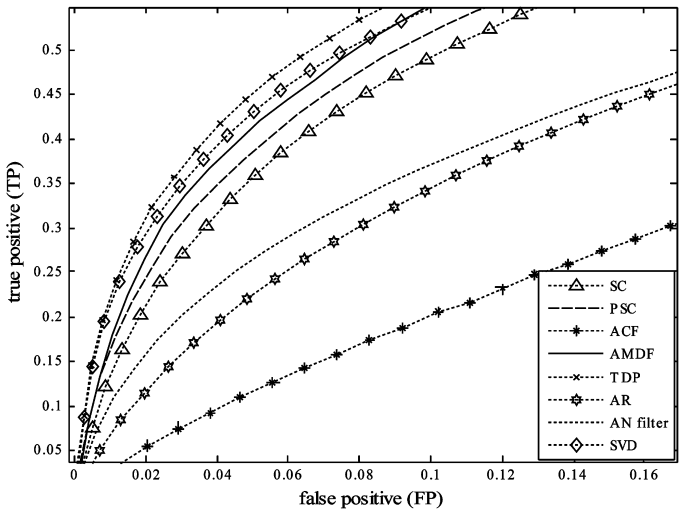

Fig. 7. ROC curves for period-3 detection, using the HMR195 dataset.

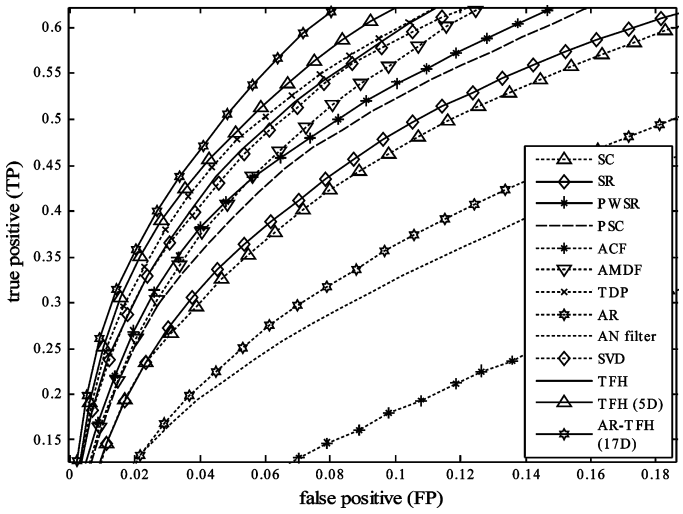

Fig. 8. ROC curves for period-3 detection, using the GENSCAN test set.

identification of period-3 regions, and a likely cause is the lack of bandpass filtering as used in the AMDF and TDP methods, as discussed in Section III-B3. In results using the HMR195 dataset, the AMDF and TDP also give improved performance compared with other methods, similar to the results obtained using Burset/Guigo1996 dataset.

Since the Burset/Guigo1996 (vertebrate) and HMR195 (mammalian) datasets contain mixed genomic sequences (i.e., sequences taken from different organisms), the SR, PWSR, and TFH methods, which require training data, can not be applied to these datasets in a straightforward manner. Hence, for comparison between all methods, the GENSCAN test set was employed. It is quite clear from the results in Fig. 8 that the data-driven, DFT-based PWSR measure outperforms well-known 1-D frequency-domain methods, giving consistently fewer false positives (and higher levels of specificity) at each sensitivity level and improved nucleotide detection. By comparison with other DFT-based measures, the PWSR method reveals relative improvements of 15.2% and 10.7%, respectively, over the SC and SR measures in the detection of exonic nucleotides at a 10% false positive rate. The recently proposed paired spectral content (PSC) method also improves on the SC and SR measures. One reason for DFT-based methods (i.e., SC, SR, PWSR, and PSC measures) giving poorer accuracy than time-domain algorithms (e.g., AMDF, and TDP), is their relatively large window size (351). Recent investigations [62] suggest that the optimum window length for DFT-based methods depends to a large extent on the average length of exon regions of the dataset being used, whereas for time-domain algorithms, this length lies within a short range. The time-frequency hybrid (TFH), which combines the complementary PWSR and AMDF methods, provides a further small gain in accuracy over the individual PWSR and AMDF methods.

Finally, the multidimensional feature-based methods give more accurate gene and exon prediction than all 1-D methods. By comparison with the best 1-D method (TFH), the recently proposed multidimensional TFH and AR-TFH methods reveal relative improvements of 4.7% and 11.4%, respectively, in the detection of exonic nucleotides at a 10% false positive rate.

C. Acceptor Splice Site Results

After training on the GENSCAN learning set, the proposed DSP-statistical hybrid acceptor site detection method from Section IV-B was compared with WMM, WAM and WWAM using the GENSCAN test set. Fig. 9 shows ROC curves for all methods using the GENSCAN test set, from which it can be observed that the ROC curve for the proposed method exhibits better discrimination power for the detection of acceptor splice sites. The DSP-based method alone is notably poorer than data-driven methods, presumably due to the fact that it relies solely on the accurate identification of the period-3 behaviour on one side of the "*AG*" junction (i.e., consensus dinucleotide for acceptor sites). The periodicity of three in exons is often weak, and existing DSP-based methods are not well equipped to identify this periodicity over a short length of the sequence (e.g., 69 in our case). However, DSP-based methods combined with data-driven methods still improve the accuracy of prediction, because the two approaches exploit different information. Table IV summarizes the comparison, giving results for AUC,

TABLE III
SUMMARY OF PERIOD-3 DETECTION RESULTS ON THREE DATASETS

Period-3 Detection Method	Data-driven (Y/N)	Burset/Guigo1996					HMR195					GENSCAN test set				
		Area under ROC curve	% impr. Over SC	% of exonic nucleotides detected at false positive			Area under ROC curve	% impr. Over SC	% of exonic nucleotides detected at false positive			Area under ROC curve	% impr. Over SC	% of exonic nucleotides detected at false positive		
				10%	20%	30%			10%	20%	30%			10%	20%	30%
SC	N	0.7634	—	42.9	59.8	70.5	0.8008	—	49.1	65.0	75.0	0.7778	—	46.7	61.6	71.0
SR	Y	—	—	—	—	—	—	—	—	—	—	0.7800	0.29	48.6	62.9	72.4
PWSR	Y	—	—	—	—	—	—	—	—	—	—	0.8123	4.44	53.8	68.7	77.3
PSC	N	0.7702	0.88	46.2	61.0	70.7	0.8061	0.66	52.0	66.9	75.8	0.8114	4.32	52.3	68.3	77.6
ACF	N	0.5795	–24.09	15.5	29.4	41.4	0.6340	–20.83	20.3	35.1	47.9	0.6218	–20.06	18.4	33.3	47.1
AMDF	N	**0.8065**	**5.64**	**52.2**	**67.2**	**76.4**	**0.8323**	**3.93**	55.1	70.5	**79.7**	0.8338	7.20	56.2	72.9	81.7
TDP	N	0.7876	3.17	50.5	63.8	72.6	0.8258	3.12	**57.5**	**70.6**	78.0	0.8335	7.16	59.9	72.5	79.6
AR	N	0.6632	–13.13	29.0	43.5	54.4	0.7128	–11	34.1	50.0	61.3	0.7057	–9.27	35.9	51.7	63.2
AN filter	N	0.6735	–11.78	31.5	45.6	56.0	0.7118	–11.12	37.1	51.2	61.3	0.6875	–11.60	32.5	47.6	58.1
SVD	N	0.7729	1.24	48.0	61.8	70.6	0.8152	1.79	54.8	68.2	76.6	0.8252	6.09	58.7	70.9	78.5
TFH	Y	—	—	—	—	—	—	—	—	—	—	0.8448	8.62	59.5	74.9	81.6
TFH (5D)	Y	—	—	—	—	—	—	—	—	—	—	0.8542	9.82	62.3	76.2	83.7
AR–TFH (17D)	Y	—	—	—	—	—	—	—	—	—	—	**0.8739**	**12.36**	**66.3**	**80.5**	**87.0**

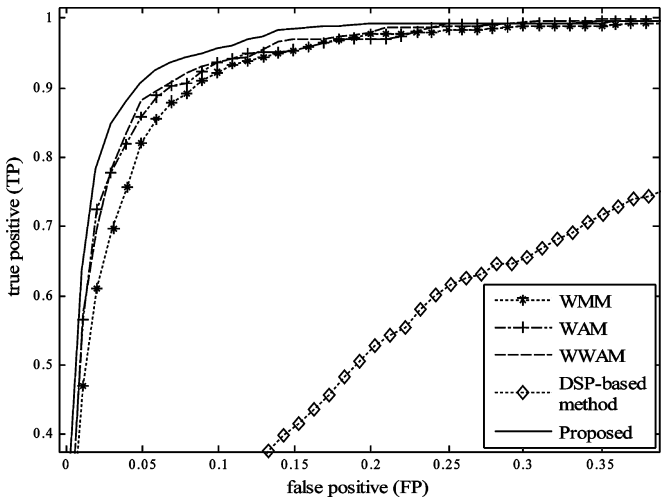

Fig. 9. ROC plot for acceptor site detection, using GENSCAN test set.

false positive and percentage specificity at different levels of percentage sensitivity, for all methods, using the GENSCAN test set. The proposed DSP-statistical hybrid method clearly achieves a larger area under the ROC curve, consistently fewer false positives and higher percent specificities compared with all three existing methods. By comparison with the WWAM method used in gene-finding program GENSCAN [6], the number of false positives across different sensitivity levels in the proposed method shows an average relative improvement of 43%.

According to the results of subsection B, further gains might be expected from using multidimensional feature-based methods; however, these require suitable training data for estimating the GMM parameters and, hence, suffer similar drawbacks to existing data-driven techniques in terms of requiring sufficient organism-specific training data.

TABLE IV
SUMMARY OF ACCEPTOR SPLICE SITE DETECTION RESULTS USING GENSCAN TEST SET

Method	Area under ROC curve	Sensitivity Level					
		25%		50%		75%	
		FP	S_p	FP	S_p	FP	S_p
WMM	0.9655	122	39.4	571	21.7	1724	12.1
WAM	0.9729	102	43.7	389	29.0	993	19.3
WWAM	0.9750	86	48.0	326	32.7	1113	17.6
DSP-based method	0.7354	3654	2.1	8324	1.9	17091	1.4
Proposed	**0.9818**	**31**	**71.9**	**211**	**42.9**	**780**	**23.4**

VII. CONCLUSION

In summary, a number of digital signal processing-based methods for eukaryotic gene prediction have been proposed, and these have been evaluated alongside many other DSP-based methods. Firstly, DNA symbolic-to-numeric mappings were compared in terms of both computational complexity and relative accuracy for the gene and exon prediction problem. From these experiments, the recently proposed paired numeric representation was shown to give an improvement of 2% over the Voss binary indicator sequences in terms of area under the ROC curve for gene and exon prediction, with 75% less downstream processing, when evaluated on the GENSCAN test set.

All 1-D output feature methods for gene and exon prediction were then evaluated on the standard genomic datasets Burset/Guigo1996, HMR195 and GENSCAN. In terms of gene and exon prediction accuracy, the recently proposed TFH, AMDF, TDP, SVD, PWSR, and PSC methods exhibited relatively more accurate gene and exon prediction, improving on the well-known DFT based-SC measure by 4% to 9% in terms of area under the ROC curve. In light of the weaknesses

and strengths of the 1-D genomic period-3 detection methods, we recommend the AMDF and TFH for nondata-driven and data-driven gene detection, respectively. Furthermore, recently proposed multidimensional output feature methods were shown to give improved gene and exon prediction over their 1-D counterparts. By comparison with the most accurate 1-D measure, the multidimensional methods yielded improvements of 5% to 11% in terms of relative increase in exonic nucleotides detected at a 10% false positive rate, when evaluated on the GENSCAN test set. Evaluations of all schemes herein have been performed on large databases and using metrics calculated at the nucleotide level, in contrast to much of the previous literature on the topic.

Finally, we have also proposed a new DSP-statistical hybrid technique for acceptor splice site detection. Results show that DSP-based approaches to gene and exon prediction alone are unlikely to rival current data-driven techniques such as GENSCAN or AUGUSTUS. The proposed DSP-statistical combination for the detection of acceptor splice sites, which achieves a performance improvement of 43% over WWAM (used in GENSCAN), is illustrative of the potential DSP-based techniques still offer in terms of improving the state of the art. Future directions may include more accurate identification of exonic/intronic end-point signals (i.e., start codon, donor splice site, acceptor splice site, and stop codons) using multidimensional DSP-based features, and combining signal processing based work with data-driven methods to advance the state of the art in eukaryotic gene prediction algorithms.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and editor whose helpful suggestions resulted in substantial improvement of this paper.

REFERENCES

[1] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern, "AUGUSTUS: A web server for gene finding in eukaryotes," *Nucl. Acids Res., Web Server Issue*, vol. 32, pp. W309–W312, 2004.

[2] V. V. Solovyev, A. A. Salamov, and C. B. Lawrence, "Identification of human gene structure using linear discriminant functions and dynamic programming," in *Proc. 3rd Int. Conf. Intelligent Systems for Molecular Biology*, 1995, pp. 367–375.

[3] G. Parra, E. Blanco, and R. Guigo, "GeneID in drosophila," *Genome Res.*, vol. 10, no. 4, pp. 511–515, 2000.

[4] A. V. Lukashin and M. Borodovsky, "GeneMark.hmm: New solutions for gene finding," *Nucl. Acids Res.*, vol. 26, no. 4, pp. 1107–1115, 1998.

[5] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, "A generalized hidden Markov model for the recognition of human genes in DNA," in *Proc. 4th Int. Conf. Intelligent Systems for Molecular Biology*, 1996, pp. 134–142.

[6] C. Burge and S. Karlin, "Prediction of complete gene structure in human genomic DNA," *J. Mol. Biol.*, vol. 268, no. 1, pp. 78–94, 1997.

[7] A. Krogh, "Two methods for improving performance of an HMM and their applications for gene-finding," in *Proc. 5th Int. Conf. Intelligent Systems for Molecular Biology*, 1997, pp. 179–186.

[8] S. Salzberg, A. L. Delcher, K. H. Fasman, and J. Henderson, "A decision tree system for finding genes in DNA," *J. Comput. Biol.*, vol. 5, no. 4, pp. 667–680, 1998.

[9] M. Q. Zhang, "Identification of protein coding regions in the human genome by quadratic discriminant analysis," *Proc. Nat. Acad. Sci.*, vol. 94, no. 2, pp. 565–568, 1997.

[10] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, pp. 353–367, 1996.

[11] S. Rogic, A. K. Mackworth, and B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome Res.*, vol. 11, no. 5, pp. 817–832, 2001.

[12] V. Makarov, "Computer programs for eukaryotic gene prediction," *Briefings Bioinf.*, vol. 3, no. 2, pp. 195–199, 2002.

[13] A. Nagar, S. Purushothaman, and H. Tawfik, "Evaluation and fuzzy classification of gene finding programs on human genome sequences," *FSKD*, pp. 821–829, 2005.

[14] S. Logeswaran, E. Ambikairajah, and J. Epps, "A method for detecting short initial exons," in *Proc. IEEE Workshop Genomic Signal Processing and Statistics*, 2006, pp. 61–62.

[15] Y. Saeys, P. Rouze, and Y. V. de Peer, "In search of the short ones: Improved prediction of short exons in vertebrates, plants, fungi and protists," *Bioinformatics*, vol. 23, no. 4, pp. 414–420, 2007.

[16] K. Murakami and T. Takagi, "Gene recognition by combination of several gene-finding programs," *Bioinformatics*, vol. 14, no. 8, pp. 665–675, 1998.

[17] V. Pavlovic, A. Garg, and S. Kasif, "A Bayesian framework for combining gene predictions," *Bioinformatics*, vol. 18, no. 1, pp. 19–27, 2002.

[18] D. Anastassiou, "Genomic signal processing," *IEEE Signal Process. Mag.*, vol. 18, no. 4, pp. 8–20, Apr. 2001.

[19] X. Zhang, F. Chen, Y. Zhang, S. C. Agner, M. Akay, Z. Lu, M. M. Y. Waye, and S. K. Tsui, "Signal processing techniques in genomic engineering," *Proc. IEEE*, vol. 90, no. 12, pp. 1822–1833, Dec. 2002.

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

[21] R. Zhang and C. T. Zhang, "Z curves, an intuitive tool for visualizing and analyzing the DNA sequences," *J. Biomol. Struct. Dyn.*, vol. 11, no. 4, pp. 767–782, 1994.

[22] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *J. Theor. Biol.*, vol. 118, pp. 295–300, 1986.

[23] P. D. Cristea, "Genetic signal representation and analysis," in *Proc. SPIE Inf. Conf. Biomedical Optics Symp.*, 2002, vol. 4623, pp. 77–84.

[24] A. K. Brodzik and O. Peters, "Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences," in *Proc. IEEE ICASSP*, 2005, vol. 5, pp. v/373–v/376.

[25] J. Ning, C. N. Moore, and J. C. Nelson, "Preliminary wavelet analysis of genomic sequences," in *Proc. IEEE Bioinformatics Conf.*, 2003, pp. 509–510.

[26] G. L. Rosen, "Signal processing for biologically-inspired gradient source localization and DNA sequence analysis," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, 2006.

[27] A. S. S. Nair and T. Mahalakshmi, "Visualization of genomic data using inter-nucleotide distance signals," presented at the IEEE Int. Conf. Genomic Signal Processing, 2005.

[28] E. Coward, "Equivalence of two Fourier methods for biological sequences," *J. Math. Biol.*, vol. 36, pp. 64–70, 1997.

[29] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 628–634, Mar. 2002.

[30] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Phys. A*, vol. 249, pp. 511–516, 1998.

[31] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucl. Acids Res.*, vol. 10, pp. 5303–5318, 1982.

[32] S. Tiwari, S. Ramaswamy, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Comput. Appl. Biosci.*, vol. 13, pp. 263–270, 1997.

[33] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions," *Genome Res.*, vol. 18, pp. 1930–1937, 2003.

[34] N. Rao and S. J. Shepherd, "Detection of 3-periodicity for small genomic sequences based on AR techniques," in *Proc. IEEE Int. Conf. Comm., Circuits Syst.*, 2004, vol. 2, pp. 1032–1036.

[35] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using allpass-based filters," presented at the IEEE Workshop Genomic Signal Processing and Statistics, Raleigh, NC, 2002.

[36] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, 2nd ed. Singapore: McGraw-Hill, 2002.

[37] G. Goertzel, "An algorithm for the evaluation of finite trigonometric series," *Amer. Math. Monthly*, vol. 65, no. 1, pp. 34–35, 1958.

[38] W. Li and T. G. Marr, "Understanding long-range correlations in DNA sequences," *Phys. D*, vol. 75, pp. 392–416, 1994.

[39] 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.

[40] W. Li, "The study of correlation structure of DNA sequences: A critical review," *Comput. Chem.*, vol. 21, no. 4, pp. 257–271, 1997.
[41] S. J. Sangwine, "The discrete quaternion Fourier transform," in *Proc. 6th Int. Conf. Image Processing and its Applications*, 1997, vol. 2, pp. 790–793.
[42] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," presented at the IEEE Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland, 2007.
[43] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279–303, 2002.
[44] S. Datta and A. Asif, "A fast DFT based gene prediction algorithm for identification of protein coding regions," in *Proc. IEEE ICASSP*, 2005, vol. 5, pp. 653–656.
[45] C. Burge, "Identification of genes in human genomic DNA," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1997.
[46] M. Akhtar and J. E. E. Ambikairajah, "Time and frequency domain methods for gene and exon prediction in eukaryotes," in *Proc. IEEE ICASSP*, 2007, pp. 573–576.
[47] M. Akhtar, J. Epps, and E. Ambikairajah, "Paired spectral content measure for gene and exon prediction in eukaryotes," in *Proc. IEEE Int. Conf. Information and Emerging Technologies*, 2007, pp. 127–130.
[48] E. Ambikairajah, J. Epps, and M. Akhtar, "Gene and exon prediction using time-domain algorithms," in *Proc. IEEE 8th Int. Symp. Signal Processing and its Applications*, 2005, pp. 199–202.
[49] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-22, no. 5, pp. 353–362, May 1974.
[50] E. Ambikairajah and M. J. Carey, "The time-domain periodogram algorithm," *Signal Process.*, vol. 5, pp. 491–513, 1983.
[51] P. P. Kanjilal, J. Bhattacharya, and G. Saha, "Robust method for periodicity detection and characterization of irregular cyclical series in terms of embedded periodic components," *Phys. Rev. E*, vol. 59, no. 4, pp. 4013–4025, 1999.
[52] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," in *Proc. IEEE Int. Conf. Emerging Technologies*, 2005, pp. 13–17.
[53] M. Akhtar, E. Ambikairajah, and J. Epps, "GMM-based classification of genomic sequences," in *Proc. IEEE 15th Int. Conf. Digital Signal Processing*, 2007, pp. 103–106.
[54] M. Akhtar, E. Ambikairajah, and J. Epps, "Comprehensive autoregressive modeling for classification of genomic sequences," presented at the IEEE 6th Int. Conf. Information, Communications, Signal Processing, 2007.
[55] P. Pollastro and S. Rampone, "$\mathrm{HS^3D}$: Homo sapiens splice site data set," *Nucl. Acids Res. Annu. Database Issue*, 2003.
[56] R. Staden, "Computer methods to locate signals in nucleic acid sequences," *Nucl. Acids Res.*, vol. 12, pp. 505–519, 1984.
[57] C. Burge, "Modeling dependencies in pre-mRNA splicing signals," in *Computational Methods in Molecular Biology*, S. L. Salzberg, D. B. Searls, and S. Kasif, Eds. New York: Elsevier, 1998, ch. 8, pp. 129–164.
[58] M. Q. Zhang and T. G. Marr, "A weight array method for splicing signal analysis," *CABIOS*, vol. 9, no. 5, pp. 499–509, 1993.
[59] S. L. Salzberg, "A method for identifying splice sites and translational start sites in eukaryotic mRNA," *Comput. Appl. Biosci.*, vol. 13, no. 4, pp. 365–376, 1997.
[60] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers HP Laboratories, 2003 [Online]. Available: http://www.hpl.hp.com/personal/Tom Fawcett/papers/ROC101
[61] A. Rushdi and J. Tuqan, "Gene identification using the Z-curve representation," in *Proc. IEEE ICASSP*, 2006, vol. 2, pp. 1024–1027.
[62] M. Akhtar, E. Ambikairajah, and J. Epps, "Optimizing period-3 methods for eukaryotic gene prediction," in *Proc. IEEE ICASSP*, 2008, pp. 621–624.

Mahmood Akhtar received the B.Sc. (Honors) degree in electrical engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, in 2003, and the M.S. degree in computer engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2005. He is currently pursuing the Ph.D. degree in the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia.

His main research interest is in genomic signal processing with specific focus on DNA representations, feature extractions, and classifications of genomic protein coding and noncoding regions. Other research interests lie in the areas of public health bio-surveillance modeling, speech, audio, and image processing. He has authored or coauthored around 16 publications.

Julien Epps (M'99) received the B.E. and Ph.D. degrees from the University of New South Wales, Australia, in 1997 and 2001, respectively.

After an appointment as a Postdoctoral Fellow at the University of New South Wales, he worked on speech recognition and language processing research as a Research Engineer at Motorola Labs and then as a Senior Researcher at National ICT Australia. He joined the UNSW School of Electrical and Telecommunications as a Senior Lecturer in 2007. He has authored or coauthored around 80 publications and has served as a reviewer for several IEEE, IET, and other journals and numerous conferences. His research interests include speaker verification, speech recognition, speech and audio coding, auditory modeling, speech enhancement, and genomic signal processing.

Eliathamby Ambikairajah (M'90) received the Ph.D. degree from Keele University, U.K.

He was appointed as Head of Electronic Engineering and later Dean of Engineering at the Athlone Institute of Technology, Ireland. He was an invited Research Fellow with British Telecom Laboratories (BTL), Martlesham Heath, U.K., for ten years (1989–1999). He joined the University of New South Wales, Australia, in 1999, where he is currently the Deputy Head of School and the Director of Academic Studies in the School of Electrical Engineering and Telecommunications. His research interests include speech and audio compression, speech enhancement and recognition, biometric technology, and biomedical signal processing. He has authored and coauthored around 150 conference and journal papers and is also a regular reviewer for several IEEE, IET, and other journals and conferences.

Prof. Ambikairajah received the Vice-Chancellor's Award for Teaching Excellence in April 2004 for his innovative use of educational technology. He is currently a Fellow and a Chartered Engineer of IET (UK) and EA (Australia).