Ann. Inst. Statist. Math.
Vol. 54, No. 1, 201–223 (2002)

LOCAL SPECTRAL ENVELOPE: AN APPROACH USING DYADIC TREE-BASED ADAPTIVE SEGMENTATION

DAVID S. STOFFER[1]*, HERNANDO C. OMBAO[2]** AND DAVID E. TYLER[3]***

[1]*Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.*
[2]*Department of Statistics and Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.*
[3]*Department of Statistics, Rutgers University, New Brunswick, NJ 08903, U.S.A.*

(Received February 19, 2001; revised July 22, 2001)

Abstract. The concept of the spectral envelope was introduced as a statistical basis for the frequency domain analysis and scaling of qualitative-valued time series. A major focus of this research was the analysis of DNA sequences. A common problem in analyzing long DNA sequence data is to identify coding sequences that are dispersed throughout the DNA and separated by regions of noncoding. Even within short subsequences of DNA, one encounters local behavior. To address this problem of local behavior in categorical-valued time series, we explore using the spectral envelope in conjunction with a dyadic tree-based adaptive segmentation method for analyzing piecewise stationary processes.

Key words and phrases: DNA sequences, gene detection, spectral envelope, dyadic-tree based methods, adaptive segmentation, categorical-valued time series, time-varying spectrum, optimal scaling, Fourier analysis, signal detection.

1. Introduction

The concept of spectral envelope for the spectral analysis and scaling of categorical time series was first introduced in Stoffer *et al.* (1993*a*). Subsequently, Stoffer *et al.* (1993*b*) explored the utility of the methodology for analyzing long DNA sequences. In that article, it was noted that there may be local behavior within a single gene (coding sequence). In this article, we combine dyadic tree-based adaptive segmentation (TBAS) and spectral envelope methodologies to develop a local spectral envelope.

Before discussing the spectral envelope and adaptive segmentation methodologies, we focus on the special problems encountered when analyzing DNA sequence data and in general, categorical-valued time series. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group. There are four different bases that can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A) and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the 5′ carbon of one sugar linked to the 3′ carbon of

*This work was supported, in part, by grant DMS-0102511 from the NSF.
**This work was supported, in part, by grants MH30915 and MH59817 from NIMH, and DMS-0102511 from NSF.
***This work was supported, in part, by grant IRI-9987695 from NSF.

Table 1. Part of the Epstein-Barr virus DNA sequence (read across and down).

AGAATTCGTC	TTGCTCTATT	CACCCTTACT	TTTCTTCTTG	CCCGTTCTCT	TTCTTAGTAT
GAATCCAGTA	TGCCTGCCTG	TAATTGTTGC	GCCCTACCTC	TTTTGGCTGG	CGGCTATTGC
CGCCTCGTGT	TTCACGGCCT	CAGTTAGTAC	CGTTGTGACC	GCCACCGGCT	TGGCCCTCTC
ACTTCTACTC	TTGGCAGCAG	TGGCCAGCTC	ATATGCCGCT	GCACAAAGGA	AACTGCTGAC
ACCGGTGACA	GTGCTTACTG	CGGTTGTCAC	TTGTGAGTAC	ACACGCACCA	TTTACAATGC
ATGATGTTCG	TGAGATTGAT	CTGTCTCTAA	CAGTTCACTT	CCTCTGCTTT	TCTCCTCAGT
CTTTGCAATT	TGCCTAACAT	GGAGGATTGA	GGACCCACCT	TTTAATTCTC	TTCTGTTTGC
ATTGCTGGCC	GCAGCTGGCG	GACTACAAGG	CATTTACGGT	TAGTGTGCCT	CTGTTATGAA

the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inwards. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand. Thus, a strand of DNA can be represented as a sequence of letters, termed base pairs (*bp*), from the finite alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA. A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of noncoding (which makes up most of the DNA). Table 1 shows part of the Epstein-Barr virus (EBV) DNA sequence. The data are taken from the EMBL data base. The entire EBV DNA sequence consists of approximately 172,000 bp.

If we are interested in discovering patterns in a DNA sequence, we could assign numbers (*scales*) to the nucleotides and then proceed with a spectral analysis of the resulting numerical sequence. One could try scaling according to the pyrimidine-purine alphabet, that is $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$, however, this is not necessarily of interest for every CDS of EBV. There are numerous possible alphabets of interest, for example, one might focus on the strong-weak hydrogen bonding alphabet $\mathtt{C} = \mathtt{G} = 0$ and $\mathtt{A} = \mathtt{T} = 1$. While model calculations as well as experimental data strongly agree that some kind of periodic signal exists in certain DNA sequences, there is a large disagreement about the exact type of periodicity. In addition, there is disagreement about which nucleotide alphabets are involved in the signals (for example, compare Ioshikhes *et al.* (1992) with Satchwell *et al.* (1986)).

If we consider the naive approach of arbitrarily assigning numerical values to the categories and then proceeding with a spectral analysis, the result will depend on the particular assignment of numerical values. For example, consider the artificial sequence ACGTACGTACGT.... Then, setting $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$, yields the numerical sequence 010101010101..., or one cycle every two base pairs ($\omega = 1/2$). Another interesting scaling is $\mathtt{A} = 1$, $\mathtt{C} = 2, \mathtt{G} = 3$, and $\mathtt{T} = 4$, which results in the sequence 123412341234..., or one cycle every four bp ($\omega = 1/4$). In this example, both scalings, $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} = \{0, 1, 0, 1\}$ and $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} = \{1, 2, 3, 4\}$, are interesting and bring out different properties of the sequence. It should be clear that one does not want to focus on only one scaling. Instead, the focus should be on finding scalings that bring out all of the interesting features in the data. Moreover, because of heterogeneity (see e.g. Karlin and Macken (1991)), it may be the case that if one scaling works well in one region of a DNA sequence that

same scaling may work poorly in another region. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a categorical time series of virtually any length in a quick and automated fashion.

Although it is well known that DNA is heterogeneous, in Stoffer *et al.* (1993*b*) we found that heterogeneities can exist within short subsequences of a single gene. In this article, we describe a methodology that will automatically divide a DNA sequence into smaller stationary segments and then extract the pertinent information from these segments. Our methodology will be an adaptation of the TBAS method given in Adak (1998) and Ombao *et al.* (2001). This methodology was specifically developed for real-valued nonstationary time series and has been successfully applied to a bivariate EEG data set recorded during an epileptic seizure. In the next section we introduce the concept of spectral envelope for stationary categorical sequences. Then, we establish a theory for the analysis of locally stationary categorical sequences. Finally we discuss fast and automatic estimation of the local spectral envelope via dyadic TBAS methodology and present some examples.

2. The spectral envelope for stationary categorical time series

As a general description, the spectral envelope is a frequency based, principal components technique applied to a multivariate time series. In this section we will focus on the basic concept and its use in the analysis of categorical time series. Technical details can be found in Stoffer *et al.* (1993*a*).

In establishing the spectral envelope for categorical time series, the basic question of how to efficiently discover periodic components in categorical time series was addressed. This was accomplished via nonparametric spectral analysis as follows. Let X_t, $t = 0, \pm 1, \pm 2, \ldots$, be a categorical-valued time series with finite state-space $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. Assume that X_t is stationary and $p_j = \mathrm{pr}\{X_t = c_j\} > 0$ for $j = 1, 2, \ldots, k$. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)' \in \boldsymbol{R}^k$, denote by $X_t(\boldsymbol{\beta})$ the real-valued stationary time series corresponding to the scaling that assigns the category c_j the numerical value β_j, $j = 1, 2, \ldots, k$. We assume the existence of $f_X(\omega; \boldsymbol{\beta})$, the spectral density of $X_t(\boldsymbol{\beta})$. The goal is to find scalings $\boldsymbol{\beta}$ so that the spectral density is in some sense interesting, and to summarize the spectral information by what we called the spectral envelope.

We chose $\boldsymbol{\beta}$ to maximize the power (variance) at each frequency ω, across frequencies $\omega \in [-1/2, 1/2]$, relative to the total power $\sigma^2(\boldsymbol{\beta}) = \mathrm{var}\{X_t(\boldsymbol{\beta})\}$. That is, we chose $\boldsymbol{\beta}(\omega)$, at each ω of interest, so that

$$\lambda(\omega) = \sup_{\boldsymbol{\beta}} \left\{ \frac{f_X(\omega; \boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})} \right\}, \tag{2.1}$$

for $\boldsymbol{\beta} \not\propto \mathbf{1}_k$, the $k \times 1$ vector of ones. Note that $\lambda(\omega)$ is not defined if $\boldsymbol{\beta} \propto \mathbf{1}_k$ because such a scaling corresponds to assigning each category the same value; in this case $f_X(\omega; \boldsymbol{\beta}) \equiv 0$ and $\sigma^2(\boldsymbol{\beta}) = 0$. The optimality criterion $\lambda(\omega)$ possesses the desirable property of being invariant under location–scale changes of $\boldsymbol{\beta}$.

As in most scaling problems for categorical data, it was useful to represent the categories in terms of the unit vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$, where $\boldsymbol{e}_j$ represents the $k \times 1$ vector with a one in the j-th row, and zeros elsewhere. We then defined a k-dimensional stationary time series $\boldsymbol{Y}_t$ by $\boldsymbol{Y}_t = \boldsymbol{e}_j$ when $X_t = c_j$. The time series $X_t(\boldsymbol{\beta})$ can be obtained from the $\boldsymbol{Y}_t$ time series by the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}' \boldsymbol{Y}_t$. Assume that the

vector process $\boldsymbol{Y}_t$ has a continuous spectral density denoted by $f_Y(\omega)$. For each ω, $f_Y(\omega)$ is, of course, a $k \times k$ complex-valued Hermitian matrix. Note that the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$ implies that $f_X(\omega;\boldsymbol{\beta}) = \boldsymbol{\beta}' f_Y(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f_Y^{re}(\omega)\boldsymbol{\beta}$, where $f_Y^{re}(\omega)$ denotes the real part of $f_Y(\omega)$. The optimality criterion can thus be expressed as

$$\lambda(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbf{1}_k} \left\{ \frac{\boldsymbol{\beta}' f_Y^{re}(\omega)\boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\} \tag{2.2}$$

where V is the variance-covariance matrix of $\boldsymbol{Y}_t$. The resulting scaling $\boldsymbol{\beta}(\omega)$ is called the optimal scaling.

The $\boldsymbol{Y}_t$ process is a multivariate point process, and any particular component of $\boldsymbol{Y}_t$ is the individual point process for the corresponding state (for example, the first component of $\boldsymbol{Y}_t$ indicates whether or not the process is in state c_1 at time t). For any fixed t, $\boldsymbol{Y}_t$ represents a single observation from a simple multinomial sampling scheme. It readily follows that $V = D - \boldsymbol{p}\boldsymbol{p}'$, where $\boldsymbol{p} = (p_1, \ldots, p_k)'$, and D is the $k \times k$ diagonal matrix $D = \text{diag}\{p_1, \ldots, p_k\}$. Since, by assumption, $p_j > 0$ for $j = 1, 2, \ldots, k$, it follows that $\text{rank}(V) = k-1$ with the null space of V being spanned by $\mathbf{1}_k$. For any $k \times (k-1)$ full rank matrix Q whose columns are linearly independent of $\mathbf{1}_k$, $Q'VQ$ is a $(k-1) \times (k-1)$ positive definite symmetric matrix.

With the matrix Q as previously defined, and for $\omega \in [-1/2, 1/2]$, define $\lambda(\omega)$ to be the largest eigenvalue of the determinantal equation

$$|Q' f_Y^{re}(\omega) Q - \lambda Q' V Q| = 0, \tag{2.3}$$

and let $\boldsymbol{u}(\omega) \in \boldsymbol{R}^{k-1}$ be any corresponding eigenvector, that is,

$$Q' f_Y^{re}(\omega) Q \boldsymbol{u}(\omega) = \lambda(\omega) Q' V Q \boldsymbol{u}(\omega). \tag{2.4}$$

The eigenvalue $\lambda(\omega) \geq 0$ does not depend on the choice of Q. Although the eigenvector $\boldsymbol{u}(\omega)$ depends on the particular choice of Q, the equivalence class of scalings associated with $\boldsymbol{\beta}(\omega) = Q\boldsymbol{u}(\omega)$ does not depend on Q. A convenient choice of Q is $Q = [\boldsymbol{I}_{k-1} \mid \mathbf{0}]'$, where $\boldsymbol{I}_{k-1}$ is the $(k-1) \times (k-1)$ identity matrix and $\mathbf{0}$ is the $(k-1) \times 1$ vector of zeros. For this choice, $Q' f_Y^{re}(\omega) Q$ and $Q'VQ$ are the upper $(k-1) \times (k-1)$ blocks of $f_Y^{re}(\omega)$ and V, respectively. This choice corresponds to setting the last component of $\boldsymbol{\beta}(\omega)$ to zero.

The value $\lambda(\omega)$ itself has a useful interpretation; specifically, $\lambda(\omega)d\omega$ represents the largest proportion of the total power that can be attributed to the frequencies within a $d\omega$ neighborhood of ω for any particular scaled process $X_t(\boldsymbol{\beta})$, with the maximum being achieved by the scaling $\boldsymbol{\beta}(\omega)$. Because of its central role, $\lambda(\omega)$ was defined to be the *spectral envelope* of a stationary categorical time series.

The name spectral envelope is appropriate because $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. That is, given any $\boldsymbol{\beta}$ normalized so that $X_t(\boldsymbol{\beta})$ has total power one, $f(\omega;\boldsymbol{\beta}) \leq \lambda(\omega)$ with equality if and only if $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$.

Although the law of the process $X_t(\boldsymbol{\beta})$ for any one-to-one scaling $\boldsymbol{\beta}$ completely determines the law of the categorical process X_t, information is lost when one restricts attention to the spectrum of $X_t(\boldsymbol{\beta})$. Less information is lost when one considers the spectrum of $\boldsymbol{Y}_t$. Dealing directly with the spectral density $f_Y(\omega)$ itself is somewhat cumbersome since it is a function into the set of complex Hermitian matrices. Alternatively, one can view the spectral envelope as an easily understood, parsimonious tool for exploring the periodic nature of a categorical time series with a minimal loss of information.

3. Local spectral envelope

In the previous section we assumed stationarity. But, as we have indicated, long DNA sequences are heterogeneous and hence there is a need to establish methods to investigate local behavior. In particular, as discussed in the introduction, the genetic model is that CDS are segments of DNA that are dispersed throughout the sequence and separated by regions of noncoding or noise. Because genetic information is contained in segments, piecewise stationarity appears to be a suitable model.

A $k \times 1$ vector-valued *piecewise stationary process*, $\{\boldsymbol{Y}_{s,T}\}_{s=0}^{T-1}$, for $T \geq 1$, is defined to be

$$\boldsymbol{Y}_{s,T} = \sum_{b=1}^{B} \boldsymbol{Y}_{s,b}\, \mathcal{I}(s/T, U_b); \tag{3.1}$$

here, $\boldsymbol{Y}_{s,b}$ are stationary processes with continuous $k \times k$ spectral matrices $f_{Y,b}(\omega)$, where $U_b = [u_{b-1}, u_b) \subset [0,1)$ is an interval, and $\mathcal{I}(s/T, U_b)$ is an indicator that takes the value 1 if $s/T \in U_b$, and 0 otherwise. For ease of notation, we rescale time in each block so that

$$\{\boldsymbol{Y}_{s,b} : s/T \in U_b\} \mapsto \{\boldsymbol{Y}_{t,b} : t = 0, \ldots, M_b - 1\}$$

where the number of observations in segment b is M_b and $\sum_{b=1}^{B} M_b = T$. This rescaling of time represents a simple time shift to the origin wherein $\boldsymbol{Y}_{s,b} \mapsto \boldsymbol{Y}_{t,b}$ for $s/T \in U_b$ with $t = s - \sum_{i=1}^{b-1} M_i$.

We shall say that a categorical time series, $\{X_{s,T}\}$, on a finite state-space and with nonzero marginal probabilities (as discussed in Section 2), is *piecewise stationary* if the corresponding $k \times 1$ point process, $\{\boldsymbol{Y}_{s,T}\}$, is piecewise stationary. To assure that more observations fall within each stationary segment (or block) upon sampling the process $X_{s,T}$, we assume that the lower bound, M, for the number of observations in each block, b, satisfies $M \to \infty$ as $T \to \infty$. We remark that DNA is truly a discrete-time process, so it would be unrealistic to consider an infill asymptotic situation wherein we assume we are able to obtain more observations in a segment as the number of observations grows. In our case, we rely on increasing-domain asymptotics to approximate the behavior of the estimated spectral envelope for suitably large segments. For small segments, simple Monte Carlo simulations can be used to approximate the small sample null distribution of the spectral envelope estimator.

If $X_{s,T}$ is a piecewise stationary categorical time series, we define the local spectral envelope as follows. The local analogue of the optimality criterion in (2.2) is

$$\lambda_b(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbf{1}_k} \left\{ \frac{\boldsymbol{\beta}' f_{Y,b}^{re}(\omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V_b \boldsymbol{\beta}} \right\}, \tag{3.2}$$

for $b = 1, \ldots, B$, where V_b is the variance-covariance matrix of $\boldsymbol{Y}_{t,b}$. Analogous to Section 2, we define $\lambda_b(\omega)$ to be the *local spectral envelope* and the corresponding eigenvector $\boldsymbol{\beta}_b(\omega)$ to be the *local optimal scaling* of block b and frequency ω.

Next, we present some asymptotic ($T \to \infty$) results for estimators of the local spectral envelope and the corresponding local scaling vectors. In this section we assume that the (stationary) segmentation is known. In the next section, we deal with the problem of estimating the local spectral envelope and optimal scalings when the exact segmentation is not known.

Suppose we observe a finite realization of the piecewise stationary categorical time series $X_{s,T}$, or equivalently, the multinomial process $\boldsymbol{Y}_{s,T}$, for $s = 0, \ldots, T-1$. When the stationary segmentation is known, the theory for estimating the local spectral density of a multivariate, real-valued time series follows from well established results (e.g. Brillinger, (1981); Hannan (1970); Rosenblatt (1959)), and can be applied to estimating $f_{Y,b}(\omega)$, the local spectral density of $\boldsymbol{Y}_{t,b}$. In view of (3.2), given an estimate $\hat{f}_{Y,b}(\omega)$ of $f_{Y,b}(\omega)$, an estimate of the local spectral envelope and the corresponding scalings, denoted $\hat{\lambda}_b(\omega)$ and $\hat{\boldsymbol{\beta}}_b(\omega)$, respectively, can then be defined in a manner analogous to (2.3)–(2.4).

To avoid excessive notation and without loss of generality, we will henceforth use the following conventions (which are typical for categorical data). Let Q be the matrix defined below (2.4), namely, $Q = [\boldsymbol{I}_{k-1} \mid \boldsymbol{0}]'$, and let $\hat{V}_b$ be the sample variance-covariance matrix obtained from the data in segment b, $\{\boldsymbol{Y}_{s,T} : s/T \in U_b\}$, or equivalently, $\{\boldsymbol{Y}_{t,b} : t = 0, \ldots, M_b - 1\}$. We set

$$\boldsymbol{Y}_{t,b} \stackrel{\text{def}}{=} Q'\boldsymbol{Y}_{t,b};$$

this operation has the effect of removing the k-th element from $\boldsymbol{Y}_{t,b}$ so that it is now a $(k-1) \times 1$ vector. In this case, we denote

$$\hat{V}_b \stackrel{\text{def}}{=} Q'\hat{V}_b Q \quad \text{and} \quad \hat{f}_{Y,b}(\omega) \stackrel{\text{def}}{=} Q'\hat{f}_{Y,b}(\omega)Q;$$

note that $\hat{V}_b$ and $\hat{f}_{Y,b}(\omega)$ are now the upper $(k-1)\times(k-1)$ blocks of the previously defined $\hat{V}_b$ and $\hat{f}_{Y,b}(\omega)$ matrices, respectively. In addition, we will use the same convention for the population values V_b and $f_{Y,b}(\omega)$.

For simplicity and without loss of generality, we define the *local sample spectral envelope*, $\hat{\lambda}_b(\omega)$, to be the largest eigenvalue of $\hat{g}_b^{re}(\omega)$ where

$$\hat{g}_b(\omega) = \hat{V}_b^{-1/2}\hat{f}_{Y,b}(\omega)\hat{V}_b^{-1/2}. \tag{3.3}$$

The *local sample optimal scaling*, $\hat{\boldsymbol{\beta}}_b(\omega)$, is then defined by $\hat{\boldsymbol{\beta}}_b(\omega) = \hat{V}_b^{-1/2}\hat{\boldsymbol{u}}_b(\omega)$, where $\hat{\boldsymbol{u}}_b(\omega)$ is the eigenvector of $\hat{g}_b^{re}(\omega)$ associated with the root $\hat{\lambda}_b(\omega)$. The scale corresponding to the k-th category is held fixed at zero. Furthermore, let $\hat{\boldsymbol{u}}_b(\omega)$ be normalized so that $\hat{\boldsymbol{u}}_b'(\omega)\hat{\boldsymbol{u}}_b(\omega) = 1$, and with the first nonzero entry of $\hat{\boldsymbol{u}}_b(\omega)$ taken to be positive.

To allow for the application of a general theory in obtaining asymptotic distributions for the estimates of the local spectral density $f_{Y,b}(\omega)$, we assume throughout this section that $\boldsymbol{Y}_{t,b}$ is strictly stationary for each block b, and that all local cumulant spectra, of all orders, exist for each series $\boldsymbol{Y}_{t,b}$. The assumption of the existence of all local cumulant spectra is not restrictive in the categorical case because the elements of $\boldsymbol{Y}_{t,b}$ take on only two values, zero or one. Rather than introduce excessive notation, we refer to Brillinger ((1981), Assumption 2.6.1). The *local periodogram* of the data $\{\boldsymbol{Y}_{s,T} : s/T \in U_b\}$ in block b is given by

$$I_b(\omega) = \boldsymbol{d}_b(\omega)\boldsymbol{d}_b^*(\omega), \tag{3.4}$$

where

$$\boldsymbol{d}_b(\omega) = M_b^{-1/2} \sum_{t=0}^{M_b-1} \boldsymbol{Y}_{t,b} \exp\{-2\pi i t\omega\} \tag{3.5}$$

is the finite Fourier transform of the data $\{\boldsymbol{Y}_{s,T} : s/T \in U_b\}$.

Under the assumption of piecewise stationarity, and in the case that the stationary blocks are known, the following results regarding estimation follow from Stoffer *et al.*

(1993*a*). The results are stated here for completeness. All limiting statements are taken as $T \to \infty$; recall that $M_b \to \infty$ as $T \to \infty$. For simplicity, the distinct frequencies ω_j, for $j = 1, \ldots, J$, are assumed to be strictly between 0 and 1/2. Let $W(p, \nu, \Sigma)$ denote the Wishart distribution of dimension p on ν degrees of freedom and with $p \times p$ covariance parameter Σ; similarly, $W_c(p, \nu, \Sigma)$ denotes the complex Wishart distribution (see Brillinger, (1981), §4.2 for details).

LEMMA 3.1. *Under the established notation and conditions, $I_b(\omega_j)$, for $j = 1, \ldots, J$, converges in distribution to independent $W_c[k, 1, f_{Y,b}(\omega_j)]$, for $j = 1, \ldots, J$.*

Since $\hat{V}_b$ converges in probability to V_b, we have $\hat{g}_b(\omega_j)$, for $j = 1, \ldots, J$, are asymptotically independent $W_c[k-1, 1, g_b(\omega_j)]$, for $j = 1, \ldots, J$, where

$$g_b(\omega) = V_b^{-1/2} f_{Y,b}(\omega) V_b^{-1/2}$$

is the population version of (3.3). Since the eigenvalues and eigenvectors are continuous functions of a matrix argument, at least almost everywhere with respect to Lebesgue measure, the asymptotic distributions of the sample spectral envelope $\hat{\lambda}_b(\omega)$ and the sample scalings $\hat{\beta}_b(\omega)$ follow from Lemma 3.1.

THEOREM 3.1. *Under the established notation and conditions, and for $\hat{f}_{Y,b}(\omega) = I_b(\omega)$, the collection $\{\hat{\lambda}_b(\omega_j), \hat{\boldsymbol{\beta}}_b(\omega_j) : j = 1, \ldots, J\}$, converges in distribution to $\{\lambda_{b,j}, \boldsymbol{\beta}_{b,j} : j = 1, \ldots, J\}$, where $\boldsymbol{\beta}_{b,j} = V_b^{-1/2} \boldsymbol{u}_{b,j}$ and $\{\lambda_{b,j}, \boldsymbol{u}_{b,j} : j = 1, \ldots, J\}$, are the largest eigenvalue and eigenvector of independent $W_c^{re}[k-1, 1, g_b(\omega_j)]$ matrices, $j = 1, \ldots, J$, with $\boldsymbol{u}_{b,j}$ normalized so that $\boldsymbol{u}'_{b,j} \boldsymbol{u}_{b,j} = 1$ and the first nonzero entry of $\boldsymbol{u}_{b,j}$ being positive.*

The above theorem gives a representation for the limiting distribution of the local sample spectral envelope and the corresponding sample scalings. Although the distribution of the largest root of a Wishart matrix or of a complex Wishart matrix has been well studied, we are not aware of any results on the distribution of the largest root of the real part of a complex Wishart matrix other than Stoffer *et al.* (1993*a*). Except for special cases, the form of the distribution of the largest root of a Wishart matrix or of a complex Wishart matrix is not tractable and contains the other roots of the matrix argument as nuisance parameters (Muirhead (1982)). The distribution of the largest root of the real part of a complex Wishart matrix is more problematic since the distribution $W_c^{re}(p, \nu, \Sigma)$ is not Wishart itself, and depends not only on Σ^{re} but also on Σ^{im}, the imaginary part of Σ.

A special case of fundamental importance is the case where $\boldsymbol{Y}_{t,b}$ is white noise wherein $g_b(\omega) = \boldsymbol{I}_{k-1}$, the $(k-1) \times (k-1)$ identity matrix, for $-1/2 < \omega \leq 1/2$. In this case, the distribution of the largest root of a $W_c^{re}[k-1, 2, g(\omega)]$ matrix, which arises in Theorem 3.1, has a relatively simple form; see Stoffer *et al.* (1993*a*, Theorem 3.2).

THEOREM 3.2. *Under the established notation and conditions, if $\boldsymbol{Y}_{t,b}$ is white noise, then for $\hat{f}_{Y,b}(\omega) = I_b(\omega)$, the collection $\{\hat{\lambda}_b(\omega_j) : j = 1, \ldots, J\}$, converges in distribution to $\{\lambda_{b,j} : j = 1, \ldots, J\}$, where the $\lambda_{b,j}$, for $j = 1, \ldots, J$, are independent and identically distributed with*

$$\mathrm{pr}(\lambda_{b,1} < x) = \mathrm{pr}(\chi^2_{2(k-1)} < 4x) - \pi^{1/2} x^{(k-2)/2} \exp(-x) \mathrm{pr}(\chi^2_k < 2x) \,/\, \Gamma[(k-1)/2],$$

for $x > 0$.

If the spectral estimate $\hat{f}_{Y,b}(\omega)$ is chosen to be the averaged local periodogram estimate

$$\hat{f}_{Y,b}(\omega) = (2m+1)^{-1} \sum_{\ell=-m}^{m} I_b(\omega + \ell/M_b),$$

(the size of m can and should depend on M_b but, for simplicity, we do not display this dependence) then Lemma 3.1 holds with $I_b(\omega_j)$ and $W_c[k, 1, f_{Y,b}(\omega_j)]$ replaced by $\hat{f}_{Y,b}(\omega_j)$ and $W_c[k, 2m+1, f_{Y,b}(\omega_j)]/(2m+1)$, respectively (Brillinger, (1981), Theorem 7.3.3). Consequently, Theorem 3.1 holds when adjusted analogously. Theorem 3.2 also holds when $I_b(\omega_j)$ is replaced by the estimate $\hat{f}_{Y,b}(\omega_j)$, in which case the distribution of $\lambda_{b,1}$ is that of the largest root of a $W(k-1, 4m+2, \boldsymbol{I}_{k-1})/(4m+2)$ matrix. We refer the reader to Muirhead ((1982), §9.7) for a discussion of the largest root of a $W(p, \nu, \boldsymbol{I}_p)$ matrix.

Finally, we consider local consistent window spectral estimates. Consider a window function (which may be different in each block) $W_b(\alpha)$, $-\infty < \alpha < \infty$, that is real-valued, even, of bounded variation, with $\int_{-\infty}^{\infty} W_b(\alpha)d\alpha = 1$, and $\int_{-\infty}^{\infty} |W_b(\alpha)|d\alpha < \infty$. Define

$$\hat{f}_{Y,b}(\omega) = M_b^{-1} \sum_{\ell=0}^{M_b-1} W_{M_b}(\omega - \ell/M_b) I_b(\ell/M_b), \tag{3.6}$$

where $W_{M_b}(\alpha) = B_{M_b}^{-1} \sum_{j=-\infty}^{\infty} W_b(B_{M_b}^{-1}[\alpha + j])$ and B_{M_b} is a bounded sequence of non-negative scale parameters such that $B_{M_b} \to 0$ and $B_{M_b} M_b \to \infty$ as $T \to \infty$. The limiting distribution for (3.6) is a special case of Brillinger ((1981), Theorem 7.4.4) and we state the result as Lemma 3.2. Based on the limiting distribution of (3.6), we may establish the asymptotic distribution of the local spectral envelope and the corresponding scalings obtained from local window spectral estimates. To this end, define

$$\nu_{M_b} = (B_{M_b} M_b)^{1/2} \left(\int_{-\infty}^{\infty} W_b(\alpha)^2 d\alpha \right)^{-1/2}.$$

LEMMA 3.2. *Under the stated conditions and assumptions, for $\hat{f}_{Y,b}(\omega)$ defined by (3.6), $\{\nu_{M_b}[\hat{f}_{Y,b}(\omega_j) - f_{Y,b}(\omega_j)] : j = 1, \ldots, J\}$ converges in distribution to $\{Z_{b,j} : j = 1, \ldots, J\}$ where the Z_j are mutually independent $k \times k$ complex matrices with $(Z_{b,j}^{re}, Z_{b,j}^{im})$ having a multivariate normal distribution with mean zero and covariance structure not dependent on the window $W_b(\alpha)$.*

If the largest root of $g_b^{re}(\omega_j)$ is distinct, the delta method can be used to argue that $\hat{\lambda}_b(\omega_j)$ and $\hat{\boldsymbol{\beta}}_b(\omega_j)$ are jointly asymptotically normal. This statement follows because the maximum eigenvalue of a symmetric matrix and the corresponding eigenvector are analytic in a neighborhood of an argument with a distinct maximum root. Let $\boldsymbol{u}_b(\omega)$ be the normalized eigenvector corresponding to the largest eigenvalue of $g_b^{re}(\omega)$; that is, $\boldsymbol{u}_b(\omega)'\boldsymbol{u}_b(\omega) = 1$, and the first nonzero element of $\boldsymbol{u}_b(\omega)$ is positive. Then, using Lemma 3.2 and the calculations given in Stoffer *et al.* ((1993*a*), Appendix) we have the following main result.

THEOREM 3.3. *Under the stated conditions and assumptions, and for $\hat{f}_{Y,b}(\omega)$ defined by (3.6), if for each $j = 1, \ldots, J$, the largest root of $g_b^{re}(\omega_j)$ is distinct, then*

$\{\nu_{M_b}[\hat{\lambda}_b(\omega_j) - \lambda_b(\omega_j)]/\lambda_b(\omega_j); \nu_{M_b}[\hat{\boldsymbol{\beta}}_b(\omega_j) - \boldsymbol{\beta}_b(\omega_j)] : j = 1, \ldots, J\}$ converges jointly in distribution to $\{z_j; \boldsymbol{y}_j : j = 1, \ldots, J\}$ with z_j and $\boldsymbol{y}_j$ being independent for $j = 1, \ldots, J$. Furthermore, for each $j = 1, \ldots, J$, z_j has a standard normal distribution and is independent of $\boldsymbol{y}_j$ which is multivariate normal with mean zero. The covariance matrix of $V_b^{1/2}\boldsymbol{y}_j$ is given by

$$\{\lambda_b(\omega_j)H_b(\omega_j)^+ g_b^{re}(\omega_j)H_b(\omega_j)^+ - \boldsymbol{a}_b(\omega_j)\boldsymbol{a}_b(\omega_j)'\}/2, \tag{3.7}$$

where $H_b(\omega_j) = g_b^{re}(\omega_j) - \lambda_b(\omega_j)\boldsymbol{I}_{k-1}$, $\boldsymbol{a}_b(\omega_j) = H_b(\omega_j)^+ g_b^{im}(\omega_j)V_b^{1/2}\boldsymbol{u}_b(\omega_j)$, and $H_b(\omega_j)^+$ refers to the Moore-Penrose inverse of $H_b(\omega_j)$.

Asymptotic normal confidence intervals and tests for $\lambda_b(\omega)$ can be readily constructed using Theorem 3.3. For $\boldsymbol{\beta}_b(\omega)$, asymptotic confidence ellipsoids and chi-square tests can also be constructed. A simpler asymptotic test statistic can be constructed by replacing the term $\boldsymbol{a}_b(\omega)$ in (3.7) by zero; details follow analogously to stationary case presented in Stoffer *et al.* (1993*a*). We note that the asymptotic distribution of $\hat{\lambda}_b(\omega_j)$ and $\hat{\boldsymbol{\beta}}_b(\omega_j)$ is considerably more complicated whenever the largest root of $g_b^{re}(\omega_j)$ is not distinct. In this case, we refer the reader to the techniques given in Tyler (1981) and Eaton and Tyler (1991).

For practical purposes, when the number of observations in block b, namely M_b, is large, searching for peaks in the local spectral envelope estimate can be aided using the following approximations. Using a first order Taylor expansion we have

$$\log \hat{\lambda}_b(\omega) \approx \log \lambda_b(\omega) + \frac{\hat{\lambda}_b(\omega) - \lambda_b(\omega)}{\lambda_b(\omega)}, \tag{3.8}$$

so that $\nu_{M_b}[\log \hat{\lambda}_b(\omega) - \log \lambda_b(\omega)]$ is approximately standard normal under the conditions for which Theorem 3.3 is true. It also follows that $E[\log \hat{\lambda}_b(\omega)] \approx \log \lambda_b(\omega)$ and $\text{var}[\log \hat{\lambda}_b(\omega)] \approx \nu_{M_b}^{-2}$. If there is no signal present in block b, we expect $\lambda_b(j/M_b) \approx 2/M_b$ for $1 < j < M_b/2$. Simulations show that M_b must be very large before this approximation holds, and at typical block sizes, the average value of $\hat{\lambda}_b(j/M_b)$ is closer to $2.5/M_b$ when there is no signal present. Using this recommended value, when there is no signal presently, approximately $(1 - \alpha) \times 100\%$ of the time, $\log \hat{\lambda}_b(\omega)$ will be less than $\log(2.5/M_b) + (z_\alpha/\nu_{M_b})$ where z_α is the $(1 - \alpha)$ upper tail cutoff of the standard normal distribution. Exponentiating, the α critical value for $\hat{\lambda}_b(\omega)$ becomes $(2.5/M_b)\exp(z_\alpha/\nu_{M_b})$. From our experience, thresholding at very small values of α relative to the sample size works well provided M_b is not too small. Our experience with DNA suggests that asymptotic approximations work well when M_b is at least 2^8.

4. Tree-based adaptive segmentation for the local spectral envelope

In the previous section we assumed that the exact segmentation is known. Although it may be possible to know the exact segmentation on visual inspection and careful microanalysis of a DNA sequence, we want to focus on the problem of fast and automatic detection of CDS dispersed throughout a long DNA sequence. We do not claim that our method will locate every CDS in a sequence, but we do believe that our method can find the approximate location of many of the genes in a DNA sequence. In this section, we will describe an algorithm for automatically segmenting a long DNA sequence. In

addition, this technique can be generalized to categorical sequences other than DNA. The strategy adopted is to divide the sequence into small blocks and then to recombine adjacent blocks whose estimated local spectral envelopes are sufficiently similar. The basic idea is that adjacent blocks with similar local spectral envelope estimates give similar genetic information. The main feature of the algorithm is it divides the sequence in a dyadic manner using a measure of distance (or discrepancy) between the genetic coding information contained at two adjacent blocks. Our method is inspired by the algorithm in Adak (1998).

We now give the algorithm.

1. *Set the maximum level J.* The value of J determines the smallest possible size of the segmented blocks. For a sequence of length T, the smallest blocks have length $T/2^J$. Ideally, the block sizes should be small enough so that one can separate useful genetic information unique to that block from the noncoding material (noise). One should be careful, however, about making the blocks too small. Blocks have to be large enough to give good estimates of the local spectral envelope. Our recommendation is that the block size should be at least 2^8.

2. *Form the blocks.* At each level $j = 0, \ldots, J$, divide the data sequence into 2^j blocks. Denote $B(j, \ell)$ to be the ℓ-th block on level j, where $\ell = 1, \ldots, 2^j$. The first block on level j is denoted as $B(j, 1)$ and the last as $B(j, 2^j)$. The "inner" blocks at level j are $B(j, \ell)$, (where $\ell = 2, \ldots, 2^j - 1$). For any level $j = 0, \ldots, J$, block $B(j, \ell)$, for $\ell = 1, \ldots, 2^j$, consists of the $M_j = T/2^j$ elements $\{X_{[(\ell-1)T/2^j]}, \ldots, X_{[\ell T/2^j - 1]}\}$.

3. *Estimate the spectral envelope.* Compute an estimate of the local spectral envelope, $\hat{\lambda}_{j,\ell}(\omega_k)$, at each fundamental frequency $\omega_k = k/M_j$ $(k = 0, \ldots, M_j/2)$ in each block $B(j, \ell)$ where $j = 0, \ldots, J$, and $\ell = 1, \ldots, 2^j$.

4. *Create a table of distances.* Let $\delta[\cdot, \cdot]$ be a distance (discrepancy) measure between the spectral envelope estimates of two children blocks. We will discuss choosing such a measure after the algorithm is presented. Using the distance measure, create a table of distances corresponding to each block, $B(j, \ell)$, namely,

$$D(j, \ell) = \delta[\hat{\lambda}_{j+1,2\ell-1}(\omega), \hat{\lambda}_{j+1,2\ell}(\omega)],$$

for $\ell = 1, \ldots, 2^j$, and for each level $j < J$.

5. *Mark the blocks for final segmentation.* Mark all the blocks $B(J-1, \ell)$, at level $J-1$ for $\ell = 1, \ldots, 2^{J-1}$. For $j = J-2$, and $\ell = 1, \ldots, 2^j$, if

$$D(j, \ell) \leq D(j+1, 2\ell - 1) + D(j+1, 2\ell)$$

then mark the block $B(j, \ell)$ and leave $D(j, \ell)$ unchanged. Otherwise, leave the block $B(j, \ell)$ as unmarked and set

$$D(j, \ell) = D(j+1, 2\ell - 1) + D(j+1, 2\ell).$$

Iterate this procedure for $j = J-3, J-4, \ldots, 0$. The *final segmentation* of the DNA sequence is the set of highest marked blocks: $\{B(j, \ell)$ such that $B(j, \ell)$ is marked and its parent block and ancestor blocks are not marked$\}$.

6. *Classification.* For the final segmentation, use the information in the estimated local spectral envelope to classify a segment as (i) highly likely to contain CDS (ii) highly likely to contain noncoding, or (iii) uncertain. A specific classification method is discussed below.

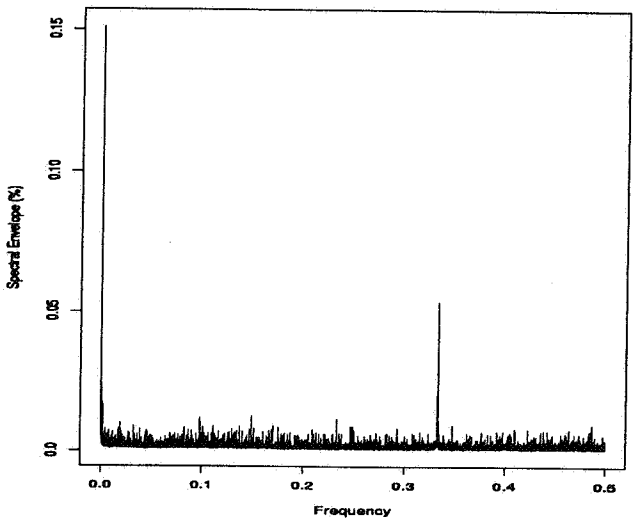

Fig. 1. Estimated spectral envelope of the entire EBV sequence.

Choosing a distance measure and classification rule for gene detection

Our choice for a distance measure in Step 4, and a classification rule in Step 6, of the algorithm is based on our extensive experience with the Fourier analysis of DNA sequences (e.g. Stoffer *et al.* (1993*b*)), and the research of others (e.g. Cornette *et al.* (1987), or Tiwari *et al.* (1997)). The consensus is that a CDS typically contains the frequency $\omega = 1/3$. Other frequencies, such as $\omega = 1/10$ may also be present (e.g. Satchwell *et al.* (1986)), and repeat regions may have many spectral peaks. Introns (junk DNA) behave generally as second order noise and sometimes fractional noise (also known as $1/f$ noise, e.g. see Voss (1992)), so the spectral envelope will either be flat or will have spectral power at or near the zero frequency in these regions. If a block contains coding only, then generally, a spectral peak will appear at the 1/3 frequency. If a block contains coding and noise, then a spectral peak at 1/3 will be present, but there may also be power at the zero frequency, indicating long memory or $1/f$ spectra.

We demonstrate these concepts using the EBV DNA sequence that was described in Section 1. Figure 1 shows the estimated spectral envelope of the entire sequence. We note the pronounced peaks at the zero and 1/3 frequencies. Also, the possibility that there are other significant peaks (such as the presence of some spectral power at the 1/10 frequency) are visible in the estimated spectral envelope of the entire sequence. Figure 2 shows the estimated spectral envelopes of the first half and the second half of the sequence. In this case, we note that the estimated spectral envelopes, though different, display the same peak frequencies as the entire sequence. Figure 3 shows the estimated spectral envelope of the first 4096 bp of the EBV sequence and of the gene BNRF1 (bp 1736-5689). The first 4096 bp of the EBV sequence contains noncoding and coding (specifically, BNRF1 is the first gene of EBV); note the estimated spectral envelope shows peaks at the zero and 1/3 frequencies (the general appearance is similar to the entire sequence). In contrast, the estimated spectral envelope of the gene BNRF1 shows a peak at the 1/3 frequency, but does not exhibit a peak at the zero frequency. Finally, Fig. 4 shows examples of subsequences in the EBV sequence that exhibit the properties of white noise and of fractional noise. The results of these examples are typical

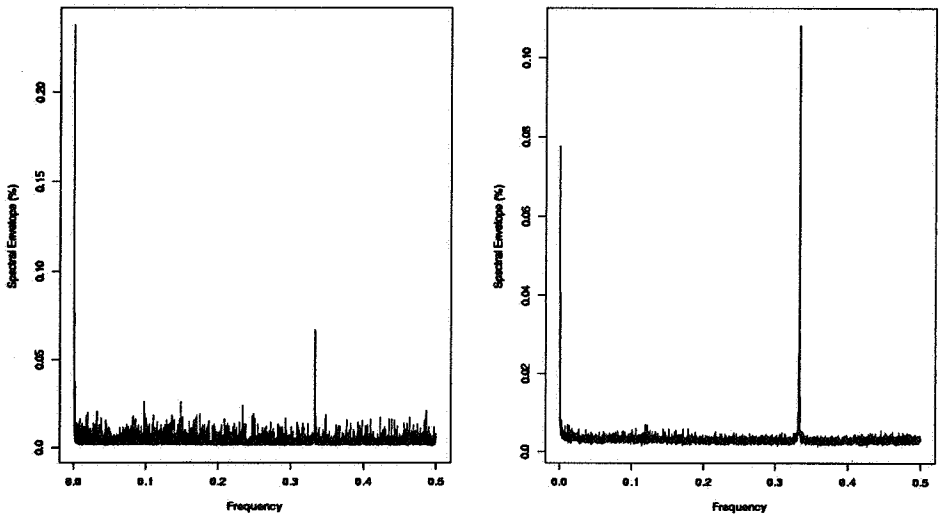

Fig. 2. Estimated spectral envelope of the first half [left] and second half [right] of the EBV sequence.

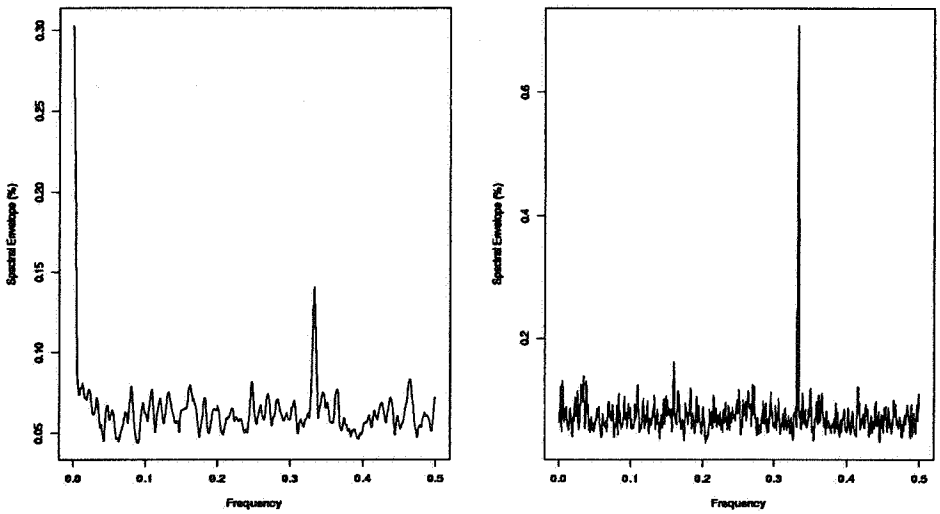

Fig. 3. Estimated spectral envelope of the first 4096 bp of the EBV sequence [left] and of the gene BNRF1 (bp 1736-5689) [right].

of the spectral analysis of DNA sequences discussed in the literature and are the basis of our distance measure and classification rule. To summarize, we emphasize the following observations: (i) If a block contains only coding, the spectral envelope should exhibit a peak at frequency $1/3$, and possibly other nonzero frequencies; (ii) if a block contains both coding and noncoding, the spectral envelope will exhibit a peak at (or near) the zero frequency as well as a peak at frequency $1/3$, and possibly other nonzero frequencies; (iii) if a block contains noncoding in the form of noise, the spectral envelope will be flat or will indicate $1/f$ noise; (iv) if a block contains other interesting features (e.g. repeat

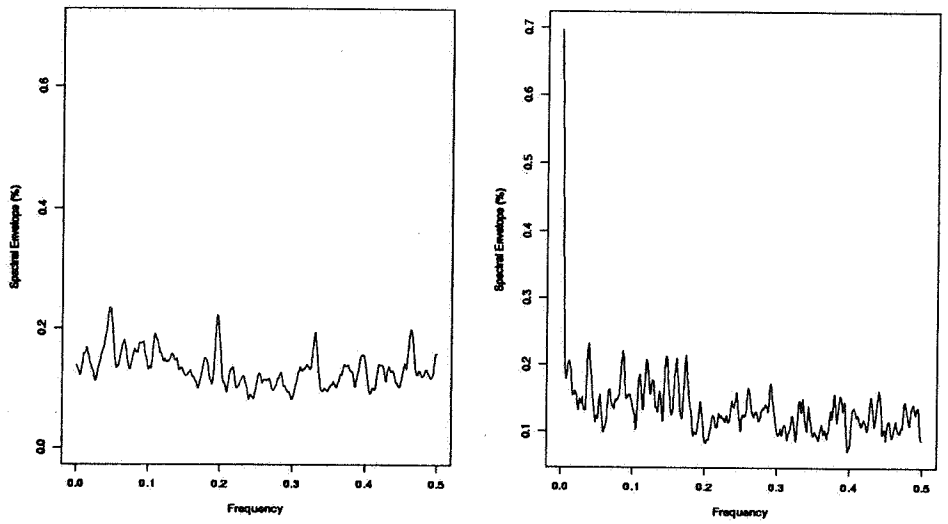

Fig. 4. Estimated spectral envelope of 2048 bp from the EBV sequence known to be noncoding in the form of white noise [left] and in the form of fractional noise [right].

regions) the spectral envelope may exhibit several nonzero peaks other than 1/3.

Using these ideas, our recommended *distance measure* is as follows:

1. Define the threshold at level j to be α_j. A discussion of significance levels and thresholding is given in the paragraph containing equation (3.8).

2. Compute the *peak information function* at block $B(j,\ell)$ as follows. Let κ be a positive integer (which may depend on j) such that $2\kappa << M_j/2$, and consider the partition of the interval of frequencies, $[0, 1/2]$, given by

$$\left\{\Omega_j(v) = \left[\frac{v}{2\kappa}, \frac{v+1}{2\kappa}\right);\ \text{for } v = 0, 1, \ldots, \kappa - 1\right\},$$

with the convention that the last partition contains 1/2. We define the peak information function to be

$$P_{j,\ell}(v) = \begin{cases} 1, & \text{if } \hat{\lambda}_{j,\ell}(\omega_k) \geq \alpha_j \text{ for any } \omega_k \in \Omega_j(v) \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

for $v = 0, 1, \ldots, \kappa - 1$, where $\omega_k = k/M_j$ for $k = 0, 1, \ldots, M_j/2$, are the fundamental frequencies. In other words, $P_{j,\ell}(v)$ is assigned the value of 1 if there is a significant peak in the estimated spectral envelope in the band of frequencies $\Omega_j(v)$. If there are no significant peaks in the particular band, then $P_{j,\ell}(v)$ is assigned the value zero. The partition is arbitrary, but we have found the partition with endpoints $\{.00, .01, .02, \ldots, .50\}$, obtained by rounding the fundamental frequencies to two decimal places, works well.

3. Compute the distance, D, between two children blocks. For $j = 0, \ldots, J-1$, and $\ell = 1, \ldots, 2^j$:

$$\begin{aligned} D(j,\ell) = &\sum_{v>0} |P_{j+1,2\ell-1}(v) - P_{j+1,2\ell}(v)| \\ &+ (J - j - 1)[P_{j+1,2\ell-1}(0) + P_{j+1,2\ell}(0)]. \end{aligned} \tag{4.2}$$

Our distance measure is based on the location of peaks in the spectral envelope. The indicator $P_{j,\ell}(v)$ identifies whether or not a significant local peak exists in the estimated spectral envelope and similar spectral envelopes in children blocks will yield a low value in the sum in (4.2). The second part of (4.2) is a penalty term that guards against combining large blocks in which there is a significant peak at the zero frequency, which, in the presence of other nonzero peaks, may indicate inhomogeneity (i.e. coding and noncoding in the same block).

Once the final segmentation is determined, our *classification rule* uses the points previously discussed.

1. A block is designated as containing only coding if the local estimated spectral envelope exhibits a peak at frequency 1/3 (and possibly other nonzero frequencies), but no peak exists at the zero frequency (see Fig. 3).

2. A block is designated as containing both coding and noncoding if the spectral envelope exhibits a peak at (or near) the zero frequency as well as a peak at frequency 1/3, and possibly other nonzero frequencies (see Fig. 2).

3. A block is designated as containing noncoding (noise) if the spectral envelope is either flat, indicating white noise, or has a peak at, or near, the zero frequency and no other peaks, indicating fractional noise (see Fig. 4).

4. A block is designated as containing other interesting features (e.g. repeat regions) if spectral envelope exhibits several nonzero peaks other than 1/3.

5. If adjacent blocks are classified in the same way, they may be recombined (see the data example in Section 6).

Before proceeding to our examples, we make the following remarks.

- *On dyadic segmentation*: Dyadic tree-structured based methods are widely used and well accepted in the statistics literature. One example of a dyadic tree-based method is CART (Classification and Regression Trees) of Breiman *et al.* (1984). In the time series literature, we now have well developed methods and theory that are based on dyadic segmentation; see, for example, Mallat *et al.* (1998), Adak (1998), Donoho *et al.* (1998) and Ombao *et al.* (2001). The Auto-SLEX method in Ombao *et al.* (2001) was applied successfully to a nonstationary EEGs recorded during an epileptic seizure. The goal was to estimate the time-varying spectra of the EEGs and coherence between the two EEGs. It was clearly demonstrated in Ombao *et al.* (2001) that the Auto-SLEX method, which is dyadic-based, does not suffer even when applied to biological signals that do not necessarily have a dyadic structure. The only conditions that need to be satisfied are that the length of the blocks at the finest level J goes to infinity at a rate that is slower than the length of the whole time series.

- *On the distance*: The distance $D(j,\ell)$ counts the number of peaks occurring at different nonzero frequencies between the children blocks. If the peaks at the two children blocks occur at different frequencies then the two children blocks are said to contain different genetic information. The magnitude of the difference in genetic information is captured by the distance $D(j,\ell)$. It may, however, be the case that if two segments have the same local spectral envelope, the signals may be due to different alphabets (i.e., the scales at the common peak frequencies may be different). In this case, a distance measure should indicate a difference. How to incorporate this information efficiently is currently under investigation.

- *Efficient computation*: It is necessary to use computationally efficient methods when analyzing very long time series data sets. Dyadic transforms are useful tools

for developing computationally efficient methods. The algorithm presented is efficient because it uses two computationally efficient methods. In computing the estimates of the spectral envelope, we used the Fast Fourier Transform. Moreover, the selection of the final segmentation was delivered by the Best Basis Algorithm (BBA) of Coifman and Wickerhauser (1992). Wickerhauser (1994) devotes a chapter to BBA and related cost measures.

• *Relationship to the Adak algorithm*: The Adak (1998) method is useful for estimating the time-varying spectrum of a univariate nonstationary process. Spectral estimates are compared between children blocks and distance measures were proposed. The algorithm presented in this article is for multivariate piecewise stationary categorical time series. Instead of spectral estimates, we compare the estimated spectral envelopes between blocks. Moreover, our distance function is different than the one developed in Adak (1998). The distance function used in our algorithm is specific for time series data sets whose spectra have power concentrated at very narrow bands of frequencies.

5. Some simulation results

To test our methods, we conducted two simulation studies. The first simulation divides the time domain of various signal and noise processes in a dyadic manner. The second simulation is similar to the first, but we use a more realistic situation where the piecewise processes have lengths that are nondyadic. In both cases, the signals are generated by discretizing two sinusoidal processes

$$X_1(t) = \cos(2\pi t/10) + \cos(2\pi t/3) + 1.5\epsilon_1(t) \tag{5.1}$$

$$X_2(t) = \cos(2\pi t/3) + \epsilon_2(t), \tag{5.2}$$

where $\epsilon_1(t)$ and $\epsilon_2(t)$ are Gaussian white noise processes with unit variance. These processes are discretized into four categories (representing the nucleosomes A, C, G, and T, with about 65% of the observations corresponding to C and G). The generated categorical sequences are used to represent CDS, the first signal contains the signature 1/3 frequency as well as the additional presence of a 1/10 frequency, whereas the second signal contains only the signature 1/3 frequency. For notation, we use $S_1(t)$ and $S_2(t)$ to represent the categorical sequences obtained by discretizing $X_1(t)$ and $X_2(t)$, respectively. Also, we use $N(t)$ to represent, generically, an iid sequence of discrete random variables taking on one of four possible values (representing noncoding, again with about 65% of the observations corresponding to C and G).

5.1 *Dyadic example*

In this example, let X_t represent a simulated DNA sequence of length $T = 4096$, generated in the following manner:

Table 2. Distances (dyadic example).

level	$D(j,\ell)$							
$j=0$	3							
$j=1$	3				1			
$j=2$	2		0		0		0	
$j=3$	0	0	0	0	0	0	0	0

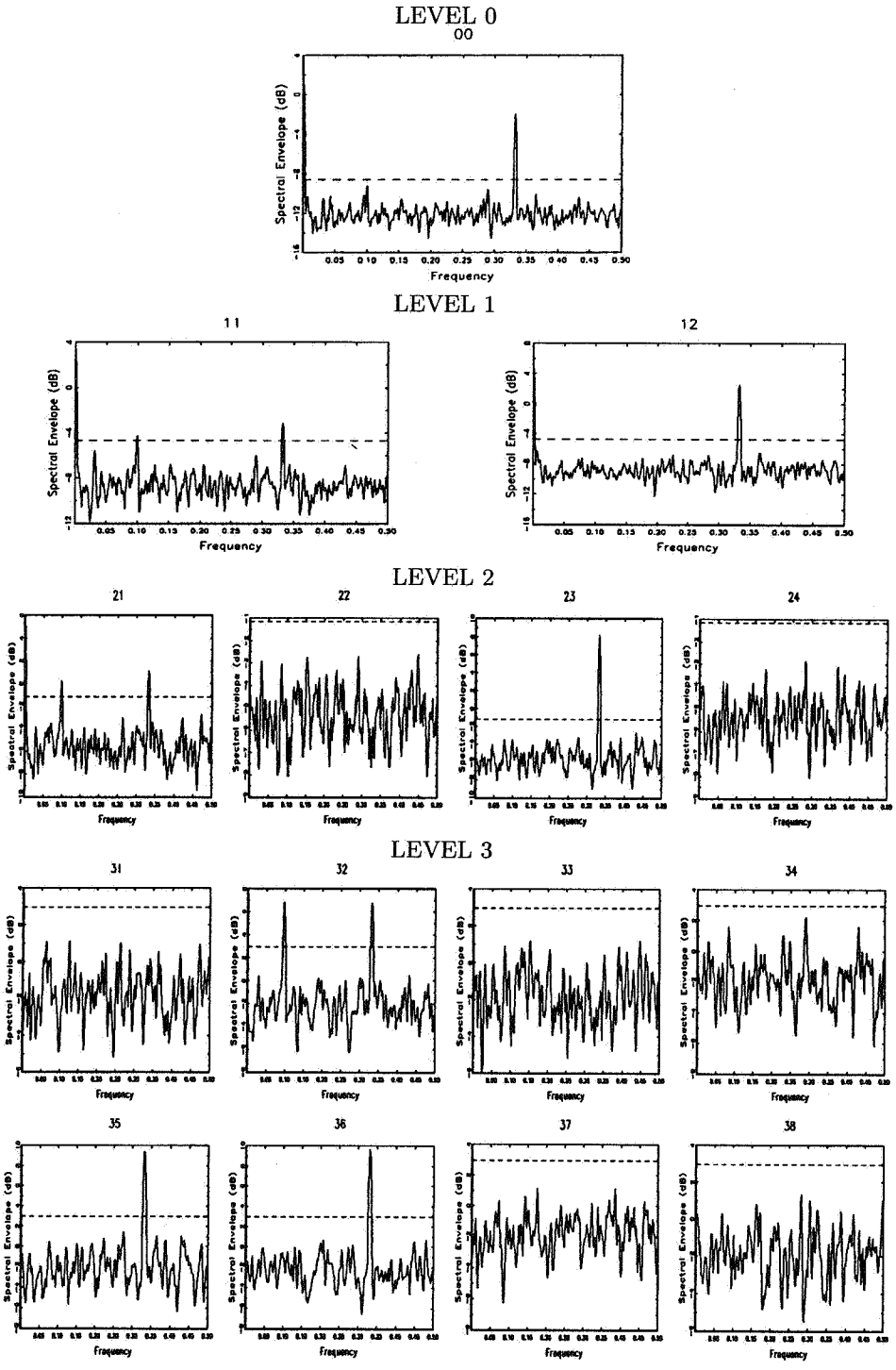

Fig. 5. Estimated spectral envelopes (in decibels) for the dyadic example.

Table 3. Recomputed distances and best segmentation (dyadic example). Bold face indicates block is marked. Asterisks indicate best segmentation.

level	$D(j,\ell)$							
$j=0$	0							
$j=1$	0				0			
$j=2$	0		**0***		**0***		**0***	
$j=3$	**0***	**0***	**0**	**0**	**0**	**0**	**0**	**0**

$$X_t = \begin{cases} N_1(t), & 1 \le t \le 512 \\ S_1(t), & 513 \le t \le 1024 \\ N_2(t), & 1025 \le t \le 2048 \\ S_2(t), & 2049 \le t \le 3072 \\ N_3(t), & 3073 \le t \le 4096 \end{cases} \tag{5.3}$$

We can visualize the decomposition of this simulated sequence using the following diagram. Here, each block represents 512 observations, N denotes noncoding and S denotes coding. The double lines separate the piecewise stationary segments.

N	S_1	N	N	S_2	S_2	N	N

(5.4)

The estimated spectral envelopes at each level, $j = 0, 1, 2, 3$, are given in the Fig. 5. In this example, the lowest level is $J = 4$ wherein the smallest data length considered is $2^8 = 256$; in general, level j corresponds to a window of length 2^{12-j}. The dashed line in each graph represents a significance threshold based on the discussion in the paragraph containing (3.8), with α_j set at $1/2^{12-j}$ for each level j. Special attention should be paid to the presence or absence of the zero frequency. Table 2 shows the distances computed using (4.2), and Table 3 shows the recomputed distances and the best segmentation based on Step 5 of the algorithm. We note that the best segmentation corresponds precisely to the segmentation of the generated data as visualized in (5.4). In addition, using our classification rule, segment (3,1) is classified as noise, segment (3,2) is classified as coding with frequencies 1/10 and 1/3 predominant, segment (2,2) is classified as noise, segment (2,3) is classified as coding with frequency 1/3 predominant, and segment (2,4) is classified as noise. This classification corresponds precisely to the way the data were generated.

5.2 *Non-dyadic example*

This example is similar to the dyadic example except that the segmentation is non-dyadic, and the signal lengths are a factor of 3 (to be more like an actual CDS). In this case, X_t was generated as follows:

$$X_t = \begin{cases} N_1(t), & 1 \le t \le 564 \\ S_1(t), & 565 \le t \le 1023 \\ N_2(t), & 1024 \le t \le 2199 \\ S_2(t), & 2200 \le t \le 3024 \\ N_3(t), & 3025 \le t \le 4096 \end{cases} \tag{5.5}$$

To visualize the segmentation, consider the following display. As before each block represents 512 observations, N denotes noncoding and S denotes coding. The double

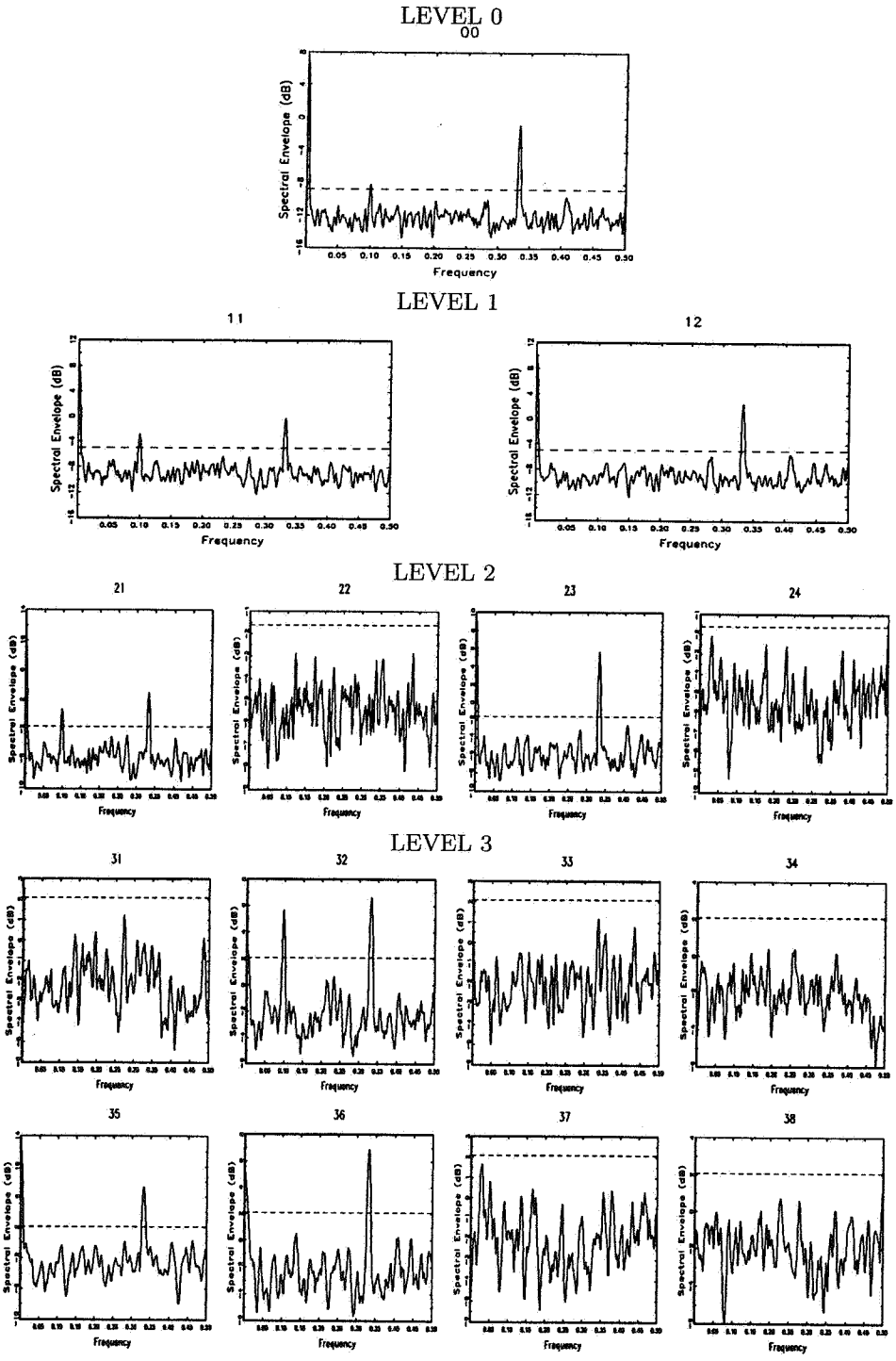

Fig. 6. Estimated spectral envelopes (in decibels) for the non-dyadic example.

Table 4. Distances (non-dyadic example).

level	$D(j,\ell)$							
$j=0$	7							
$j=1$	4				3			
$j=2$	3		0		2		0	
$j=3$	0	1	0	0	2	2	1	0

Table 5. Recomputed distances and best segmentation (non-dyadic example). Bold face indicates block is marked. Asterisks indicate best segmentation.

level	$D(j,\ell)$							
$j=0$	3							
$j=1$	1				2			
$j=2$	1		**0***		**2***		**0***	
$j=3$	**0***	**1***	**0**	**0**	**2**	**2**	**1**	**0**

lines indicate the best segmentation with the smallest block size being 512. For example, the first block contains all noncoding whereas the second block contains coding and noncoding.

(5.6)

N	S_1/N	N	N	S_2/N	S_2/N	N	N

Based on the estimated spectral envelopes in Fig. 6, the distance tables in this example are presented in Tables 4 and 5. As in the dyadic case, our segmentation and classification rule correctly identify the decomposition of the data. In particular, according to our classification rule, segment (3,1) is classified as noise, segment (3,2) is classified as coding with frequencies 1/10 and 1/3 predominant but with some noncoding present (notice the significant peak at the zero frequency), segment (2,2) is classified as noise, segment (2,3) is classified as containing coding with frequency 1/3 predominant and noncoding (note the peak at zero), and segment (2,4) is classified as noise.

6. Application: analysis of the EBV DNA sequence

We applied our algorithm to a data set that is a subsequence of the EBV DNA sequence. The subseries consists of bp 46001 to 54192; the length of the series is $T = 2^{13} = 8192$. Below is a list of the interesting portions of this subsequence taken directly from the EMBL data file:

```
CDS             46333..47481
                /note="BWRF1 reading frame 12"
CDS             48386..50032
                /note="Coding exon for EBNA-2"
repeat_region   50578..52115
                /note="12 x "125bp" repeats"
```

Note that the segment contains two coding sequences (CDS), one from bp 46333 to 47481, and another from bp 48386 to 50032. Also notable is a large repeat region from bp 50578 to 52115; repeat regions are highly repetitive regions DNA.

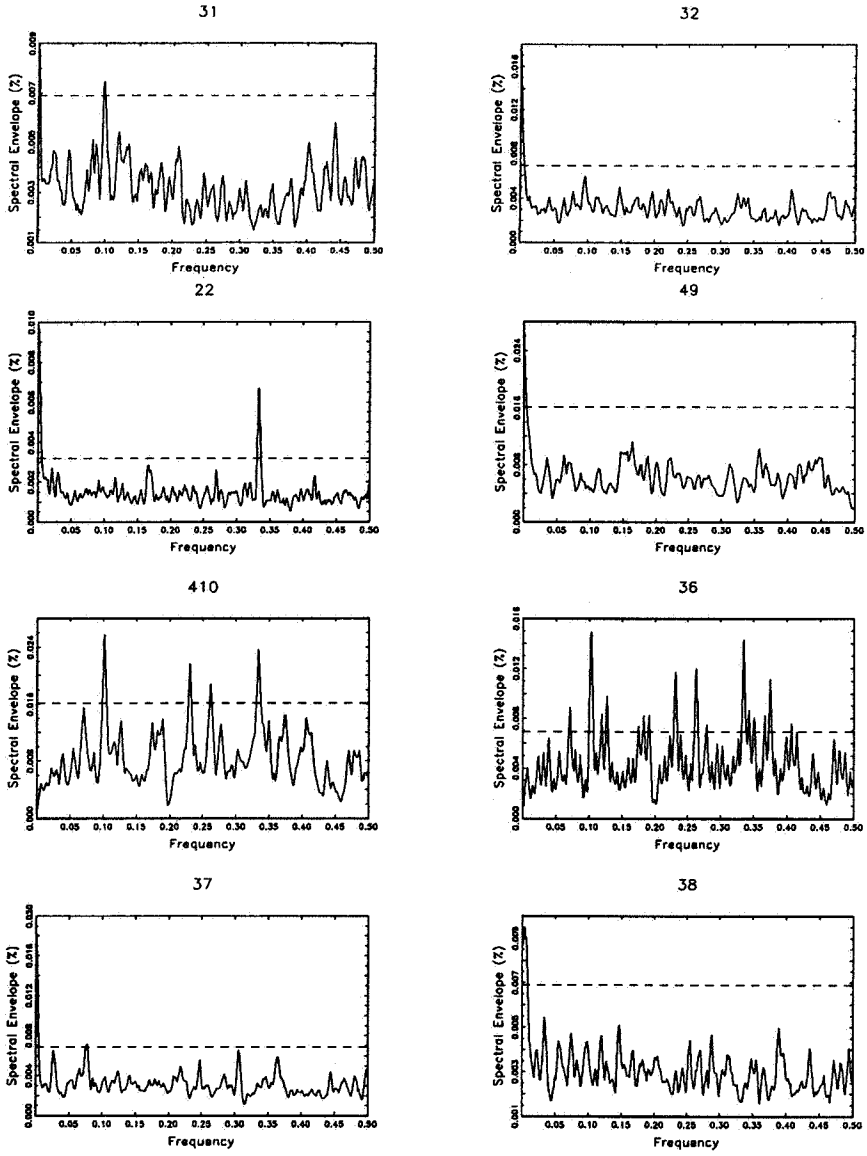

Fig. 7. Estimated spectral envelopes in the EBV example.

In our analysis, we set the lowest level at $J = 5$ so that the smallest blocks have 256 elements, and we set the threshold using $\alpha_j = 1/2^{13-j}$, for $j = 1, \ldots, 5$. The best segmentation and classifications are shown in Table 6. We note that adjacent blocks with the same classification can be recombined; this situation happens with blocks $B(4, 10)$ and $B(3, 6)$, and with blocks $B(3, 7)$ and $B(3, 8)$. The spectral envelopes for the final segmention are displayed in Fig. 7. Our algorithm locates approximately the three interesting segments of the DNA sequence considered here. In particular, block $B(3, 1)$, which includes bp 46001 to 47042, correctly identifies a CDS (as previously noted, the actual location is bp 46333 to 47481). Block $B(2, 2)$, which includes bp 48049 to 50096 correctly identifies a CDS (the actual location is bp 48386 to 50032). Finally, the

Table 6. Best segmentation and classifications. The best segmentation is marked with a letter indicating the classification: **C** = CDS, **N** = Noise, **R** = Repeat region.

level	$B(j,\ell)$															
$j=0$																
$j=1$																
$j=2$					**C**											
$j=3$	**C**		**N**								**R**		**N**		**N**	
$j=4$									**N**	**R**						

combination of blocks $B(4,10)$ and $B(3,6)$, which includes bp 50609 to 52144, correctly identifies a large repeat region (the actual location is bp 50578 to 52115).

7. Discussion and conclusion

In this article we have extended the concept of the spectral envelope for a stationary categorical time series to the situation where the time series is stationary only over intervals or subsequences. DNA sequences exhibit this kind of behavior. In particular, the genetic model presumes that genetic information comes in pieces with definite starting points (or start codons) and ending points (or stop codons). We have presented a method to aid in the identification of coding sequences that are dispersed throughout the DNA and separated by regions of noncoding. To address this problem we explored using the notion of a local spectral envelope in conjunction with a dyadic tree-based adaptive segmentation method. Our focus was on the problem of fast and automatic detection of the approximate location, rather than the precise location of a CDS. Our hope is that this information will allow molecular biologists to focus on small portions of a given sequence. We do not claim that our method will locate every CDS in a sequence, but we do believe that our method can find the approximate location of many of the genes in a DNA sequence. We presented an algorithm for automatically segmenting a long DNA sequence. The strategy adopted was to divide the sequence into small blocks and then to recombine adjacent blocks whose estimated local spectral envelopes are sufficiently similar. The basic idea is that adjacent blocks with similar local spectral envelope estimates give similar genetic information. We provided two simulation studies and an actual data analysis that demonstrated the viability of our methodology. This research is ongoing and some fine tuning of the methodology will certainly be developed in the future. For example, we will investigate other distance measures and classification rules; moreover, we will focus on including the estimated optimal scalings into the algorithm.

In terms of the general problem (i.e. not specific to any particular analysis), we make the following concluding remarks. The spectral envelope could come under the general title of spectral domain principal component analysis of multiple time series. This topic is discussed in detail in Chapter 9 of Brillinger (1981) and there is a connection between Brillinger's work and the spectral envelope. Specifically, Brillinger's approach can be viewed as a scaling problem with complex-valued scales. In the spectral envelope approach, we restrict attention to the case where the scales are real and the vector time series of interest is the multiple indicator process associated with a categorical-valued process.

While the piecewise stationary assumption is reasonable for DNA, we would like to broaden the scope of our technique to evolutionary stationary processes. In the

theoretical development of an evolutionary spectral envelope, we can use the model of a locally stationary process of Dahlhaus (1997, 1999) or its special case given in Chiann and Morettin (1999). A k-dimensional zero-mean random process $\boldsymbol{Y}_{t,T}$, for $t = 0, \ldots, T-1$ is defined in Chiann and Morretin (1999) to be Dahlhaus locally stationary if it admits the spectral representation

$$\boldsymbol{Y}_{t,T} = \int_{-1/2}^{1/2} \exp(2\pi i\omega) A(t/T, \omega) d\boldsymbol{Z}(\omega), \tag{7.1}$$

where $\boldsymbol{Z}(\omega)$ is a vector stochastic process whose increments are orthogonal and satisfy regularity conditions on its cumulants. The $k \times k$ matrix $A(\cdot, \cdot)$ is the time-varying filter. Under this model, the $k \times k$ evolutionary spectral density matrix is defined to be $f_Y(u, \omega) = A(u, \omega)A^*(u, \omega)$. We note that the first argument of $A(t/T, \omega)$ is rescaled to live on the unit interval. Increasing the number of observations, T, does not mean looking into the future. Rather, this asymptotic framework, i.e. the concept of a doubly-indexed sequence of processes $\{\boldsymbol{Y}_{t,T}\}$, allows for more data to be observed at a local structure and to do asymptotic inference starting from a single realization rather than using replications of $\boldsymbol{Y}_{t,T}(t = 0, \ldots, T-1)$. Dahlhaus also proposed a method for estimating $f_Y(u, \omega)$ in his model. The estimators are consistent but the method is not computationally efficient and can be problematic when the time series is long. In Ombao *et al.* (2001), we proposed estimators of the evolutionary spectrum that are based on dyadic segmentation framework. The methods used are computationally efficient and the estimators are mean square consistent.

Under the Dahlhaus model, an evolutionary spectral envelope could be defined analogously to the spectral envelope for stationary processes discussed in Section 2. That is, let

$$\lambda(u, \omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbf{1}_k} \left\{ \frac{\boldsymbol{\beta}' f_Y^{re}(u, \omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V(u) \boldsymbol{\beta}} \right\}, \tag{7.2}$$

where $V(u)$ is the variance-covariance matrix of $\boldsymbol{Y}_{t,T}$ at time $u = t/T$, as defined in Dahlhaus (1999). We can define $\lambda(u, \omega)$ and $\boldsymbol{\beta}(u, \omega)$ to be the spectral envelope and the resulting optimal scaling, respectively, at time u and frequency ω. Although the time-varying spectral envelope $\lambda(u, \omega)$ could be estimated by using the windowing method for estimating $f_Y(u, \omega)$ in Dahlhaus (1999), as previously stated, this approach is not computationally efficient and is impractical for large data sets. We can, however, use the principles in Ombao *et al.* (2001) to devise a computationally efficient method for estimating the evolutionary spectral envelope. We expect that our estimators will be consistent under known segmentation. In addition, we will still need to address over-all consistency of our method given that the segmentation is usually not known and has to be selected using the data-driven BBA.

The dyadic segmentation framework is computationally efficient and provides a remedy to the problem of efficient estimation. It is in the tradition of the growing body of work used in regression and signal processing. Under the dyadic segmentation framework, consistent estimators for the Dahlhaus time-varying spectrum are formed when the segmentation is known. This result is given in Ombao *et al.* (2001). Thus, one conjecture that can be given at this point is that under known segmentation, one can also form a consistent estimator for the true spectral envelope if the evolving spectral envelope follows the same smoothness assumptions of the Dahlhaus evolving spectrum. The next step is to rigorously define that evolving spectral envelope. Moreover, the over-all

consistency still has to be addressed given that the segmentation has to be selected from the data.

REFERENCES

Adak, S. (1998). Time dependent spectral analysis of nonstationary time series, *J. Amer. Statist. Assoc.*, **93**, 1488–1501.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth and Brooks/Cole, Monterey, California.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*, 2nd ed., Holden-Day, Oakland, California.

Chiann, C. and Morettin, P. (1999). Estimation of time varying linear systems, *Statistical Inference for Stochastic Processes*, **2**, 253–285.

Coifman, R. and Wickerhauser, M. (1992). Entropy based algorithms for best basis selection, *IEEE Trans. Inform. Theory*, **32**, 712–718.

Cornette, J. L., Cease, K. B., Margaht, H., Spouge, J. L., Berzofsky, J. A. and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, *Journal of Molecular Biology*, **195**, 659–685.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes, *Ann. Statist.*, **25**, 1–37.

Dahlhaus, R. (1999). A likelihood approximation for locally stationary processes, Beiträge zur Statistik 56, Universität Heidelberg.

Donoho, D., Mallat, S. and von Sachs, R. (1998). Estimating covariances of locally stationary processes: Rates of convergence of best basis methods, Tech. Report 517, Department of Statistics, Stanford University.

Eaton, M. L. and Tyler, D. E. (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix, *Ann. Statist.*, **19**, 260–271.

Hannan, E. J. (1970). *Multiple Time Series*, Wiley and Sons, New York.

Ioshikhes, I., Bolshoy, A. and Trifonov, E. N. (1992). Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences, *Journal of Biomolecular Structure and Dynamics*, **9**, 1111–1117.

Karlin S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data, *J. Amer. Statist. Assoc.*, **86**, 27–35.

Mallat, S., Papanicolau, G. and Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes, *Ann. Statist.*, **26**, 1–17.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Ombao, H., Raz, J., von Sachs, R. and Malow, B. (2001). Automatic statistical analysis of bivariate nonstationary time series, *J. Amer. Statist. Assoc.*, **96**, 543–560.

Rosenblatt, M. (1959). Statistical analysis of stochastic processes with stationary residuals, *Probability and Statistics* (ed. U. Grenander), 246–275, Wiley, New York.

Satchwell, S. C., Drew, H. R. and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA, *Journal of Molecular Biology*, **191**, 659–675.

Stoffer, D. S., Tyler, D. E. and McDougall, A. J. (1993*a*). Spectral analysis for categorical time series: Scaling and the spectral envelope, *Biometrika*, **80**, 611–622.

Stoffer, D. S., Tyler, D. E., McDougall, A. J. and Schachtel, G. (1993*b*). Spectral analysis of DNA sequences (with discussion). *Bulletin of the International Statistical Institute*, Bk I, 345–361 (Discussion: ibid. (1994). Bk IV, 63–69).

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1997). Prediction of probable genes by fourier analysis of genomic sequences, *Computer Applications in the Biosciences*, **13**, 263–270.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors, *Ann. Statist.*, **9**, 725–736.

Voss, R. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, *Phys. Rev. Lett.*, **68**(25), 3805–3808.

Wickerhauser, M. (1994). *Adapted Wavelet Analysis from Theory to Software*, IEEE Press, Wellesley, Massachusetts.