Computing Linear Transforms of Symbolic Signals

Wei Wang and Don H. Johnson, *Fellow, IEEE*

***Abstract*—Signals that represent information may be classified into two forms: numeric and symbolic. Symbolic signals are discrete-time sequences that at, any particular index, have a value that is a member of a finite set of symbols. Set membership defines the only mathematical structure that symbolic sequences satisfy. Consequently, symbolic signals cannot be directly processed with existing signal processing algorithms designed for signals having values that are elements of a field (numeric signals) or a group. Generalizing an approach due to Stoffer, we extend time–frequency and time-scale analysis techniques to symbolic signals and describe a general linear approach to developing processing algorithms for symbolic signals. We illustrate our techniques by considering spectral and wavelet analyses of DNA sequences.**

***Index Terms*—DNA sequences, linear transforms, symbolic signal.**

I. Introduction

MUCH of signal processing has focused on analyzing numeric information. Real and complex-valued data satisfy the mathematical properties of a *field*. Arithmetic operations between numbers can subsequently be derived from the definition of a field [8]. Specifically, the values of most data form an ordered field. Signal processing techniques all rely on this mathematical structure.

Information is often written in the language of symbols. Each element in a sequence of symbols is a member of a set. Symbolic signals differ from numeric signals in that symbolic sets have no additional mathematical structure. A set of symbols is not a field because algebraic operations on symbols are usually not meaningful. For example, addition, multiplication, and numeric ordering cannot be performed on symbols. At the fundamental level, digital signals are composed from the set $\{0, 1\}$,[1] text files consist of a sequence of characters, and multiple neuron discharge patterns may be represented by symbols indicating the occurrence of action potentials [12]. The arbitrary assignment of a number to each symbol would impose a mathematical structure not present in the original data. Therefore, numeric signal processing algorithms cannot be readily applied to symbolic signals. The specific application we will take here is the genetic code. DNA sequences are built from the set of nucleotides represented by the symbols $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$, and sequence index indicates location along the genome. See [9] and [14] for engineering-oriented overviews. Despite the fact that we concentrate here on analyzing DNA sequences, they serve only as an example; we describe a general approach to the spectral and scale analysis of symbolic data.

Symbolic signals can have a rich statistical structure that is the focus of many signal processing algorithms. For example, stochastic symbolic signals are discrete random processes with an unknown amplitude distribution (probability mass function) and a correlation structure. The amplitude distribution can be easily estimated as a histogram and, thus, forms a type [6]. It is the correlation structure we want to elucidate here, primarily using frequency-domain or wavelet-domain analysis. Applying a transform technique requires mapping the symbolic domain into the numeric domain in such a way that no additional structure is placed on the symbolic sequence beyond that inherent to it. For example, one arbitrary map for the DNA sequence application would be to assign the alphabetically sorted nucleotides to an increasing sequence of integers: $\mathtt{A} \leftrightarrow 1, \mathtt{C} \leftrightarrow 2, \mathtt{G} \leftrightarrow 3, \mathtt{T} \leftrightarrow 4$. This mapping would suggest that one nucleotide is somehow greater than another, which is a property the symbolic set does not possess. A recent paper [2] uses a much better, but arbitrary assignment of nucleotide to number to calculate spectrograms of DNA sequences.

Approaches have been developed for processing signals within restricted mathematical domains. For example, wavelet transform algorithms have been defined over groups and finite fields [5], [18]. In the statistical literature, symbolic data are termed categorical data [1]. We use the terminology "symbolic signals" rather than "categorical time series." The latter term arises when data are drawn from categories, such as the kind of animals going into a vetenarian facility. Sequences of symbols arise in information theory, wherein information sources produce symbols (like letters of the alphabet) that need to be encoded into bit strings, which are themselves simple symbols. We find the term "symbolic signal" more appropriate from a signal processing viewpoint, but many symbolic algorithms go under the heading of categorical time series. Markov chains and link-function-based regression models have been examined for time-domain analysis of categorical time series. However, a methodology for the frequency domain analysis of symbolic signals is needed. Symbols have been associated with vectors so that spectral analysis could be performed [2], [7], but again, this assign-

Manuscript received April 20, 2000; revised November 16, 2001. The associate editor coordinating the review of this paper and approving it for publication was Prof. Abdelhak M. Zoubir.

W. Wang was with the Computer and Information Technology Institute, Department of Electrical and Computer Engineering, Rice University Houston, TX 77251-1892 USA. She is now with the University of California, Berkeley, CA 94720 USA.

D. H. Johnson is with the Computer and Information Technology Institute, Department of Electrical and Computer Engineering, Rice University, Houston, TX 77251-1892 (e-mail: dhj@rice.edu).

Publisher Item Identifier S 1053-587X(02)01326-0.

[1]In this paper, we use a different font to denote symbols such as $\{\mathtt{0}, \mathtt{1}\}$. Note how this differs from the numeric values $\{0, 1\}$.

ment is arbitrary. Some authors have exploited the complementary structure of DNA sequences [19]. Because A is paired with T and C with G on the parallel strands of the double helix, these pairs are grouped together to create a new set of two symbols $\{(\mathtt{A},\mathtt{T}),(\mathtt{C},\mathtt{G})\}$. We show later that such binary-symbol sequences pose no challenge in applying numeric signal processing techniques. Stoffer developed an approach for spectral analysis of categorical time series [16], [17] that we elaborate here. He proposed a mapping of a symbolic sequence to a numeric one in a nonarbitrary manner that emphasizes any periodic feature that might exist in the categorical process. However, he assumed that the categorical time series was stationary. We expanded Stoffer's approach to processing nonstationary symbolic signals in the frequency domain. We describe time-frequency and time-scale (wavelet) domain analysis techniques for symbolic-valued signals. Our approach to symbolic signal processing is sufficiently general to allow any linear transform of symbolic data to be calculated.

II. Methods

The procedure requires two parts: mapping the symbolic data to a numeric form in a nonarbitrary manner and calculating the transform of that numeric sequence. In passing from symbolic to numeric data (as shown in Fig. 1), the set of symbols $\chi = \{\mathtt{c}_1, \mathtt{c}_2, \ldots, \mathtt{c}_K\}$ is first mapped to a set of indicator vectors $\xi = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$. Each indicator vector $\mathbf{e}_i$ has dimension K equal to the size of the symbolic set, and it has a "1" in the row corresponding to the occurrence of a symbol and a "0" otherwise. The symbolic sequence $\mathtt{x}_n$ can thus be mapped to a sequence of indicator vectors $\mathbf{y}_n$ by replacing each symbol with its corresponding indicator vector $\mathbf{y}_n = \mathbf{e}_i$ when $\mathtt{x}_n = \mathtt{c}_i$. Symbol set indexing (and, thus, rows of the indicator vector) has no any significance in the data processing. Note that vectors of dimension greater than one cannot be ordered, and any nontrivial arithmetic operation on an indicator vector would not result in an indicator vector. Thus, the set of indicator vectors does not constitute a field, and the sequence $\mathbf{y}_n$ is also symbolic in nature. To obtain a numeric sequence z_n, we evaluate the inner product of each element of the sequence $\mathbf{y}_n$ with a weight vector $\mathbf{w}$ (dimension $(K \times 1)$). Determination of the weight vector is therefore critical. In the example shown in Fig. 1, the weighting is arbitrary and seems to imply that T > A and C > G. Our algorithm determines the weight vector according to signal processing criteria to maximally highlight structure in the data. In this way, the data and the type of analysis being performed determine the weight vector.

Assume for the moment that we have a weight vector and have converted the symbolic sequence into a discrete numeric sequence z_n to which we can apply time-frequency and time-scale analyses.We will discuss two transforms: the short-time Fourier transform and the wavelet transform. Developing a meaningful transform is the same for each; we use the Fourier transform to exemplify the process. Taking a discrete linear transform of the discrete-time numeric sequence

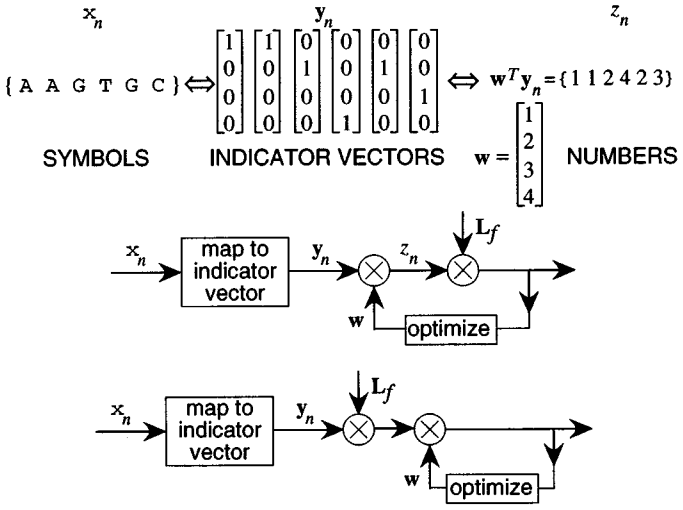

Fig. 1. First step is mapping the symbolic sequence to a sequence of indicator vectors. Using the genetic code as an example, $\mathtt{x}_n$ is the sequence consisting of the four symbols $\{\mathtt{A},\ \mathtt{G},\ \mathtt{C},\ \mathtt{T}\}$, making $\mathbf{y}_n$ a sequence of (4×1) indicator vectors. Here, we assigned indicator vectors according to col$[1\ 0\ 0\ 0] \leftrightarrow \mathtt{A}$, col$[0\ 1\ 0\ 0] \leftrightarrow \mathtt{G}$, col$[0\ 0\ 1\ 0] \leftrightarrow \mathtt{C}$, col$[0\ 0\ 0\ 1] \leftrightarrow \mathtt{T}$. The upper panel shows how a specific nucleotide sequence would be assigned an indicator vector sequence with respect to this choice. We used the arbitrarily selected weight vector $\mathbf{w} =$ col$[1\ 2\ 3\ 4]$, creating the depicted numeric sequence z_n. However, this weighting is arbitrary and seems to imply that $\mathtt{T} > \mathtt{A}$ and $\mathtt{C} > \mathtt{G}$. A better weight choice is to optimize some aspect of the transformed sequence according to some criterion. For linear transforms, which are represented in the middle and lower panels by a multiplication by $\mathbf{L}_f$, we can interchange the multiplications by $\mathbf{w}$ and $\mathbf{L}_f$ so that the weight-vector optimization can be performed on the transform of the indicator vector sequence.

z_n is computationally equivalent to a matrix multiplication (see Fig. 1). Because the symbolic-to-numeric mapping and the transform are linear operations, we can interchange their order. At any particular frequency f, we represent the column vector of transform coefficients by $\mathbf{L}_f$. Collecting a sequential set of N values of z_n starting at index m together into a *row* vector $\mathbf{z}_m = [z_m, z_{m+1}, \ldots, z_{m+N-1}]$, we want to transform this segment, advance the segment's time origin m by some amount, resegment, etc. The transform of the segmented numeric sequence can be expressed by $Z_{m,f} = \mathbf{z}_m \mathbf{L}_f$. Therefore, the Fourier transform of the numeric equivalent of the symbolic sequence can be written as

$$Z_{m,f} = \mathbf{z}_m \mathbf{L}_f = (\mathbf{w}^t \mathbf{Y}_m)\mathbf{L}_f = \mathbf{w}^t(\mathbf{Y}_m \mathbf{L}_f) = \mathbf{w}^t \mathcal{Y}_{m,f} \quad (1)$$

where $\mathbf{Y}_m$ denotes the matrix formed by collecting the sequence of indicator vectors together: $\mathbf{Y}_m = [\mathbf{y}_m, \ldots, \mathbf{y}_{m+N-1}]$. Thus, we transform the sequence of indicator vectors first, creating the vector $\mathcal{Y}_{m,f}$ at each frequency, and then apply the weight vector. The transform calculation thus becomes a multichannel Fourier transform, with the transform of each row of the indicator vector sequence $\mathcal{Y}_{m,f}$ computed separately. This reordering of applying the weight vector and transforming allows us to determine the weight vector *after* transforming the signal to its frequency representation. Thus, the weight vector can be optimized according to signal processing criteria at *each* analysis frequency [17]. In the time domain, no single weight vector results as the equation $z_n = \mathbf{w}^t \mathbf{y}_n$ would

suggest. Because we determine the weight vector according to the signal's frequency domain structure within a particular time window, an equivalent time-domain weight vector does not exist. Because of this data-dependent optimization, the weight vector depends on the data, making the calculation of the transform of symbolic sequences nonlinear.

[2] The wavelet transform does not involve windowing, but the idea of interchanging application of the weight vector and the wavelet transform operation still applies. Here, we optimize the weight vector according to scale and analysis time. For *any* linear transform, we optimize the weight vector according to the signal's structure in the transform's domain.

We used the optimization criterion of maximizing the energy of the transform output at each frequency. More specifically, we maximize the magnitude squared of the transform output in the transform domain $|\mathbf{w}^t\mathcal{Y}_{m,f}|^2 = \mathbf{w}^t\langle\mathcal{Y}_{m,f}\mathcal{Y}'_{m,f}\rangle\mathbf{w}$,[3] imposing the normalization constraint that $\|\mathbf{w}\|^2 = 1$. This constrained energy optimization corresponds to the unconstrained maximization of the Rayleigh quotient $(\mathbf{w}^t\langle\mathcal{Y}_{m,f}\mathcal{Y}'_{m,f}\rangle\mathbf{w})/(\mathbf{w}^t\mathbf{w})$, which is maximized when $\mathbf{w}$ is proportional to the eigenvector corresponding to the largest eigenvalue of $\langle\mathcal{Y}_{m,f}\mathcal{Y}'_{m,f}\rangle$ [11]. Thus, the weight vector depends on the analysis frequency f and on the time origin m about which the averaging operation took place. We indicate this dependence explicitly by $\mathbf{w}(m,f)$. The largest eigenvalue equals the power of the transform output at a particular frequency, and thus provides the power spectral values we seek.

Note that the indicator vector sequences $\mathbf{y}_n$ and $\mathbf{Y}_m$ are rank deficient by one. Any particular row is completely determined from the other rows because each column of the matrix can only hold a single 1 with the remaining entries equaling 0. Thus, we need only compute the transform of $(K-1)$ rows of the indicator vector sequence $\mathbf{Y}_m$, where K equals the number of symbols. For a two-symbol sequence, we need only compute the transform of one row of the matrix $\mathbf{Y}_m$, and either will do. Because the transforms considered here are linear, the selected row can be multiplied by any real number without affecting spectral structure. In this case, we can map from a two-symbol set to the real numbers arbitrarily without imposing an ordering. Assign one of the two symbols the value zero and the other any real number. No special algorithms are needed for this case.

We implemented this approach by developing Fourier and wavelet transforms for symbolic sequences. Because our data are nonstationary, we computed the short-time Fourier transform for time-frequency analysis of symbolic signals: $\mathrm{STFT}(m,f) = \sum_{n=m}^{m+N-1} z_n g_{n-m} e^{-j2\pi fn}$, where g_n represents a moving window [15]. We window each row of the matrix $\mathbf{Y}_m$ and compute the Fourier transform of each. At each frequency, we form the outer product $\mathcal{Y}_{m,f}\mathcal{Y}'_{m,f}$, creating the so-called cross-spectral matrix [13, ch. 14]. It is the eigenstructure of the cross-spectral matrix we seek. Because a matrix consisting of a single outer product has rank one, we average these outer products over several windowed segments to create $\langle\mathcal{Y}_{m,f}\mathcal{Y}'_{m,f}\rangle$. The largest eigenvalue of this matrix constitutes the power spectrum of the symbolic sequence at the time m and the frequency f. We use these values to create a spectrogram of the symbolic sequence.

Wavelet analysis has the desirable characteristic that it makes no stationarity assumptions on the signal it processes, and it has adaptive time-scale resolution. Since the usual wavelet transform steps along the time axis by the length of the wavelet basis, a short duration pattern in the data could easily be missed completely by the correlating wavelet basis if the two are not exactly in phase. We used the *redundant* wavelet transform because it is shift-invariant (no down-sampling occurs after highpass and lowpass filtering at each wavelet decomposition scale): $\mathrm{WT}(j,k) = \sum_n z_n 2^{-j/2}\psi(2^j(n-k))$. The choice of wavelet basis ψ further adapts the wavelet transform to the particular signal; we used the Haar basis because of its square-wave waveform [4]. It is worth noting that the Haar basis is the only appropriate basis for symbolic sequences because the sequence of indicator vectors is binary valued. We calculate the optimal weight vector independently at each scale of the wavelet decomposition by averaging across different length time windows according to the time-frequency resolution tiling of the wavelet transform. We then display the power of the time-scale content of the symbolic signal in a scalogram. Because the shift-invariant transform does not down sample, we normalize the scalogram display at increasing scales by dividing by powers of $\sqrt{2}$.

By displaying the short-time Fourier or wavelet transforms as spectrograms or scalograms, respectively, local structure can be visually detected. The frequency or scale of the display indicates the periodicity of the structure. Once structure is located, the weight vectors provide additional information about the structure of the pattern. Recall that the weight vectors are optimized to maximize transform power. Thus, the relative weights given to each symbol indicates the level of participation of each symbol in the detected pattern. Repeating symbols receive higher weight values, while nonrepeating symbols are suppressed. We found that by multiplying weight vector sequences by the spectrogram or scalogram as appropriate, we can create a display that highlights regions of periodicity and clearly shows the symbolic structure. Thus, for spectrograms we computed $\mathrm{STFT}(m,f_0)\cdot\mathbf{w}(m,f_0)$, what we call the weighted spectrogram, and displayed the components of the resulting vector as a function of segment time origin.

III. Results

We present as examples some spectrograms, scalograms, and weight vector displays of real and simulated DNA data and show how these displays provide information about the correlation

[2]The fact that the transform calculation is nonlinear does not obviate the manipulations given in (1), which depend on the properties of matrix multiplication.

[3]The notation $\mathbf{w}^t$ means the transpose of $\mathbf{w}$ and $\mathcal{Y}'_{m,f}$, which is the conjugate-transpose of $\mathcal{Y}_{m,f}$. The notation $\langle\cdot\rangle$ refers to an averaging window that is used to derive nontrivial solutions for the weight vector. We describe the reason behind the averaging subsequently.

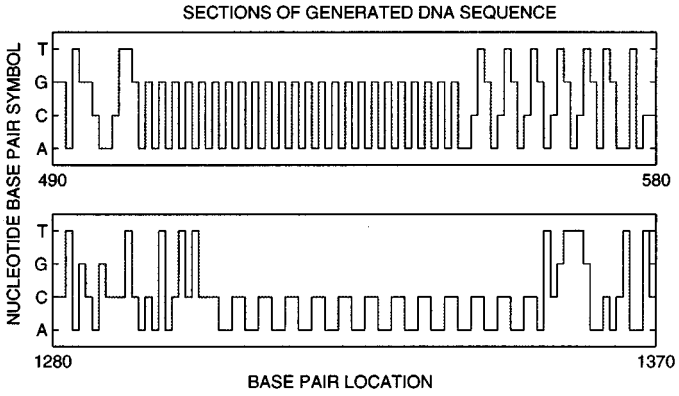

Fig. 2. We created a sequence of random noise (or randomly distributed {A, G, C, T} symbols) and interspersed short sections of patterns with varying periodicities and at varying intervals. One pattern we created, shown in the top panel, for our pseudo DNA sequence consisted of the period-2 pattern GAGA ... immediately followed by a period-4 pattern ACTGACTG .... In the bottom panel, we show the period-4 pattern AACC ⋯.

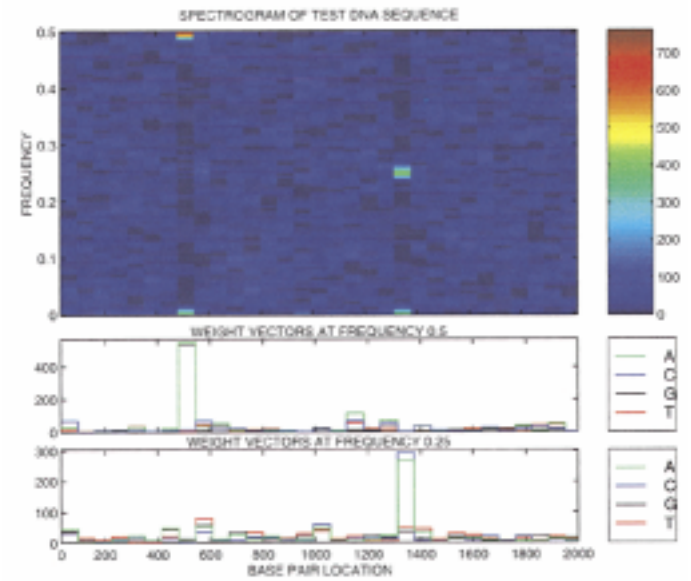

Fig. 3. Spectrogram displays the power of the symbolic short-time Fourier transform of the test DNA sequence. High power is found at frequencies of 0.5 and 0.25 at indexes of approximately 500 and 1300, respectively, accompanied by no power at all other frequencies at those locations except at zero frequency. These high-power regions indicate the detection of a pattern with a periodicity of 2 at base pair index around 500 and a pattern with a periodicity of 4 at around 1300. These regions and patterns indeed correspond to the patterns shown in Fig. 2. Weight vectors are calculated independently at each frequency of the Fourier transform of test DNA. The entries of the weight vector corresponding to each symbol are then multiplied by the power density spectrum to create the weighted spectrogram that displays what symbols formed the patterns that produced power in the short-time spectrum. The figure indicates that the first high-power burst around index 500 corresponds to a period–2 pattern of alternating A and G, followed by a period-4 pattern involving A, C, T, and G. The second high-power region at index 1300 corresponds to a period–4 pattern of A and C. Because no power is present at frequency 0.5, the pattern must be AACC. These regions and patterns correspond to the patterns shown in Fig. 2.

structure of the symbolic signals. Structure corresponds to regions of high energy in the spectrogram and scalogram. The sequence index locates the pattern while the highlighted frequency or scale indicates the periodicity of the structure: $T = 1/f_0$ or 2^j. Since we used the Haar basis in the wavelet transform, high

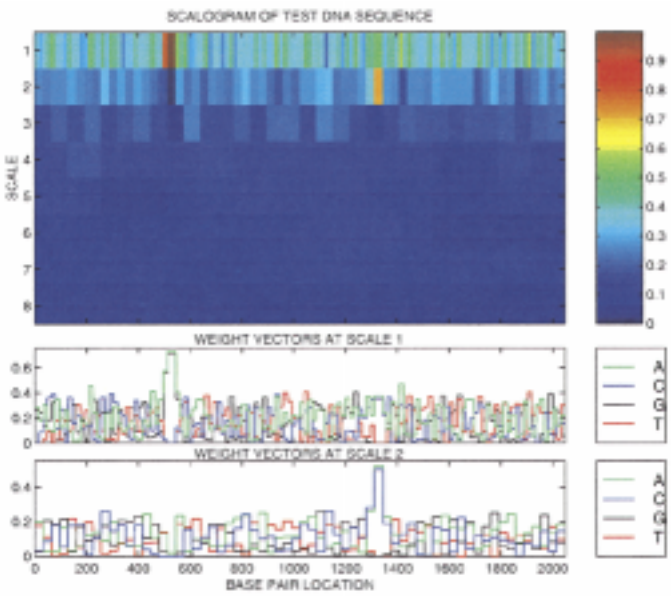

Fig. 4. Scalogram displays the power of the cross-spectral matrix of the shift-invariant wavelet transform of test DNA. We normalize at scale s by dividing by $2^{s/2}$. High power values are found at scale 1 at index location 500 and at scale 2 at index location 1300. Note that if we zoomed in, we can locate the pattern to the exact DNA base pair index because of the higher time index resolution of the wavelet transform. Since we used a Haar basis (with period 2), this corresponds to patterns of period 2 and 4 at scales 1 and 2, respectively. The weight vector is calculated independently at each scale of the shift-invariant wavelet transform using varying length averaging windows according to the desired time-scale resolution. The elements of the weight vector associated with each symbol are multiplied with the power of the wavelet transform to create the weighted scalogram. The display shows that the first pattern detected involves the symbols A and G, whereas the second pattern involves A and C. These regions and patterns correspond to the patterns shown in Fig. 2.

power at scale j indicates the occurence of a pattern with period 2^j. Conceptually, a symbolic pattern with periodicity L indicates that a symbol repeats at regular intervals of length L. An example of a period 2 pattern in the symbol A would be {A G A G A C A T}.

We generated some simple "pseudo DNA" data to show how the spectrogram and scalogram display temporal periodic structure. Fig. 2 shows the symbolic patterns used, whereas Figs. 3 and 4 display the spectrogram and scalogram of the pseudo DNA, respectively. The spectrogram and scalogram each detect the high energy sections of pattern with the frequencies $1/2$ and $1/4$, or scales 1 and 2. The weighted spectrogram and scalogram clearly identify the corresponding period 2 GA and period 4 ACTG and AACC patterns.

We applied the method to analyze E. *coli* and human hemoglobin DNA sequences. The spectrogram and weighted spectrogram displays of E. *coli* are given in Fig. 5. The DNA sequences corresponding to high-energy spectral components are shown in Fig. 6. The scalogram and weighted scalogram displays of human hemoglobin are given in Fig. 7, and Fig. 8 shows the sequence found in one high-energy segment. These displays allow visual detection and location of structure in symbolic sequences.

Generally, the wavelet transform gives finer time index resolution such that patterns can be exactly located, whereas the

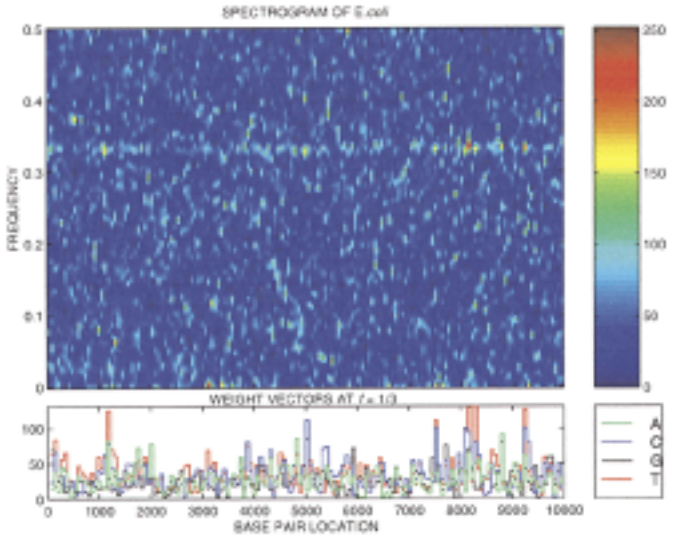

Fig. 5. The spectrogram of a 10 000 length section of E. *coli* DNA displays a faint line at a frequency of $1/3$ across the entire sequence, indicating a weak recurring pattern with period 3. The presence of this line indicates that DNA coding is occurring. The weighted spectrogram shown in the bottom panel displays the behavior of the individual symbols. For example, at index locations of around 1200 and 8200, the pattern seems dominated by T. There also appears to be structure at index 8000 involving C and T. Examination of these areas (Fig. 6) indeed shows that these symbols do repeat every third symbol.

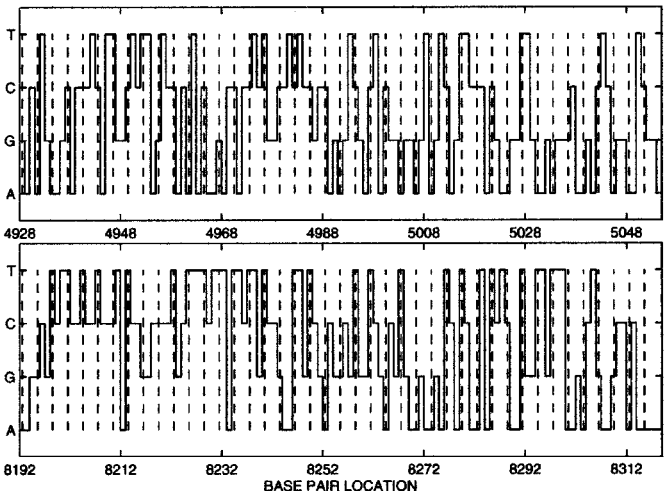

Fig. 6. Two base pair segments are shown around locations where a high degree of synchrony to a frequency of $1/3$ is found in the E. *coli* spectrogram. A frequency of $1/3$ corresponds to a symbol recurring every third position. The dashed line markers indicate precise synchrony of this nature. In the top panel, synchrony of the symbols C and G (the first in particular) is found; the bottom one reveals synchrony to T.

Fourier transform only indicates the region in which the pattern can be found. However, for lengthy symbolic sequences (on the order of tens of thousands), this high time index resolution may be too much information to be useful for visual detection. If a pattern has a periodicity that is not a power of 2, the wavelet transform energy will not "live" in a single scale but will spill across adjacent scales. This effect could be observed in the E. *coli* data that contained a structure with period 3. The Fourier transform does not have this limitation (see Fig. 5).

IV. Conclusion

Computing the power spectrum for numeric signals usually ends with a display of the spectrum. For symbolic signals, the spectrum is not enough. When we find a peak in the Fourier spectrum of a numeric signal, we know that the underlying signal at that point in time contained an additive sinusoidal component. A peak in a symbolic signal's spectrum merely indicates that one or several symbols are periodically repeating. (Note that the frequency at which symbols repeat must be rational.) Which symbols contribute to this repetition are not expressed by the spectrum. Furthermore, multiple peaks do not mean that several symbolic components are added; they cannot be because symbols cannot be added. Instead, the components must interleave, which means that some spectral combinations—periods that are mutually prime—cannot occur.

When periodicities do occur, we can determine which symbols contribute by considering the weight vector. Because the weight vector maximizes spectral power, those weight vector components that have the larger values must correspond to the symbols contributing to the periodicity and smaller ones do not. Thus, our symbolic spectral or wavelet algorithm has two phases: We compute the spectrum first and then reveal the weight vectors corresponding to interesting time intervals.

Fourier and wavelet methods may not elucidate structure in symbolic signals. For example, where does the word "symbol" occur in this paper, or where does the sequence ATAGCT occur in a DNA chain? Especially in the case of DNA sequences, interesting patterns occur in blocks separated by segments having a different pattern or none at all. Pattern matching algorithms that locate when a particular symbolic sequence occurs in a longer sequence would perform better. The Boyer-Moore algorithm [3], for example, gives good performance in such cases. In more complicated cases, symbols may not only be wrong (a situation well described by a stochastic model), but they can be missing. In signal processing terms, samples are missing without any indication of absence. Stochastic models for deletions or the sequence location problem may not be most appropriate, making signal-processing-oriented detection algorithms suspect. The transform techniques presented here would not work well as pattern-matching algorithms in such cases. On the other hand, when one wants to scan a sequence for periodic structure without specification of what symbols comprise the period, the method described here may be better suited than targeted algorithms. Another consideration in choosing algorithms is the computational complexity. While the transforms described here have low complexity, the optimization step amounts to finding the largest eigenvalue and the corresponding eigenvector of a $(K-1)\times(K-1)$ symmetric matrix at each frequency or scale. Thus, our approach's complexity increases with both number of symbols K and the number of frequencies. Furthermore, interesting patterns are not necessarily periodic. In such cases, Fourier and wavelet techniques may not work well.

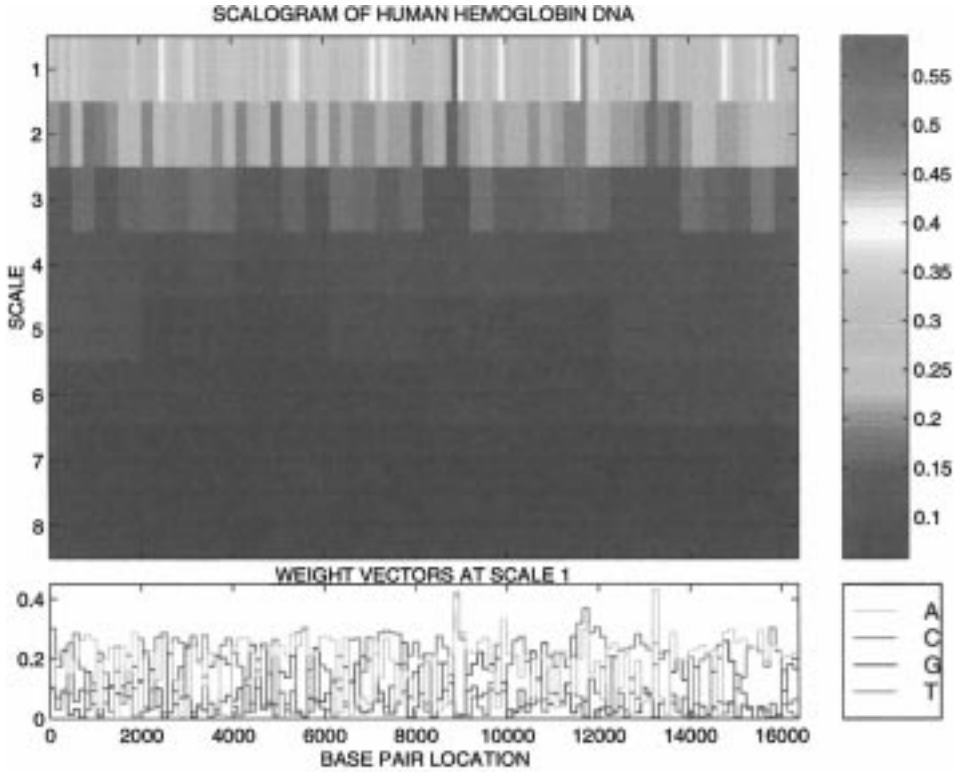

Fig. 7. Scalogram of a 16 000 length section of human hemoglobin DNA detects a period–2 pattern (scale 1) at index 8800. The weighted scalogram (see Fig. 8) shows that this pattern consists of ATAT . . .. There is a weaker period–2 pattern detected at index 13200, which seems to be comprised of interspersed AC and AT patterns.

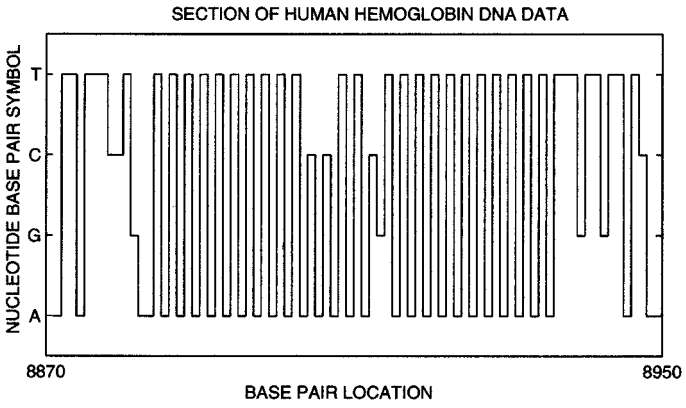

Fig. 8. Examination of a section of the human hemoglobin DNA sequence reveals the period–2 AT pattern detected in the weighted scalograms (Fig. 7).

It should be also noted that symbolic signals can be processed in many ways *without* using the techniques described here. In stochastic frameworks (for example, detection problems), the likelihood ratio can be calculated just as easily for symbolic as it can for numeric observations. The likelihood ratio depends solely on the data's hypothesized probability distributions, and these impose no ordering on the symbols and map the value of a random variable, be it symbolic or numeric, to a numeric value precisely in the way needed for optimal detection. Estimating the needed probability distribution is accomplished via histogram estimators (types). Because symbols are discrete, these estimates will converge to the true distribution; just as in any detection scenario, proper account of serial dependencies must be made to achieve optimal detection performance. Empirical detection algorithms [10], which have no hypothesized distribution and rely solely on training data, are tailored to symbolic or discrete real-valued random variables and asymptotically yield optimal performance. Thus, in the case of detection and other algorithms dependent on the probability function for data, our approach is not needed.

We have shown that signal processing algorithms requiring numeric signals can be extended to symbolic ones as well. We have used a linear, data-dependent, algorithm-dependent mapping $z_n = \mathbf{w}^t \mathbf{y}_n$ to pass from symbolic to numeric representations. Because the weight vector depends on frequency or scale, the weight vector does not exist as a single entity. Because of this data- and algorithm-dependent weight, this technique for converting from symbols to numbers does not impose an arbitrary structure. The generality of this approach allows other linear signal processing algorithms, such as filtering, to be developed. This paper's approach can be used in such cases to bridge between symbolic and numeric signals.

REFERENCES

[1] A. Agresti, *Categorical Data Analysis*. New York: Wiley, 1990.
[2] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Mag.*, vol. 18, pp. 8–20, 2001.
[3] R. S. Boyer and J. S. Moore, "A fast string matching algorithm," *Commun. ACM*, vol. 20, pp. 62–72, 1997.
[4] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*. Upper Saddle River, NJ: Prentice-Hall, 1998.

[5] G. Caire, R. L. Grossman, and H. V. Poor, "Wavelet transforms associated with finite cyclic groups," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1157–1166, 1993.
[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[7] E. Coward, "Equivalent of two Fourier methods for biological sequences," *J. Math. Biol.*, vol. 36, pp. 64–70, 1997.
[8] C. W. Curtis, *Linear Algebra*. New York: Springer-Verlag, 1984.
[9] J. P. Fitch and B. Sokhansanj, "Genomic engineering: Moving beyond DNA sequence to function," *Proc. IEEE*, vol. 88, pp. 1949–1971, 2000.
[10] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, pp. 401–408, 1989.
[11] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[12] D. H. Johnson and C. M. Gruner, "Neural ensemble processing with types," in *Proc. Computat. Neurosci.*, Bozeman, MT, 1997.
[13] S. M. Kay, *Fundamentals of Statistical Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[14] S. K. Moore, "Understanding the human genome," *IEEE Spectrum*, vol. 37, pp. 33–35, 2000.
[15] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1999.
[16] D. S. Stoffer and D. E. Tyler, "Matching sequences: Cross-spectral analysis of categorical time series," *Biometrika*, vol. 85, pp. 201–213, 1998.
[17] D. S. Stoffer, D. E. Tyler, and A. J. McDougall, "Spectral analysis for categorical time series: Scaling and the spectral envelope," *Biometrika*, vol. 80, pp. 611–622, 1993.
[18] M. D. Swanson and A. H. Tewfik, "A binary wavelet decomposition of binary images," *IEEE Trans. Image Processing*, vol. 5, pp. 1637–1650, 1996.
[19] R. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," *Phys. Rev. Lett.*, vol. 68, pp. 3805–3808, 1992.

Wei Wang received the B.S. degree in electrical and computer engineering from Rice University, Houston, TX, in 2000. She is now pursuing the Ph.D. degree at the University of California, Berkeley.

Her interests are in statistical signal processing.

Don H. Johnson (F'90) received the S.B. and S.M. degrees in 1970, the E.E. degree in 1971, and the Ph.D. degree in 1974, all in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He joined MIT Lincoln Laboratory as a Staff Member in 1974 to work on digital speech systems. In 1977, he joined the faculty of the Electrical and Computer Engineering Department at Rice University, Houston, TX, where he is currently the J.S. Abercrombie Professor and Chair of that department. At MIT and at Rice, he received several institution-wide teaching awards. His present research activities are in statistical signal processing. Particular areas of interest are non-Gaussian signal processing and the representation of information by neural signals.

Dr. Johnson is a recipient of the IEEE Millenium Medal and the Signal processing Society's Meritorious Service Award. He is a member of the Acoustical Society of America, the American Association for the Advancement of Science, and the Association for Research in Otolaryngology. He is former President of the IEEE Signal Processing Society, past chair of the IEEE Prize Paper Committee, and a member of Eta Kappa Nu and Tau Beta Pi. He is currently a member of the Signal Processing Society's Signal Processing Theory and Methods Technical Committee and a member of the ICASSP 2002 Organizing Committee.