BIOINFORMATICS

Vol. 16 no. 12 2000
Pages 1073–1081

Frequency-domain analysis of biomolecular sequences

Dimitris Anastassiou

Department of Electrical Engineering, Columbia University, 500 West 120th Street, Mail Code 4712, New York, NY 10027, USA

Received on April 18, 2000; revised on July 5, 2000; accepted on July 28, 2000

Abstract

Motivation: *Frequency-domain analysis of biomolecular sequences is hindered by their representation as strings of characters. If numerical values are assigned to each of these characters, then the resulting numerical sequences are readily amenable to digital signal processing.*

Results: *We introduce new computational and visual tools for biomolecular sequences analysis. In particular, we provide an optimization procedure improving upon traditional Fourier analysis performance in distinguishing coding from noncoding regions in DNA sequences. We also show that the phase of a properly defined Fourier transform is a powerful predictor of the reading frame of protein coding regions. Resulting color maps help in visually identifying not only the existence of protein coding areas for both DNA strands, but also the coding direction and the reading frame for each of the exons. Furthermore, we demonstrate that color spectrograms can visually provide, in the form of local 'texture', significant information about biomolecular sequences, thus facilitating understanding of local nature, structure and function.*

Availability: *All software for techniques described in this paper is available from the author upon request.*

Contact: *anastas@ee.columbia.edu*

Introduction

The main reason that the field of digital signal processing did not yet have significant impact on biomolecular sequence analysis is that the former refers to numerical sequences, while the latter refers to character strings. In this paper, we demonstrate that, by assigning proper (complex, in general) numerical values to each character, digital signal processing of biomolecular sequences provides a set of novel and useful tools. For example, we show that color spectrograms, like the one shown in Figure 2, visually provide information about the local nature of DNA stretches; and color maps, like the one shown in Figure 6 predicting the exon locations shown in Table 1, can identify exons, including their coding directions and reading frames.

Table 1. Locations and reading frames of the five exons of the gene F56F11.4

Relative position	Exon length	Reading frame
929–1135	207	2
2528–2857	330	2
4114–4377	264	1
5465–5644	180	2
7255–7605	351	1

For a DNA sequence of length N, assume that we assign the numbers a, c, t, g to the characters 'A', 'T', 'C', 'G', respectively. The resulting numerical sequence is:

$$x[n] = au_A[n] + tu_T[n] + cu_C[n] + gu_G[n], \quad n = 0, 1, 2, \ldots, N-1 \tag{1}$$

in which $u_A[n]$, $u_T[n]$, $u_C[n]$, and $u_G[n]$ are the *binary indicator sequences*, which take the value of either 1 or 0 at location n, depending on whether or not the corresponding character exists at location n (Voss, 1992).

Any three of these four binary indicator sequences are sufficient to determine the DNA character string, because

$$u_A[n] + u_T[n] + u_C[n] + u_G[n] = 1, \quad \text{for all } n. \tag{2}$$

A proper choice of the numbers a, t, c and g for a DNA segment can provide potentially useful properties to the numerical sequence $x[n]$. For example, if we choose complex conjugate pairs $t = a^*$ and $g = c^*$, then the complementary DNA strand is represented by:

$$\tilde{x}[n] = x^*[-n + N - 1], \qquad n = 0, 1, \ldots, N-1 \tag{3}$$

and, in that case, all 'palindromes' will yield 'conjugate symmetric' numerical sequences, which have interesting mathematical properties.

Systems and methods

The main computational tool that we use is the *Discrete Fourier Transform* (DFT) of a sequence $x[n]$ of length

N. The DFT is itself another sequence $X[k]$ of the same length N (Mitra, 2000; Oppenheim and Shaffer, 1999), defined by:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}, \qquad k = 0, 1, \ldots, N-1. \quad (4)$$

The sequence $X[k]$ provides a measure of the frequency content at 'frequency' k, which corresponds to an underlying 'period' of $\frac{N}{k}$ samples. It turns out that, except for finite length effects that can be corrected (Oppenheim and Shaffer, 1999), the square of the magnitude of the DFT is also a scaled version of the DFT of the autocorrelation sequence.

From equations (1) and (4) it follows that:

$$X[k] = aU_A[k] + tU_T[k] + cU_C[k] + gU_G[k], \quad k = 0, 1, \ldots, N-1. \quad (5)$$

For pure DNA character strings (i.e. without assigning numerical values), the sequences $U_A[k]$, $U_T[k]$, $U_C[k]$, and $U_G[k]$ provide a four-dimensional representation of the 'frequency spectrum' of the character string. The quantity:

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 \quad (6)$$

has been used as a measure of the total spectral content of the DNA character string, at 'frequency' k (Tiwari *et al.*, 1997; Silverman and Linsker, 1986; Li *et al.*, 1994).

From equations (2) and (4) it follows that:

$$U_A[k] + U_T[k] + U_C[k] + U_G[k] = \begin{cases} 0, & k \neq 0 \\ N, & k = 0. \end{cases} \quad (7)$$

Therefore, we can reduce the 'dimensionality' of the frequency spectrum representation from four to three, e.g. by ignoring one of the four frequency components. If we wish to reduce dimensionality in a symmetric manner with respect to all four components, we may adopt the technique (Silverman and Linsker, 1986) in which three numerical sequences x_r, x_x, and x_b are defined from corresponding coefficients (a_r, t_r, c_r, g_r), (a_g, t_g, c_g, g_g), (a_b, t_b, c_b, g_b) by considering the four three-dimensional vectors having magnitude equal to 1 and pointing to the four directions from the center to the vertices of a regular tetrahedron. For example, we can choose $(a_r, a_g, a_b) = (0, 0, 1)$, $(t_r, t_g, t_b) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right)$, $(c_r, c_g, c_b) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$, $(g_r, g_g, g_b) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right)$, hence:

$$\begin{aligned} x_r[n] &= \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]) \\ x_g[n] &= \frac{\sqrt{6}}{3}(u_C[n] - u_G[n]) \\ x_b[n] &= \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n] \end{aligned} \quad (8)$$

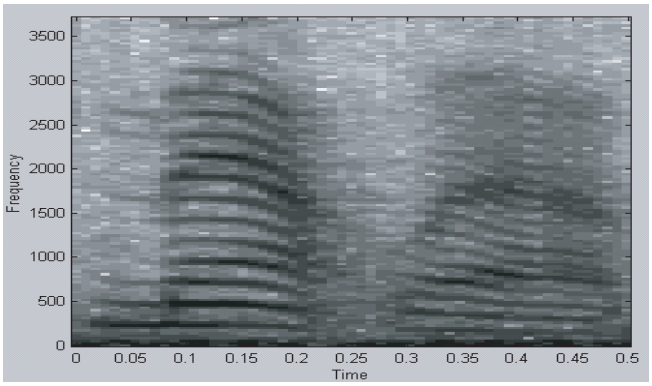

Fig. 1. Spectrogram of a speech signal.

from which we can find the DFTs $X_r[k]$, $X_g[k]$, $X_b[k]$.

We can apply a 'sliding window' of small length to a sequence, resulting in a 'sequence of DFTs', each providing a 'localized' measure of the frequency content. This is known as the *Short-Time Fourier Transform* (STFT). The display of the magnitude of the STFT is called a *spectrogram*, and it has long been used in the analysis of speech signals. For example, Figure 1 shows a spectrogram, created using MATLAB, corresponding to a sampled speech signal as a time-frequency diagram, in which image intensity is proportional to the corresponding STFT coefficient. The appearence of the spectrogram provides significant information, to the extent that trained observers can figure out the words uttered in voice signals.

Algorithms and implementation

DNA spectrograms

We introduce spectrograms of biomolecular sequences that simultaneously provide local frequency information for all frequencies and for all four bases. To achieve the latter, we reduce dimensionality form four to three (retaining all information content) in a symmetrical manner using equation (8), and we display the resulting three magnitudes by superposition of the corresponding three primary colors, red for x_r, green for x_g and blue for x_b. Thus, color conveys real information, as opposed to 'pseudocolor spectrograms', in which color is used for contrast enhancement. For example, Figure 2 shows a spectrogram using DFTs of length 60 of a DNA stretch of 4000 nucleotides from chromosome III of *C.elegans* (GenBank accession number NC 000967).

The vertical axis corresponds to the 'frequencies' k from 1 to 30, while the horizontal axis shows the relative nucleotide locations, starting from nucleotide 858 001. The DNA stretch contains three regions ('*C.elegans* telomere-like hexamer repeats') at relative locations (953–1066), (1668–1727), and (1807–2028). These three regions are

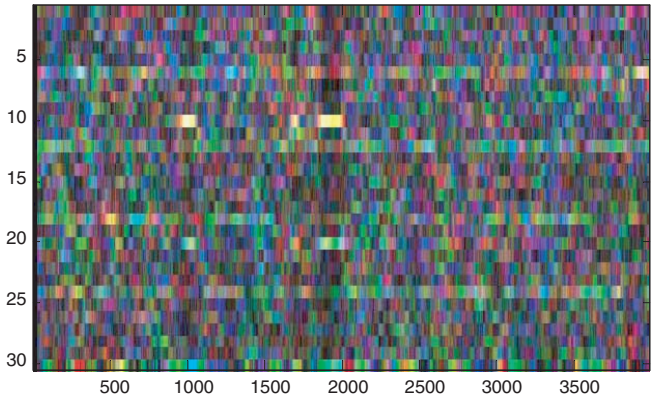

Fig. 2. Color spectrogram of a DNA stretch.

well depicted as bars of high-intensity values corresponding to the particular frequency $k = 10$ (because period 6 corresponds to $\frac{N}{6} = 10$). Other frequencies also appear to play a prominent, role in the whole region of the 4000 nucleotides.

In other examples, we have noted the presence of other periodicities (including 10–11 periodicities) and features, indicative of structural patterns. Such periodicities have been observed before (Ioshikhes *et al.*, 1996; Widom, 1996; Herzel *et al.*, 1999; Ioshikhes *et al.*, 1999; Stein and Bina, 1999; Trifonov, 1998).

In developing spectrograms, we may use 'tapered windows', in which central elements are assigned higher 'weights' compared with elements at the ends of the windowed subsequences. The width and the shape of the windowing operation play important roles in the appearance of spectrograms. Furthermore, spectrograms can be defined using the wavelet transform, rather than the DTFT. The wavelet transform has been used to analyze some fractal scaling properties of DNA sequences (Arneodo *et al.*, 1995).

Identification of protein coding DNA regions

The 'frequency' $k = \frac{N}{3}$ corresponds to a period of three samples, equal to the length of each codon. It is known (Fickett, 1982; Chechetkin and Turygin, 1995) that the spectrum of protein coding DNA typically has a peak at that frequency. This property has been used (Tiwari *et al.*, 1997) to design a gene prediction algorithm. If we define the following normalized DFT coefficients at frequency $k = \frac{N}{3}$:

$$W = \frac{1}{N} X\left[\frac{N}{3}\right]$$
$$A = \frac{1}{N} U_A\left[\frac{N}{3}\right], \quad T = \frac{1}{N} U_T\left[\frac{N}{3}\right], \tag{9}$$

$$C = \frac{1}{N} U_C\left[\frac{N}{3}\right], \quad G = \frac{1}{N} U_G\left[\frac{N}{3}\right]$$

then it follows from equation (5), with $k = \frac{N}{3}$, that:

$$W = aA + tT + cC + gG. \tag{10}$$

In other words, for each DNA segment of length N (where N is a multiple of 3), and for each choice of the parameters a, t, c and g, there corresponds a complex number $W = aA + tT + cC + gG$. We have found that, for properly chosen values of a, t, c and g, W can be an accurate predictor of not only whether or not the DNA segment is part of a protein coding region, but also, in the former case, in which reading frame it belongs, the latter information coming from the phase $\Theta = \arg\{W\}$.

For each DNA segment, there corresponds a set of complex numbers A, T, C and G, as defined in equation (9), in which $A + C + T + G = 0$, because of equation (7). These quantities can be thought of as complex random variables. They have quite different probabilistic characteristics depending on whether or not the DNA sequence is part of a protein coding region, as well as on the corresponding reading frame. Under this interpretation, the quantity W, as defined in equation (10), is a complex random variable itself, and its properties depend on the particular choice of the parameters a, t, c, and g.

To quantify the statistical properties of the random variables A, T, C, and G for protein coding regions, we arbitrarily chose chromosome XVI of *S.cerevisiae* (GenBank accession number NC 001148). We isolated all genes for which there were no introns, and for which the 'evidence' was labeled 'experimental'. If the orientation of a gene was 'complementary', then we properly transformed its values as if it were a 'forward coding' gene (i.e. starting from the codon ATG). For each of the chosen genes, we evaluated the corresponding numbers A, T, C, and G, thus creating a set of statistical samples. We found that, for that particular chromosome, the average values of A, T, C, and G, scaled by 10^3, were $8.0 - 56.3j$, $-84.1 + 37.4j$, $-46.2 - 23.2j$, and $122.3 + 42.1j$. By comparison, the magnitudes of A, T, C, and G, for nonprotein-coding regions are much smaller, typically between 1 and 2.

There have been many proposed 'protein coding measures' used for gene identification (Fickett and Tung, 1992; Claverie, 1997). In this paper, we predict whether or not a given DNA segment belongs to a reading frame from the magnitude of a properly defined W, i.e. after optimizing the values of the parameters a, t, c, and g.

We wish to maximize the 'discriminatory capability' between protein coding regions (with corresponding random variables A, T, C, and G) and 'random DNA' regions. Using a random number generator, we synthesized a random

DNA sequence, with the same number and length as the protein coding statistical sample, thus creating a different set of random variables: A_R, T_R, C_R, and G_R.

Because of the fact that $A + T + C + G = 0$, if we add any constant value to the coefficients a, t, c, and g, then the value of W in equation (10) remains the same. In order to define an optimization problem with a unique solution, we first fix one of the four coefficients (c) to the value of 0, so that $W = aA + tT + cC + gG$, and $c = 0$. (We could have reduced dimensionality in a symmetrical manner, but this would not have enhanced predictive power.)

Therefore, the following problem is naturally formulated once we have available a joint probabilistic model for the complex random variables A, T, and G (in our case coming from our measurements from chromosome XVI of *S.cerevisiae*) and for the complex random variables A_R, T_R, and G_R:

> Find the complex numbers a, t, and g maximizing the quantity:
>
> $$p(a,t,g) = \frac{E\{|aA + tT + gG|\} - E\{|aA_R + tT_R + gG_R|\}|}{\text{std}(|aA + tT + gG|\} + \text{std}(|aA_R + tT_R + gG_R|)}$$
>
> (in which std stands for standard deviation)

under the constraining conditions (because W is also invariant to rotation and scaling):

$$E\{\arg\{aA + tT + gG\}\} = 0$$
$$|a| + |t| + |g| = 1.$$

The above mathematical problems (and similar ones defined below) can potentially be solved yielding some closed-form solution as a function of certain statistical coefficients. However, there is no need for this, because conventional optimization techniques, based on iterated random perturbations starting from an initial guess, immediately converge to the optimum values.

Using the resulting random variables, we found the solution:

$$a = 0.10 + 0.12j \quad t = -0.30 - 0.20j$$
$$c = 0 \quad g = 0.45 - 0.19j \qquad (11)$$

corresponding to a value of $p(a, t, g) = 2.18$.

Using the coefficients in equation (11), we evaluated the magnitude of the 351-point STFT for a DNA stretch of *C.elegans* (GenBank accession number AF 099922), containing 8000 nucleotides starting from location 7021. The plot of its square is shown in Figure 3. The DNA stretch contains a gene (F56F11.4) with five exons, all identified by the peaks of the plot, at the following positions, relative to 7021:

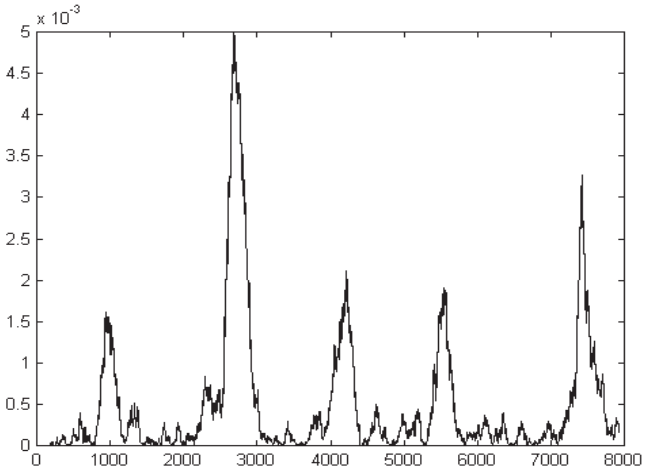

Fig. 3. Plot of $|aA + tT + cC + gG|^2$ for the five exons shown in Table 1.

Comparison with the traditional spectral content measure

We now compare the 'optimized' spectral content measure $|aA + tT + cC + gG|^2$ with the 'traditional' one from equation (6), $|A|^2 + |T|^2 + |C|^2 + |G|^2$, (or equivalently, $|aA + tT + cC + gG|$ and $\sqrt{|A|^2 + |T|^2 + |C|^2 + |G|^2}$) in terms of their capabilities to distinguish between coding and noncoding regions in DNA sequences. As mentioned previously, we used all single-exon genes with 'experimental evidence' from chromosome XVI of *S.cerevisiae* to collect the statistics from which the optimized values of a, t, c, and g were found.

We used these optimized values to compare the two measures on the set of single-exon genes with 'experimental evidence' from all remaining 15 chromosomes of *S.cerevisiae*, thus avoiding any overlap. The number of such genes from these 15 chromosomes was quite large (3088), sufficient to create a rendition for the estimated probability density functions (pdfs) shown in Figure 4. The pdfs for the optimized spectral content measure are drawn in red color; those for the traditional spectral content measure are drawn in green color.

For comparison purpose, we generated an equal number of 'nongenes' of the same length with the corresponding genes and totally random nucleotide distribution. The pdfs for 'nongenes' are drawn with dotted lines. All four curves have identical integrals, estimating probability density functions.

Because the traditional and optimized measures correspond to different units, we did not label the horizontal axis, but we scaled one of these pairs of pdfs so that the intersection point of the red curves is vertically aligned with the intersection point of the green curves. We note that the

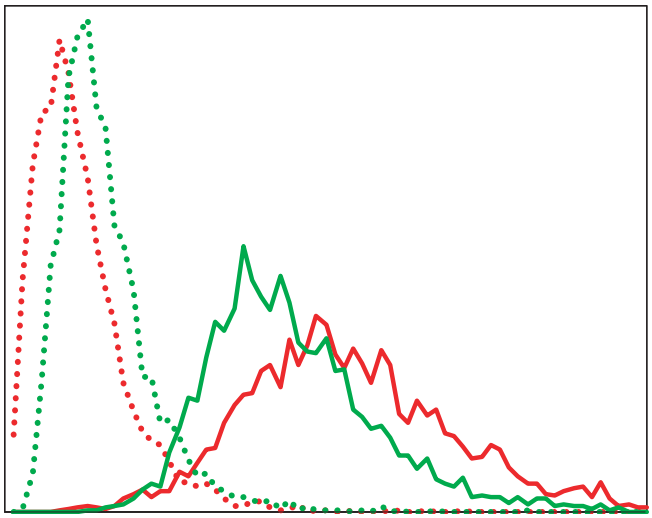

Fig. 4. Probability density functions for optimized (red) and traditional (green) spectral content measures. Solid lines indicate genes, dotted lines indicate 'nongenes'.

value of the former is less than half the value of the latter, that the solid red curve is to the right of the solid green curve, and that the dotted red curve is to the left of the dotted green curve. This indicates that the optimized measure improves upon the traditional measure because the 'distance' between the red graphs is larger than the 'distance' between the green graphs.

The improved performance of the optimized measure is also demonstrated numerically in Table 2, showing that the scores $p(a, t, g)$ of the optimized measure are higher compared to the traditional one, in each of the 15 chromosomes.

We note that, in addition to improved performance, the optimized measure may be computationally simpler than the traditional one, because it only requires the computation of one Fourier Transform of the sequence as in equation (1); the traditional technique requires the computation of four Fourier Transforms and of the sum of the squares of their magnitudes.

Reading frame identification

We label the three reading frames by the number $\mathrm{mod}(n, 3) + 1$ where n is the leftmost location of any codon triplet. According to this definition, the reading frames corresponding to each of the five exons of the gene are given in Table 1.

In order to distinguish forward-coding reading frames from reverse-coding ones, we augment the notation by including a 'tilde' on the numerical value assigned to a reverse-coding reading frame. For example, the bases at locations (0, 1, 2) form codons at either 'reading frame 1' or 'reading frame $\tilde{1}$' depending on whether the codon is 'forward-coding' or 'reverse-coding', respectively. Similarly, the bases at locations (1, 2, 3) form codons at either 'reading frame 2' or 'reading frame $\tilde{2}$', and the bases at locations (2, 3, 4) form codons at either 'reading frame 3' or 'reading frame $\tilde{3}$'.

Table 2. Comparison of scores using optimized and traditional measures of spectral content in the first 15 chromosomes of *S.Cerevisiae*

	Number of genes	Score of optimized measure	Score of traditional measure
1	56	1.82	1.37
2	256	1.81	1.52
3	87	1.71	1.20
4	444	1.72	1.32
5	163	1.84	1.56
6	67	1.79	1.36
7	306	1.73	1.44
8	147	1.56	1.30
9	113	1.93	1.54
10	209	1.76	1.37
11	187	1.73	1.41
12	277	1.68	1.41
13	255	1.78	1.44
14	219	1.80	1.44
15	302	1.67	1.34

It has been known (Shepherd, 1981) that the different reading frames exhibit different statistical characteristics. In this paper, we use the phase Θ of the complex random variable W as the reading frame predictor, making use of the following fact, which can be proved from the DFT definition:

> Assuming that a DNA segment is part of a forward coding region (reverse coding will be addressed later), we define the angles ϕ_1, ϕ_2, and ϕ_3 to be expected values of the phase of the random variable $W = aA + tT + cC + gG$ corresponding to the reading frames 1, 2, and 3, respectively. Then, $\mathrm{mod}(\phi_2 - \phi_1) = \mathrm{mod}(\phi_3 - \phi_2) = \mathrm{mod}(\phi_1 - \phi_3) = -\frac{2\pi}{3}$.

If, for example, $\phi_1 = 34°$, then $\phi_3 = 154°$ and $\phi_2 = 274°$, or, equivalently, $\phi_2 = -86°$. Therefore:

> If, for a particular choice of the parameters a, t, c, and g, the phase Θ of the complex random variable W has small variance, then the angle Θ will probably take values that will be close to one out of three possible ones, ϕ_1, ϕ_2, and ϕ_3, differing from each other by 120°.

In order to maximize predictive power, it is desirable to select the parameters a, c, t, and g minimizing some

measure of the 'variability' (such as the statistical variance) of Θ.

The definition of a unique meaningful statistical variance of the phase of a complex random variable is complicated by the fact that the phase is not uniquely specified unless restricted to an interval of length 2π, in which case the two ends of the interval correspond to equivalent values. Therefore, we have chosen instead the almost equivalent task of maximizing the magnitude of the expected value of the random variable $\frac{W}{|W|} = e^{j\Theta} = \cos\Theta + j\sin\Theta$. We would like that number to be as large, and as close to 1, as possible, because if it is only slightly less than 1, this will imply that $e^{j\Theta}$ is 'concentrated' on a tiny area in the periphery of the unit circle in the complex plane.

As in the previous optimization problem, we reduce the dimensionality of the problem by setting $c = 0$, and we formulate the following optimization problem:

> Find the complex numbers a, t, and g maximizing the quantity:
>
> $$q(a,t,g) = \left| E\left\{ \frac{aA + tT + gG}{|aA + tT + gG|} \right\} \right| \quad (12)$$
>
> under the constraining conditions (for unique solution):
>
> $$E\{\arg\{aA + tT + gG\}\} = 0$$
> $$|a| + |t| + |g| = 1.$$

The solution of the optimization problem, for our data, is given by:

$$a = 0.26 + 0.10j \qquad t = 0.03 - 0.17j$$
$$c = 0 \qquad\qquad g = 0.51 - 0.21j \qquad (13)$$

corresponding to a value of $q(a,t,g)$, as defined in equation (12), equal to 0.952, and to a standard deviation of the phase $\Theta = \arg\{aA + tT + gG\}$ of 18.2°. This means that the probability that Θ will be within, say, two standard deviations (36.4°) from the mean (0°) is very high.

All statistical data were collected under reading frame 1, and in that case the value of $E\{\Theta\}$ is 0°. Therefore, the value of Θ for reading frame 1 will be within $0° \pm 36.4°$, with high probability. Therefore, as explained above, if the data were corresponding to reading frame 2, then the value of Θ would have been within $-120° \pm 36.4°$, with high probability. Similarly, if the data were corresponding to reading frame 3, then the value of Θ would have been within $120° \pm 36.4°$, with high probability. There is still a significant 'gap' between any two of those angular regions.

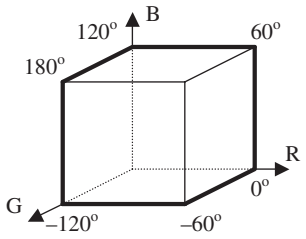

Fig. 5. Color coding of the Fourier Transform phase.

Table 3. Color-coded reading frame identification

RED	READING FRAME 1
GREEN	READING FRAME 2
BLUE	READING FRAME 3

Color-coding

Because the number of primary colors (red, green and blue) is the same as the number of possible forward coding reading frames, we have conveniently assigned a color-coding scheme, in which the value $\Theta = 0°$ is assigned the color 'red', the value $\Theta = 120°$ is assigned the color 'blue', and the value $\Theta = -120°$ is assigned the color 'green'. In-between values are color-coded in a linear manner, according to Figure 5, in which the three axes, labeled R, G, and B, correspond to the primary colors red, green, and blue, respectively.

Color maps

We use the above color coding for reading frame identification, according to Table 3. All STFT windows must be aligned at the same reading frame. Therefore, we have chosen for the sliding window to 'slide' by precisely three locations for each DFT evaluation. Furthermore, we always make sure that the window size is a multiple of 3, so that the frequency $k = \frac{N}{3}$ is well defined.

Figure 3 identifies the five exons, based on the magnitude of the STFT using the parameter values of equation (11). We now use the parameter values of equation (13) to enrich the information of Figure 3 in the form of a color map shown in Figure 6. For each nucleotide location in the color map, the color assigned obeys the rule of Figure 5, and the intensity is modulated by the square-magnitude, multiplied by 700 and clipped to the interval (0, 1).

Note that the 'color' of the third exon is closer to orange than to pure red, but the information is still sufficient to accurately identify its reading frame to be 1.

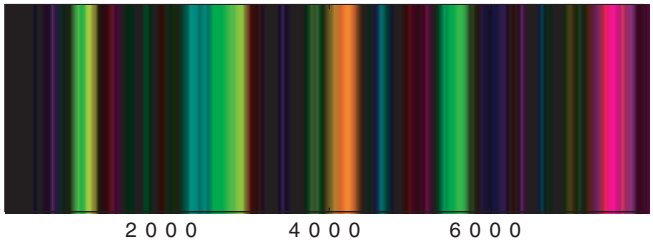

Fig. 6. Color map of reading frames for the exons of the gene of Table 1.

Complementary sequences

The binary indicator sequences of the complementary DNA strand are:

$$\tilde{u}_A[n] = u_T[-n+N-1]$$
$$\tilde{u}_T[n] = u_A[-n+N-1]$$
$$\tilde{u}_C[n] = u_G[-n+N-1]$$
$$\tilde{u}_G[n] = u_C[-n+N-1]$$

in which $u_A[n]$, $u_T[n]$, $u_C[n]$, and $u_G[n]$, $n = 0, 1, \ldots, N-1$ are the binary indicator sequences of the corresponding primary DNA strand.

We make use of the following fact, which can readily be proved from the DFT definition in equation (4):

> If two sequences $x[n]$ and $\tilde{x}[n]$ are related to each other by equation (3), i.e. if $\tilde{x}[n] = x^*[-n+N-1]$, then
>
> $$\tilde{X}\left[\frac{N}{3}\right] = e^{j\frac{2\pi}{3}} X^*\left[\frac{N}{3}\right] \qquad (14)$$

We now find the values of A, T, C, and G, as defined in equation (9), for the numerical sequence of the complementary DNA strand, for which we will use the notation $\tilde{A}, \tilde{T}, \tilde{C}$, and $\tilde{G}$, respectively. If we use the same choice for a, t, c, and g for both strands, it follows from equations (9) and (14) that:

$$\tilde{A} = e^{j\frac{2\pi}{3}} T^*, \tilde{T} = e^{j\frac{2\pi}{3}} A^*, \tilde{C} = e^{j\frac{2\pi}{3}} G^*, \tilde{G} = e^{j\frac{2\pi}{3}} C^* \qquad (15)$$

in which A, T, C, and G are the corresponding values of the primary strand. The value for $\tilde{W} = \hat{X}[\frac{N}{3}]$ is:

$$\tilde{W} = a\tilde{A} + t\tilde{T} + c\tilde{C} + g\tilde{G}. \qquad (16)$$

Now, if we define:

$$\tilde{a} = e^{-j\frac{2\pi}{3}} t^*, \tilde{t} = e^{-j\frac{2\pi}{3}} a^*, \tilde{c} = e^{-j\frac{2\pi}{3}} g^*, \tilde{g} = e^{-j\frac{2\pi}{3}} c^* \qquad (17)$$

Table 4. Locations and reading frames of six genes

Relative location	Gene length	Reading frame
761 → 1429	669	2
1687 → 3135	1449	1
3387 → 4931	1545	3
5066 ← 6757	1692	$\tilde{2}$
7147 ← 9918	2772	$\tilde{1}$
10143 ← 10919	777	$\tilde{3}$

then from equations (15), (16), and (17) it follows that $\tilde{W} = (\tilde{a}A + \tilde{t}T + \tilde{c}C + \tilde{g}G)^*$.

In conclusion: we can simulate the processing of the complementary strand in the reverse direction with parameters values a, t, c, and g, by processing the primary strand using the values of the parameters $\tilde{a}, \tilde{t}, \tilde{c}, \tilde{g}$ given by equation (17), and taking the complex conjugate of the resulting W.

It can easily be shown that the identical color code for reading frame identification, shown in Table 3, applies to reading frames $\tilde{1}, \tilde{2}$, and $\tilde{3}$ as well.

Example

We have used a DNA stretch from chromosome III of *S.cerevisiae* (GenBank accession number NC 001135). Note that there is no overlap with the collected statistics. The DNA stretch contains 12 000 nucleotides starting from location 212 041. It contains six genes (three forward coding and three reverse coding) at the locations shown in Table 4, relative to 212 040.

The major problem is that the color map for forward coding will contain some 'interference' from reverse coding regions, and vice versa (recall that the parameters a, t, c, and g were optimized to distinguish forward coding regions from noncoding regions, and not from reverse coding regions). One way of solving this problem is to partition the DNA segment into possible forward coding regions and possible reverse coding regions (this approach will fail to detect simultaneously multiple coding areas, but these are rare occasions in most organisms).

Because of equation (17), the following optimization problem is defined.

> Find the parameters a, t, c, and g maximizing the expected value of the following random variable:
>
> $$V = \left| \frac{aA + tT + cC + gG}{t^*A + a^*T + g^*C + c^*G} \right|.$$

To assure unique solution, we may simply pose the constraints $c = 0$ and $g = 1$, in which case we found

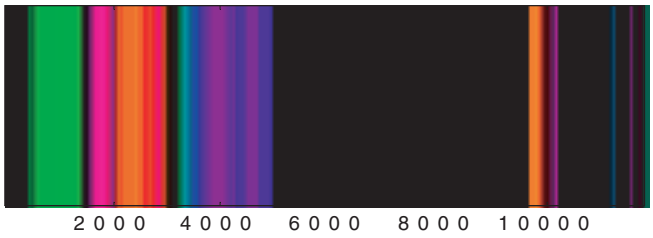

Fig. 7. Color map for forward coding after partition for the genes shown in Table 4.

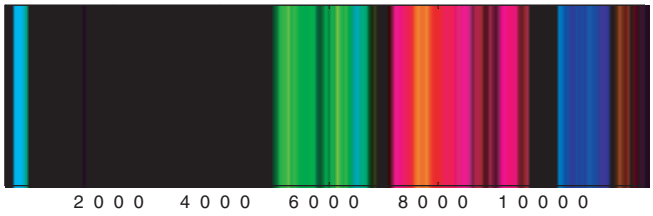

Fig. 8. Color map for reverse coding after partition for the genes shown in Table 4.

the optimal values $a = 0.049 + 0.149\,j$ and $t = -0.122 - 0.518\,j$. The criterion for partitioning is, then, whether or not V is greater or less than 1.

The resulting partitioning between forward and reverse coding is another unique feature of our proposed approach, compared with existing Fourier analysis tools. Figures 7 and 8 show the resulting color maps for forward and reverse coding. Comparing with Table 3, we see that the six genes were accurately color-coded, and we can obtain a sense of the power and the limitations of these tools.

In Anastassiou (2000), we provide more details and propose a more sophisticated 'soft partitioning' scheme by estimating the probabilities $P(F/V)$ and $P(R/V)$ that a particular location belongs to a forward (reverse) coding region, given that it belongs to either one, and given the value of V for that location.

Discussion

The 'frequency-domain' tools introduced in this paper are meant to synergistically complement 'character-string-domain' tools. For example, they can be of help in predicting precise splice sites in multiexonic genes in gene prediction programs, because reading frame identification excludes several potential splice sites due to the requirement that consecutive exons are 'in sync'.

Regarding visualization, there are several 'character-string-domain' tools for bimolecular sequences (Hamori and Ruskin, 1983; Mizrahi and Ninio, 1985; Cebrat and Dudek, 1998), including 'dot plots' and various alignment visualization tools. This is not true, however, for the frequency-domain, where the only used visualization tool, to our knowledge, is a pseudocolor rendition of the values of the traditional spectral content measure of equation (6) evaluated at $k = \frac{N}{3}$. Such a bar is included, for example, in the journal inset map recently resulted from the Genome Annotation Assessment Project (GASP) for *Drosophila melanogaster* (Reese *et al.*, 2000).

The techniques presented in this paper provide improved frequency-domain visualization tools, with use of optimized spectral content measures, and color-coded reading frame identification from the phase of Fourier Transforms.

The color maps presented here were based on parameter values that resulted from the collection of statistical data exclusively from chromosome XVI of *S.cerevisiae*, as an example. Such statistics, and the resulting optimized parameter values, are expected to vary a cross species and gene types. The assignment of optimized complex numerical values to nucleotides and amino acids provides a new computational framework (Anastassiou, 2000), which may also result in new computational techniques for the solution of useful problems in bioinformatics, including sequence alignment, macromolecular structure analysis and phylogeny (Durbin *et al.*, 1998).

An important advantage of the proposed tools is their flexibility. Spectrograms can be defined in many ways. For example, depending on the particular features that we wish to emphasize, we may wish to define spectrograms using certain values of the parameters a, t, c, and g. Once a visual pattern appears to exist, we have the opportunity to interactively modify the values of these parameters in ways that will enhance the appearance of these patterns, thus clarifying their significance. It is hoped that visual inspection of spectrograms will establish links between particular visual features (like areas with peculiar texture or color) and certain yet undiscovered motifs of biological sequences.

Acknowledgements

Appreciation is expressed to Ramakrishna Ramaswamy for helpful discussions of his gene prediction technique (Tiwari *et al.*, 1997). Much of the work in this paper was inspired by his results.

References

Anastassiou,D. (2000) Digital signal processing of biomolecular sequences. Technical Report EE000420-1, http://www.ee.columbia.edu/cgi-ee-bin/show_archive.pl.

Arneodo,A., Bacry,E., Graves,P.V. and Muzy,J.F. (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.*, **74**, 3293–3296.

Cebrat,S. and Dudek,M.R. (1998) The effect of DNA phase structure on DNA walks. *Eur. Phys. J.*, **3**, 271–276.

Claverie,J.-M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, **6**, 1735–1744.

Chechetkin,V.R. and Turygin,A.Y. (1995) Size-dependence of three-periodicity and long-range correlations in DNA sequences. *Phys. Lett.*, A **199**, 75–80.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.

Fickett,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.

Hamori,E. and Ruskin,J. (1983) H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.*, **258**, 1318–1327.

Herzel,H. and Große,I. (1995) Measuring correlations in symbol sequences. *Physica*, A **216**, 518–542.

Herzel,H., Weiss,O. and Trifonov,E.N. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and protein folding. *Bioinformatics*, **15**, 187–193.

Ioshikhes,I., Bolshoy,A., Derenshteyn,K., Borodovsky,M. and Trifonov,E.N. (1996) Nucleosomal DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, **262**, 129–139.

Ioshikhes,I., Trifonov,E.N. and Zhang,M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl. Acad. Sci. USA*, **96**, 2891–2895.

Li,W., Marr,T.G. and Kaneko,K. (1994) Understanding long-range correlations in DNA sequences. *Physica*, D **75**, 392–416.

Mitra,S.K. (2000) *Digital Signal Processing: A Computer-Based Approach*. 2nd edn, McGraw-Hill, New York.

Mizrahi,E. and Ninio,J. (1985) Graphical coding of nucleic acid sequences. *Biochimie*, **67**, 445–448.

Oppenheim,A.V. and Shaffer,R.W. (1999) *Discrete-Time Signal Processing*. 2nd edn, Prentice-Hall, Upper Saddle River, NY.

Reese,M.G., Hartzell,N.L., Harris,U., Ohler,J.F., Abril,J.F. and Lewis,S.-E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.

Shepherd,J.C. W. (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. USA*, **78**, 1596–1600.

Silverman,B.D. and Linsker,R. (1986) A measure of DNA periodicity. *J. Theor. Biol.*, **118**, 295–300.

Stein,A. and Bina,M. (1999) A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res.*, **27**, 848–753.

Tiwari,S., Ramachandran,S., Bhattacharya,A., Bhattacharya,S. and Ramaswamy,R. (1997) Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS*, **113**, 263–270.

Trifonov,E.N. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica*, A **249**, 511–516.

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

Widom,J. (1996) Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.*, **259**, 579–588.