Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 49(63), Fascicola 2, 2004

Analysis of biomolecular sequences through spectral based methods

Şerban Mereuţă [1]

Abstract – In this paper we apply a few computational and visual tools, specific to digital signal processing, to the analysis of biomolecular sequences. In particular, we prove that color spectrograms can help in visually identifying protein coding areas of the DNA strand and provide, in the form of local "texture", significant information about biomolecular sequences, thus facilitating the understanding of local nature, structure and function. We also show that the magnitude of a properly defined function in the spectral domain can be a predictor for the existence of protein coding regions in DNA sequences.

Keywords: DNA spectrograms, frequency-domain analysis, genome analysis.

I. INTRODUCTION

The mathematical treatment of macromolecular biological sequences corresponding to chains of nucleotides or amino acids is usually done by considering such sequences to be strings of characters like "*A*", "*T*", "*C*" and "*G*". If, however, we assign a numerical value to each of these abstract characters, then such sequences become numerical and amenable to digital signal processing. In this paper, we demonstrate that digital signal processing of numerical biomolecular sequences can provide a set of novel and useful tools in the field of bioinformatics.

In the case of DNA segments, assume that we assign the number a to the character "*A*", the number t to the character "*T*", the number c to the character "*C*", and the number g to the character "*G*". In general, a, t, c and g can be complex numbers.

The numerical sequence resulting from a character string of length N can be written as:

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

where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are the *binary indicator sequences*, which take on the value of either 1 or 0 at location n, depending on whether the corresponding character exists or not, respectively, at location n [1]. For example, the string ACCTG has $N = 5$, $u_A[0] = 1$, $u_T[0] = 0$, $u_C[2] = 1$ and $u_C[3] = 0$.

The four binary indicator sequences uniquely determine the character string corresponding to a DNA segment. For each n, three of the four sequences take the value of 0 and one takes the value of 1. They are a redundant, linearly dependent set of sequences because, for all n,

$$u_A[n] + u_T[n] + u_C[n] + u_G[n] = 1. \tag{2}$$

A. *Discrete Fourier Transform*

The Discrete Fourier Transform (DFT) of a sequence $x[n]$, of length N, is itself another sequence $X[k]$, of the same length N:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi k}{N}n}, \quad k = 0,1,\ldots,N-1 \tag{3}$$

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, where the maximum frequency (period 2) corresponds to $k = \frac{N}{2}$, assuming that N is even.

Using the definition in (3), the resulting sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ are the DFTs of the binary indicator sequences $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$, respectively.

If we assign numerical values a, t, c, and g, then from (1) and (3) it follows that:

$$X[k] = aU_A[k] + tU_T[k] + cU_C[k] + gU_G[k], \quad k = 0,1,\ldots,N-1 \tag{4}$$

In the case of pure DNA character strings (i.e., without assigning numerical values), the sequences

[1] Facultatea de Electronică şi Telecomunicaţii, Catedra de Telecomunicaţii
Bd. Carol I nr.11, 700506, Iaşi, e-mail: smereuta@etc.tuiasi.ro

$U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ provide a four-dimensional representation of the "frequency spectrum" of the character string. The quantity:

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 \quad (5)$$

can be used as a measure of the total power spectral content of the DNA character string at "frequency" k.

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

$$U_A[k] + U_T[k] + U_C[k] + U_G[k] = \begin{cases} 0, & k \neq 0 \\ N, & k = 0 \end{cases} \quad (6)$$

Therefore, we can reduce the "dimensionality" of the frequency spectrum representation from four to three, reflecting the same property of the binary indicator sequences [2]. One way to do this is simply to ignore one of the four frequency components. If we wish to reduce the dimensionality in a manner that is symmetric with respect to all four components, we may adopt the technique [3], in which three numerical sequences x_r, x_g, and x_b are defined from the 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), in the following way: the four three-dimensional vectors have magnitude equal to 1 and point to the four directions from the center to the vertices of a regular tetrahedron, corresponding to the four DNA bases.

For example, we can choose:

$(a_r, a_g, a_b) = (0,0,1)$, $(t_r, t_g, t_b) = (\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3})$,

$(c_r, c_g, c_b) = (-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3})$. $(g_r, g_g, g_b) =$

$= (-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3})$, resulting in

$$\begin{aligned} x_r[n] &= \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]) \\ x_g[n] &= \frac{\sqrt{6}}{3}(u_C[n] - u_G[n]) \\ x_b[n] &= \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n]) \end{aligned} \quad , \quad (7)$$

from which we can find three DFTs: $X_r[k]$, $X_g[k]$ and $X_b[k]$

B. *Short Time Fourier Transform*

Instead of evaluating the DFT of a full-length sequence, we have the option of evaluating the DFTs of several of its subsequences. This strategy makes sense particularly in the case of long sequences consisting of several segments with different characteristics.

For example, we may apply a "sliding window" of length L to a sequence of length N, where $N > L$, resulting in a "sequence of DFTs". Each of these DFTs provides a "localized" measure of the frequency content, and is an example of a location-dependent Fourier transform, known as the *short-time Fourier transform* (STFT).

II. DNA SPECTROGRAMS

The display of the magnitude of the STFT is called a *spectrogram* and it has long been used in the analysis of speech signals. The appearance of the spectrogram visually provides significant information to a trained observer about the local nature of the sound signal.

Similarly, we can use spectrograms to visually provide information about the localized frequency content of biomolecular sequences. In this case, to maximize the information content of the spectrogram, we can use color-coding to display the magnitudes of all individual sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ simultaneously, rather than merely displaying the overall magnitude as given in equation (5).

In this case, it is preferable to reduce the dimensionality from four to three (retaining all information content) using equations (7). We can then create one colored spectrogram combining all three spectrograms of the corresponding magnitudes by color-coding red for $|X_r[k]|$, green for $|X_g[k]|$, and blue for $|X_b[k]|$.

For example, Fig. 1 shows the spectrogram using DFTs of length $N = 60$. The data come from a DNA stretch of 4000 nucleotides from chromosome III of *C. elegans* (GenBank Accession number NC 000967 - http://www.ncbi.nlm.nih.gov/entrez).

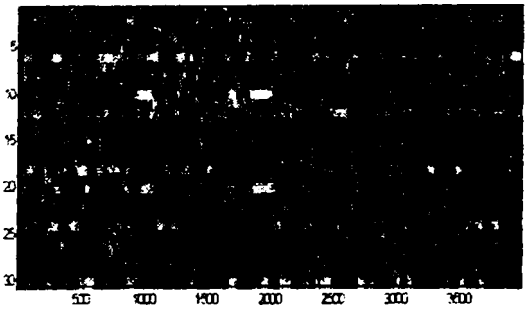

Fig. 1. Color spectrogram of a DNA stretch

The vertical axis corresponds to the "frequencies" k from 1 to 30, while the horizontal axis shows the relative nucleotide locations, starting from nucleotide 858001. The genomic annotations establish that the DNA stretch contains three regions ("*C. elegans* telomere-like hexamer repeats") at relative locations (953-1066), (1668-1727), and (1807-2028) [4]. These three regions are well depicted as bars of high-intensity values corresponding to the particular

frequency $k = 10$ (because hexamers -period 6- correspond to $\frac{N}{6} = 10$). Furthermore, the frequencies $k = 6$ (corresponding to a periodicity of 10) and its multiples, appear to play a prominent role in the whole region of the 4000 nucleotides.

For comparison purposes, Fig. 2 shows the "texture" of a spectrogram coming from a sample of "totally random" DNA, i.e., in which each type of nucleotide appears with probability 0.25 and independent of the other nucleotides.

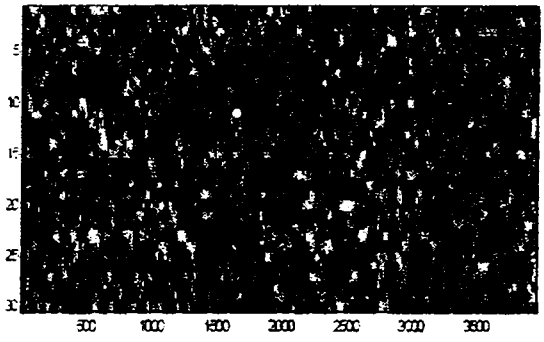

Fig. 2. Color spectrogram of "totally random" DNA.

III. PROTEIN CODING DNA REGIONS

Protein synthesis is governed by the genetic code which maps each of the 64 possible triplets (codons) of DNA characters into one of the 20 possible amino acids (or into a punctuation mark, like a stop codon, signaling termination of protein synthesis).

One of the most relevant and yet unsolved problems in bioinformatics is to accurately and automatically annotate sequences by identifying such regions using gene prediction [5], [6]. It is clear [7] that the total number of nucleotides in the protein coding area of a gene will be a multiple of three.

The "frequency" $k = \frac{N}{3}$ is of particular importance for protein coding DNA regions because it corresponds to a period of three samples, equal to the length of each codon (triplet of nucleotides).

We now show how frequency-domain analysis of DNA sequences can be a powerful tool for specifically identifying protein coding regions in DNA sequences. In Fig. 3 we have plotted the sequence $S[k]$, as defined in (5), for a coding region of length $N = 1320$ inside the genome of the baker's yeast (formally known as *S. cerevisiae*), demonstrating a peak at frequency $k = 440$ $(= N/3)$. This peak confirms the genetic findings reported for *S. cerevisiae* [4].

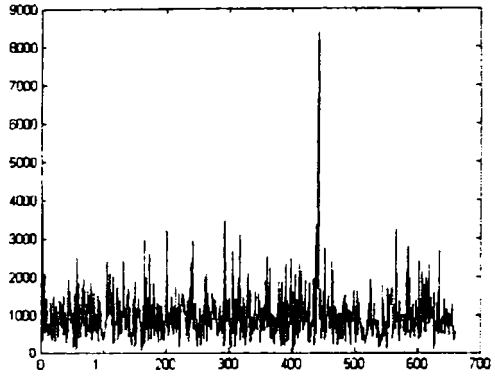

Fig. 3. Plot of the spectrum of a coding DNA region, demonstrating peak at frequency $k = N/3$.

REFERENCES

[1] R. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", *Physical Review Letters*, vol. 68(25), p. 3805-3808, 1992

[2] W. Li, T.G. Marr, K. Kaneko, "Understanding long-range correlations in DNA sequences", *Physica D.*, vol. 75, p. 392-416, 1994.

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

[4] http://www.ncbi.nlm.nih.gov/entrez

[5] J.-M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences", *Human Molecular Genetics*, vol. 6(10), p. 1735-1744, 1997.

[6] J.W. Fickett, "Recognition of protein coding regions in DNA sequences", *Nucleic Acids Research*, vol. 10, p. 5303-5318, 1982.

[7] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Essential Cell Biology*, New York, Garland Publishing, 1998.