Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 51(65), Fascicola 2, 2006

Spectral analysis for detecting protein coding regions based on a new numerical representation of DNA

Şerban Mereuţă [1]

Abstract – The major signal in coding regions of genomic sequences has a three-base periodicity. By proposing a new numerical representation for the DNA chain, our aim is to use spectral analysis for recognizing the coding regions of a gene. Since the peak at *f*=1/3 in the Fourier spectrum is a good discriminator of the coding potential of an intronless DNA strand, we utilized this feature within a sliding window in order to detect probable exons in a DNA sequence. Our technique is independent of training sets or existing database information, and thus can find general application.

Keywords: genomic signal processing, spectral analysis, exon detection.

I. INTRODUCTION

A single strand of DNA is a biomolecule consisting of many linked, smaller components called nucleotides. Each nucleotide (base) is one of four possible types designated by the letters *A*, *G*, *C* and *T* and has two distinct ends, the 5' end and the 3' end, so that the 5' end of a nucleotide is linked to the 3' end of another nucleotide by a strong chemical bond, thus forming a long, one-dimensional chain (backbone) of a specific directionality. Therefore, each DNA single strand is mathematically represented by a character string which, by convention, specifies the 5' to 3' direction when read from left to right. The double helix DNA is formed together with a complementary strand by linking *A* with *T* and vice versa, and *C* with *G* and vice versa.

The worldwide genome sequencing triggered the necessity of developing new approaches to rapidly assess the potential of a given DNA sequence. In this context, the gene identification problem through computational means is of great interest [1]-[5], [10]. But accurate gene prediction becomes complicated because of the fact that, in advanced organisms, protein coding regions in DNA are typically separated into several isolated subregions called *exons*. The regions between two successive exons are called *introns*, and they are eliminated before protein coding through a process called *splicing*.

In this paper, we investigate a spectral analysis technique based on a distinctive feature of protein coding regions of DNA sequences, i.e., the existence of short-range correlations in the nucleotide arrangement. The most prominent of these is a 3-base periodicity, which has been shown to be present in coding sequences [1], [3], [6], [10]. The signature of this periodicity (and any other) can be seen most directly, through the Fourier analysis, as a spectral peak [7].

II. DNA AS BINARY CODE INDICATOR SEQUENCES

In order to apply the techniques specific to digital signal processing, the DNA symbolic form given in the public genomic databases [8] must be represented by numerical sequences. This symbolic – numeric mapping must be done in such a manner that it doesn't distort the properties of the original DNA sequence, nor it introduces noise-like artifacts [3], [6], [7], [9].

In this paper, we propose a new numerical representation of the nucleotidic chain to be analyzed. This representation preserves the properties of the original genomic sequence and opens the possibility of an information theoretic approach, based on the source coding nature of the resulting binary string.

First, we start by collecting the statistics of the DNA sequence and compute the occurrence probabilities of the nucleotides *A*, *G*, *C* and *T*. Arranging the symbols in ascending order of probability, we assign binary code (00, 01, 10, 11) to the four bases.

For example, given the DNA sequence:

5' – *C-C-G-A-C-A-T-T-C-A* – 3',

the occurrence probabilities are $p(A)=0.3$, $p(G)=0.1$, $p(C)=0.4$ and $p(T)=0.2$. Hence, the corresponding binary code is $G\to 00$, $T\to 01$, $A\to 10$ and $C\to 11$.

In general, considering a sequence of *N* nucleotides, the numerical sequence attached can be written as:

$$x[n]=b_A[n]+b_G[n]+b_C[n]+b_T[n], \quad n=0,1,2,...,N-1 \tag{1}$$

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

where $b_A[n]$, $b_G[n]$, $b_C[n]$ and $b_T[n]$ are the *binary code indicator sequences*, which either have or haven't the code assigned to a specific nucleotide, depending on whether the corresponding character exists or not, respectively, at location *n*.

For example, in Table 1 we show the four binary code indicator sequences of a part of the previous DNA stretch. For computational reasons, in the implementation of our algorithm, we considered a level representation of the binary code.

Table 1

	C	*G*	*A*	*C*	*A*	*T*
$b_A[n]$	0 0	0 0	1 -1	0 0	1 -1	0 0
$b_G[n]$	0 0	-1-1	0 0	0 0	0 0	0 0
$b_C[n]$	1 1	0 0	0 0	1 1	0 0	0 0
$b_T[n]$	0 0	0 0	0 0	0 0	0 0	-1 1

In this manner, any DNA character string becomes a numerical sequence, having at each location *n* the code assigned to that particular nucleotide with respect to the corresponding occurrence probability.

III. ALGORITHM

There have been numerous proposed "protein coding measures" used for gene identification [1], [2], [3], [10]. In this paper, we predict whether or not a given DNA segment is a coding one using a similar methodology [3], from the magnitude of a properly defined spectral measure. We start by presenting the main tools of our algorithm.

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*:

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi \frac{k}{N} n}, \quad k = 0,1,2,\ldots,N-1 \tag{2}$$

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 (2), the resulting sequences $B_A[k]$, $B_G[k]$, $B_C[k]$ and $B_T[k]$ are the DFTs of the binary code indicator sequences $b_A[n]$, $b_G[n]$, $b_C[n]$ and $b_T[n]$, respectively.

From (1) and (2) it follows that:

$$X[k] = B_A[k] + B_G[k] + B_C[k] + B_T[k], \quad k = 0,1,2,\ldots,N-1. \tag{3}$$

Supposedly that instead of the binary code indicator sequences introduced in section II, we consider only the indicator sequences [7] – $u_A[n]$, $u_G[n]$, $u_C[n]$ and $u_T[n]$. These sequences take on the value of either 1 or 0 at location *n*, depending on whether the corresponding nucleotide exists or not at that location. Then, in the case of pure DNA character strings (i.e., without assigning numerical values), the resulting DFTs, $U_A[k]$, $U_G[k]$, $U_C[k]$ and $U_T[k]$, provide a four-dimensional representation of the "frequency spectrum" of the character string. The quantity

$$S[k] = |U_A[k]|^2 + |U_G[k]|^2 + |U_C[k]|^2 + |U_T[k]|^2 \tag{4}$$

has been used as a measure of the total power spectral content of the DNA character string, at "frequency" *k* [6], [10]. The DFT frequency $k = \frac{N}{3}$ corresponds to a period of three samples. It is known [1], [7], [10] that the spectrum of protein coding DNA typically has a peak at that frequency. For example, in Fig. 1 we have plotted the sequence $S[k]$, as defined in (4), for a coding region of length $N = 1320$ inside the genome of the baker's yeast (formally known as *Saccharomyces Cerevisiae*), demonstrating a peak at frequency $k = 440$ $(= N/3)$.

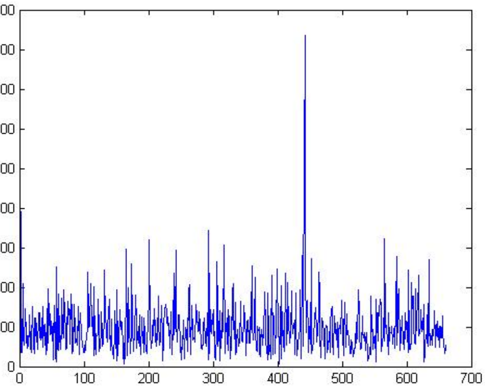

Fig. 1. Plot of the spectrum of a coding DNA region of length *N*, demonstrating peak at $k = N/3$.

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).

Implementation

Taking into consideration [3] and the previously presented tools for spectral analysis, if we define the following normalized DFT coefficients at frequency $k = \frac{N}{3}$:

$$W_{\frac{N}{3}} = \frac{1}{N} X\left[\frac{N}{3}\right]$$

$$A_{\frac{N}{3}} = \frac{1}{N} B_A\left[\frac{N}{3}\right], \quad G_{\frac{N}{3}} = \frac{1}{N} B_G\left[\frac{N}{3}\right], \qquad (5)$$

$$C_{\frac{N}{3}} = \frac{1}{N} B_C\left[\frac{N}{3}\right], \quad T_{\frac{N}{3}} = \frac{1}{N} B_T\left[\frac{N}{3}\right],$$

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

$$W_{\frac{N}{3}} = A_{\frac{N}{3}} + G_{\frac{N}{3}} + C_{\frac{N}{3}} + T_{\frac{N}{3}}. \qquad (6)$$

In other words, for each DNA segment of length N (where N is a multiple of three) it corresponds a complex number $W_{\frac{N}{3}}$.

We evaluated the magnitude of the 351-point STFT ($L = 351$) for a DNA stretch of *Caenorhabditis Elegans* (obtained by searching database [8] under 'Nucleotide', with the accession number AF099922), which contains 8040 nucleotides starting from location 7021. Inside this segment, the gene F56F11.4 is present, having five exons, with the positions relative to 7021 as in Table 2.

Table 2

Exon #	Relative position	Exon length
I	929 – 1135	207
II	2528 – 2857	330
III	4114 – 4377	264
IV	5465 – 5644	180
V	7255 – 7605	351

By collecting the statistics of the DNA sequence, it results the following occurrence probabilities of the nucleotides and their corresponding binary codewords: $p(C) = 0.157 \rightarrow 00$, $p(G) = 0.162 \rightarrow 01$, $p(T) = 0.337 \rightarrow 10$, $p(A) = 0.344 \rightarrow 11$.

In Fig. 2 it is shown the square magnitude $\left|W_{\frac{N}{3}}\right|^2 = \left|A_{\frac{N}{3}} + G_{\frac{N}{3}} + C_{\frac{N}{3}} + T_{\frac{N}{3}}\right|^2$, and all the five exons of the gene F56F11.4 are identified by the peaks of the plot at the positions shown in Table 2.

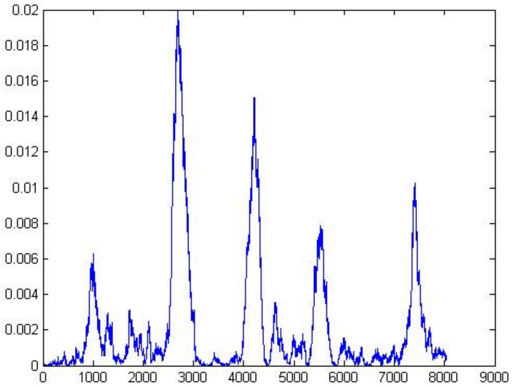

Fig. 2. Plot of $\left|W_{\frac{N}{3}}\right|^2$ for the five exons shown in Table 2

For comparison purposes, in Fig. 3 we represent the plot of $S[k]$ as defined in (4). This plot also proves that the performance of the spectral content measure $\left|W_{\frac{N}{3}}\right|^2$ is significantly superior to that of the one proposed by Tiwari et al. [10]. A demonstration of this fact was presented in [3], but in that case, for the numerical representation of DNA, Anastassiou used optimized values based on a training set of nucleotidic strings.

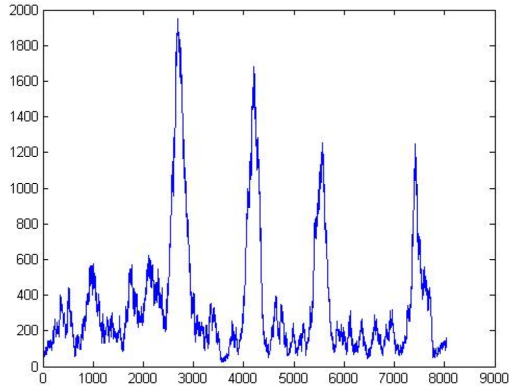

Fig. 3. Plot of $S[k]$ as defined in (4) and proposed in [10]

As it can be seen, the representation of $S[k]$ fails in detecting the first exon.

IV. CONCLUSIONS

In this paper, we presented a spectral analysis method for detecting DNA coding regions. We introduced a new numerical representation for the DNA stretch, by assigning binary codewords according to the occurrence probabilities of the nucleotides.

The advantage of our method is that it is independent of training sets or existing database information, because the nucleotidic – numeric conversion is adapted to the statistics of the investigated DNA sequence.

Our experimental results prove that the algorithm we implemented offers a very good discrimination between coding and noncoding regions in DNA stretches. The capability to distinguish exons from introns is better than what has already been proposed. We also intend to extend our analysis on a bigger set of eukaryotic DNAs, as the examples used in this paper, and also to apply our algorithm in the case of prokaryotic organisms.

Because the numerical characterization we introduced for DNA strings results from the symbolic chain itself, the total computing time is smaller and can find general application.

REFERENCES

[1] J. W. Fickett, "The gene identification problem: an overview for developers", *Computers & Chemistry*, vol. 20(1), p. 103-118, 1996.

[2] J.-M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences", *Hum. Mol. Genet.*, vol. 6, p. 1735-1744, 1997.

[3] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences", *Bioinformatics*, vol. 16 (12), p. 1073-1082, 2000.

[4] S. Seneff, C. Wang, C. B. Burge, "Gene structure prediction using an orthologous gene of known exon-intron structure", *Appl. Bioinformatics*, vol. 3, p. 81-90, 2004.

[5] B. Brejova, D. G. Brown, M. Li, T. Vinar, "ExonHunter: a comprehensive approach to gene finding", *Bioinformatics*, vol.21(1), p. 57-65, 2005.

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

[7] 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.

[8] GenBank [Online]: http://www.ncbi.nih.gov/GenBank

[9] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals", *J. Cell. Mol. Med.*, vol. 6 (2), p. 279-303, 2002.

[10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences", *CABIOS*, vol. 13 (3), p. 263-270, 1997.