DSP IN GENOMICS: PROCESSING AND FREQUENCY-DOMAIN ANALYSIS OF CHARACTER STRINGS

Dimitris Anastassiou

Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
anastas@ee.columbia.edu

ABSTRACT

We demonstrate that digital signal processing of biomolecular sequences provides powerful approaches for solving highly relevant problems in bioinformatics by properly mapping character strings into numerical sequences. As examples, we show that color spectrograms visually provide, in the form of local texture, significant information about biomolecular sequences, thus facilitating understanding of local nature, structure and function; we provide an optimization procedure predicting protein-coding regions in DNA sequences including reading frame and coding direction, using both the magnitude and the phase of properly defined Fourier transforms; and we present a digital filtering approach to the process of translating nucleic acids into proteins. These approaches result in alternative mathematical formulations and often provide improved computational techniques for the solution of useful problems in genomic information science and technology.

1. INTRODUCTION

Biomolecular sequences, like DNA and proteins, are represented by character strings, in which each element is one out of a finite number of possible "letters" of an "alphabet." In the case of DNA, the alphabet has size 4 and consists of the letters A, T, C and G. In the case of proteins, the size of the corresponding alphabet is 20. With the explosive growth of the amount of publicly available genomic data, a new field of computer science, bioinformatics, has emerged, aimed at processing the information of these "character strings" to help solve problems in molecular biology. A plethora of computational techniques familiar to the DSP community has already been used extensively and with significant success in bioinformatics [6], including such tools as hidden Markov models and neural networks.

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. If we properly map, however, a character string into one or more numerical sequences, then digital signal processing provides a set of novel and useful tools [1], [2].

1.1. Character strings described by numerical sequences

In 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 $x[n] = au_A[n] + tu_T[n] + cu_c[n] + gu_G[n]$, 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 [15]. Any three of these four binary indicator sequences are sufficient to determine the character string, because they add to 1 for all n. If we wish to reduce the dimensionality from four to three in a manner that is symmetric with respect to all four components, we may adopt the technique [12] 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), by considering four three-dimensional vectors (a_r, a_g, a_b), (t_r, t_g, t_b), (c_r, c_g, c_b), (g_r, g_g, g_b) that 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 characters A, T, C, G. For example,

$$x_r[n] = \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n])$$
$$x_g[n] = \frac{\sqrt{6}}{3}(u_C[n] - u_G[n]) \qquad (1)$$
$$x_b[n] = \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n])$$

1.2. Essential concepts from molecular biology.

A DNA single strand is a biomolecule consisting of many linked smaller components, called nucleotides. Each

nucleotide is one out of four possible types, designated by the letters A, T, C and G, 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 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.

A DNA double strand contains two single strands, called "complementary" to each other, because each nucleotide of one strand is linked to a nucleotide of the other strand by a chemical bond so that A is linked to T and vice versa; and C is linked to G and vice versa. The two strands run in opposite "directions." An example of part of a DNA double strand is:

```
5'-    A-T-T-G-C-A-A-G-A-C-T-G    -3'
3'-    T-A-A-C-G-T-T-C-T-G-A-C    -5'
```

Because each strand of a DNA double strand uniquely determines the other strand, a double-stranded DNA molecule is represented by either of the two character strings read in its 5' to 3' direction. Thus, in the example above, the character strings 'ATTGCAAGACTG' and 'CAGTCTTGCAAT' can be alternatively used to describe the same DNA double strand, but they specify two different single strands, which are said to be complementary to each other. DNA strands that are complementary to themselves are called self-complementary, or palindromes. For example 'AATCTAGATT' is a palindrome. DNA molecules store the "digital information" that constitutes the genetic blueprint of living organisms.

A protein is also a biomolecule consisting of many linked smaller components, called amino acids. There are 20 possible types of amino acids in proteins, and, just as is the case in DNA single strands, they are connected one after the other forming a long one-dimensional chain of a specific directionality. Therefore, each protein is mathematically represented by a character string as well. The length of a protein character string is relatively small, typically in the hundreds, while the length of DNA strings is typically in the millions, or even hundreds of millions. Furthermore, contrary to DNA single strands that tend to form double helices with other DNA strands, protein molecules tend to fold into complex three-dimensional structures uniquely determined by the one-dimensional character strings, and these structures in turn determine their functions.

Each protein "character string" is synthesized based on information in "genes," which are regions in DNA "character strings," according to the "genetic code," which maps each triplet ("codon") of DNA characters into one of the 20 possible amino acids (or a "punctuation mark" like a "stop codon" signaling termination of protein synthesis). This can happen in either the "forward" or the "reverse" coding direction. Therefore, there are six possible reading frames for protein coding DNA regions. For example, if the nine nucleotide pairs of the above example correspond to protein coding regions, there are six possibilities for the codons:

```
ATT GCA AGA CTG ...
.AT TGC AAG ACT G..
..A TTG CAA GAC TG.
CAG TCT TGC AAT ...
.CA GTC TTG CAA T..
..C AGT CTT GCA AT.
```

In addition to protein coding regions, DNA contains regions serving regulatory functions, as well as regions serving yet unknown functions. One of the most relevant problems in bioinformatics is to automatically annotate sequences, by identifying such regions, using gene prediction [4], [7], [13]. Accurate prediction becomes further complicated by the fact that protein coding regions are typically separated into several isolated subregions, called "exons."

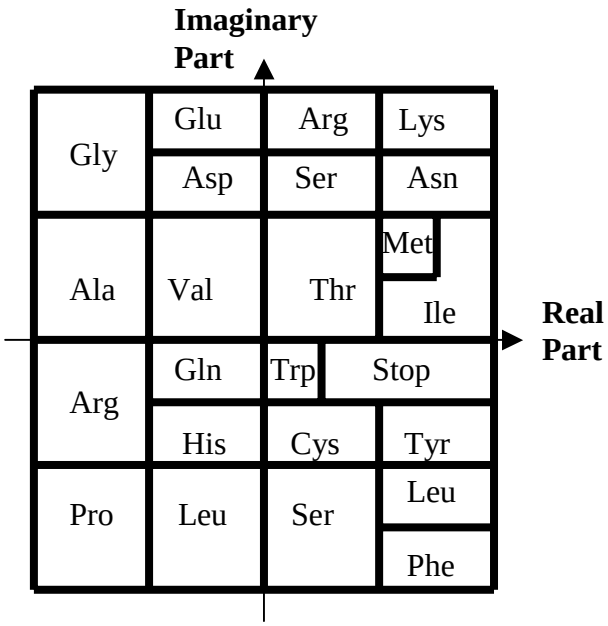
2. ASSIGNMENT OF NUMERICAL VALUES

A proper choice of the assigned 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 $\tilde{x}[n] = x^*[-n+N-1]$, n = 0, 1, …, N-1, and, in that case, all "palindromes" will yield "conjugate symmetric" numerical sequences, which have interesting mathematical properties, including generalized linear phase.

One such assignment is $a = 1+j$, $t = 1-j$, $c = -1-j$, $g = -1+j$. In that case, we may also assign numerical values to amino acids by modeling the protein coding process as an FIR "digital filter," in which the input x[n] is the numerical nucleotide sequence and the output y[n] is the resulting numerical amino acid sequence.

$$y[n] = h[0]x[n] + h[1]x[n-1] + h[2]x[n-2] \quad (2)$$

If we set $h[0] = 1$, $h[1] = \frac{1}{2}$, and $h[2] = \frac{1}{4}$, then $y[n]$ can only take one out of 64 possible values arranged as a grid on the complex plane. Furthermore, if, e.g., $x[n]$ corresponds to a forward coding DNA sequence in the first reading frame (i.e., if $x[0]$, $x[1]$, $x[2]$ correspond to the first codon), then the elements of the output subsequence $y[2]$, $y[5]$, $y[8]$, $y[11]$, …, $y[N-1]$ are complex numbers representing each of the amino acids of the resulting protein. In fact, the entire genetic code can be drawn on the complex plane as shown in the following figure. Each of the entries in the figure covers the coding results of one or more triplets (the genetic code is redundant) and corresponds to one of the 20 amino acids or to a "stop codon."



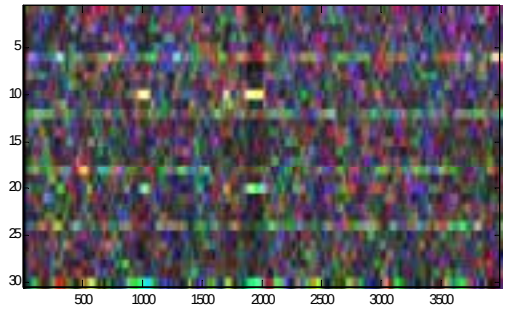
Therefore, the protein coding process can be "simulated" by a digital "low-pass" filter, followed by subsampling via a three-band polyphase decomposition, followed by a switch selecting one of the three bands ("reading frames"), followed by a "vector quantizer" as defined in the figure above.

In the frequency domain, because of (2), the Fourier transform of the sequence *y*[n] will be the product of the Fourier Transforms of *x*[n] and of the known finite-duration sequence *h*[n]. Therefore, we can use existing knowledge about the polyphase components to relate the "frequency spectra" of proteins with those of nucleic acids. Frequency domain analysis of nucleotide sequences has already been recognized as an important tool in bioinformatics by authors outside the DSP community [3], [5], [8], [9], [13], [14].

3. DNA SPECTROGRAMS

We introduce spectrograms of biomolecular sequences that simultaneously provide local frequency information for all frequencies and for all four bases. The Discrete Fourier Transform $X[k]$ of a sequence $x[n]$ provides a measure of the frequency content at "frequency" k, which corresponds to an underlying "period" of $\frac{N}{k}$ samples. Consider the DFTs $X_r[k]$, $X_g[k]$, $X_b[k]$ of the sequences defined in (1). One spectrogram simultaneously displays the three magnitudes of the Short-Time Fourier Transform, 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, the following figure shows a spectrogram using DFTs of length 60 of a DNA stretch of 4,000 nucleotides from chromosome III of *C. elegans* (GenBank Accession number NC 000967). See [2] for a color display of the figure.



The vertical axis corresponds to the "frequencies" k from 1 to 30, while the horizontal axis shows relative nucleotide locations, starting from nucleotide 858,001. The DNA stretch contains several regions, depicted as bars of high-intensity values corresponding to particular frequencies, for which there are physical explanations [1], [2].

This was a "proof-of-concept" discussion of DNA spectrograms. Of course, we may use "tapered windows," adjusting their width and shape. Furthermore, more "balanced" spectrograms can be defined using the wavelet transform, rather than the DFT. The wavelet transform has been used to analyze some fractal scaling properties of DNA sequences [3].

4. PREDICTION OF PROTEIN CODING REGIONS

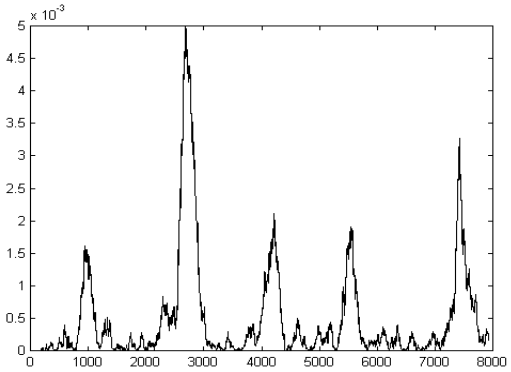
The frequency $k = \frac{N}{3}$ corresponds to a period of three samples, equal to the length of each codon. It is known [5], [7] that the spectrum of protein coding DNA typically has a peak at that frequency. We observe that for each DNA segment of length N (where N is a multiple of 3), and for each choice of the parameters a, t, c and g, there corresponds a complex number $W = aA + tT + cC + gG$, where W, A, T, C and G are the normalized (divided by N) DFT coefficients of X, U_A, U_T, U_C, U_G at frequency $k = \frac{N}{3}$. We found that, for properly chosen values of a, t, c and g, the magnitude of 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 $\Theta = \arg\{W\}$. In particular, we consider W to be a complex random variable whose properties depend on the particular choice of the parameters a, t, c, and g.

We have defined and solved [1], [2] several optimization problems, in which a, t, c, and g are selected so as to maximize the capability to distinguish between protein coding and noncoding regions. For example, the following

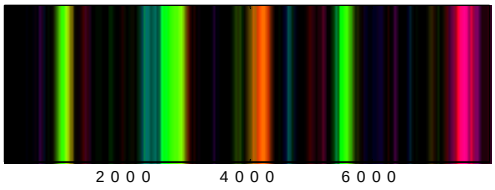
table shows the locations and the reading frames of the five exons of a particular gene (G56F11.4) of *C. Elegans*:

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

The following figure shows a plot of $|aA+tT+cC+gG|^2$ for the five exons as a function of the relative position using optimized parameter values. We have shown [1] that this method provides significant improvement in predicting protein coding regions compared with traditional [13] techniques based on Fourier analysis. Such techniques and visually tools have been proved useful; see, e.g., the pseudocolor bar labeled "Fourier measure" in the journal inset map recently resulted from the Genome Annotation Assessment Project (GASP) for *Drosophila melanogaster* [11].



In contrast, the following "color map" is a visual tool identifying the exon locations and their reading frames coming from the phase of the STFT based on a specific color mapping. See [2] for a color display of the figure.



5. CONCLUSIONS

The computational techniques and visual tools presented in this paper are meant to synergistically complement "character-string-domain" tools that have been successfully used for many years by computer scientists.

It has been said that the most significant scientific and technological endeavors of the 21st century will be related to genomics. The topic of genomics was recently featured in the IEEE Spectrum magazine [10]. We believe that there exists a unique opportunity for the DSP community to play an important role in this effort.

6. REFERENCES

[1] D. Anastassiou, "Frequency-Domain Analysis of Biomolecular Sequences," *Bioinformatics*, vol. 16, Nov. 2000, accepted for publication.

[2] D. Anastassiou, "Digital Signal Processing of Biomolecular Sequences," Technical Report EE000420-1, April 2000, http://www.ee.columbia.edu/cgi-ee-bin/show_archive.pl

[3] Arneodo, E, Bacry, P.V. Graves, and J.F. Muzy, "Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis," *Phys. Rev. Lett.*, vol. 74, pp. 3293-3296, 1995.

[4] J.-M. Claverie, "Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences," *Hum. Mol. Genet.*, vol. 6, pp. 1735-1744, 1997.

[5] V.R. Chechetkin and A.Y. Turygin, "Size-dependence of three-periodicity and long-range correlations in DNA sequences," *Phys. Lett. A*, vol. 199, pp. 75-80, 1995.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison "Biological Sequence Analysis," Cambridge University Press, Cambridge, UK, 1998.

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

[8] H. Herzel, O. Weiss, and E.N. Trifonov, "10-11 bp periodicities in complete genomes reflect protein structure and protein folding," *Bioinformatics*, vol. 15, pp. 187-193, 1999.

[9] W. Li, T.G. Marr, and K. Kaneko, "Understanding Long-Range Correlations in DNA Sequences," *Physica D.*, vol. 75, pp. 392-416, 1994.

[10] S.K. Moore, "Understanding the human genome," IEEE Spectrum, vol. 37, pp. 33-35, Nov. 2000.

[11] M.G. Reese, N.L. Hartzell, U. Harris, J.F. Ohler, J.F. Abril and S.E. Lewis, "Genome Annotation Assessment in *Drosophila melanogaster*," *Genome Res.* vol. 10, pp. 483-501, 2000.

[12] B.D. Silverman and R. Linsker, "A Measure of DNA Periodicity," *J. theor. Biol.*, vol. 118, pp. 295-300, 1986.

[13] S. Tiwari, S. Ramachandran, A. Bhattacharya., S. Bhattacharya and R. Ramaswamy, "Prediction of Probable Genes by Fourier Analysis of Genomic Sequences," *CABIOS*, vol. 113, pp. 263-270, 1997.

[14] E.N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences" *Physica A*, vol. 249, pp. 511-516, 1998.

[15] R. Voss, "Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences," *Phys. Rev. Lett.*, vol. 68, 3805-3808, 1992.