GENOME ANALYSIS USING TIME-FREQUENCY REPRESENTATIONS

Cornelia Gordan, Reiz Romulus

University of Oradea, Faculty of Electrotehnics and Informatics,
Armatei Romane Str., No. 5, 410087, Oradea, Romania

Abstract: Bioinformatics is a field of science that implies the use of techniques from mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level. Signal processing methods can also be used due to the representation of biomolecular sequences as strings of characters. The paper presents some results obtained using time frequency representations to analyze biomolecular sequences that were converted to numerical form by values assigned to each of the characters

Keywords: biocybernetics, non-stationary signals, time-frequency representations

1. INTRODUCTION

Genomics is a highly cross-disciplinary field that creates paradigm shifts in such diverse areas as medicine and agriculture. It is believed that many significant scientific and technological endeavors in the 21st century will be related to the processing and interpretation of the vast information that is currently revealed from sequencing the genomes of many living organisms, including humans.

Genomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA and proteins, have been mathematically represented by character strings, in which each character is a letter of an alphabet. In the case of DNA, the alphabet is 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.

The term DNA sequencing encompasses biochemical methods for determining the order of the nucleotide bases, adenine, guanine, cytosine, and thymine, in a DNA oligonucleotide. The sequence of DNA constitutes the heritable genetic information in nuclei, plasmids, mitochondria, and chloroplasts that forms the basis for the developmental programs of all living organisms. Determining the DNA sequence is therefore useful in basic research studying fundamental biological processes, as well as in applied fields such as diagnostic or forensic research. The advent of DNA sequencing has significantly accelerated biological research and discovery.

As the list of references shows, biomolecular sequence analysis has already been a major research topic among computer scientists, physicists, and mathematicians. The main reason that the field of signal processing does not yet have significant impact in the field is because it deals with numerical sequences rather than character strings.

However, if we properly map a character string into one or more numerical sequences, then digital signal processing (DSP) provides a set of novel and useful tools for solving highly relevant problems. For example, in the form of local texture, color spectrograms visually provide significant information about biomolecular sequences which facilitates understanding of local nature, structure, and function. Furthermore, both the magnitude and the phase of properly defined Fourier transforms can be used to predict important features like the location and certain properties of protein coding regions in DNA.

Even the process of mapping DNA into proteins and the interdependence of the two kinds of sequences

can be analyzed using simulations based on digital filtering. These and other DSP-based approaches result in alternative mathematical formulations and may provide improved computational techniques for the solution of useful problems in genomic information science and technology.

2. NUMERICAL REPRESENTATION OF DNA SEQUENCES

A DNA sequence or genetic sequence is a succession of letters representing the primary structure of a real or hypothetical DNA molecule or strand, with the capacity to carry information.

The possible letters are A, C, G, and T, representing the four nucleotide subunits of a DNA strand - adenine, cytosine, guanine, thymine bases covalently linked to phospho-backbone. Typically the sequences are printed abutting one another without gaps, as in the sequence AAAGTCTGAC, going from 5' to 3' from left to right. A succession of any number of nucleotides greater than four is liable to be called a sequence. With regard to its biological function, which may depend on context, a sequence may be sense or anti-sense, and either coding or noncoding. DNA sequences can also contain "useless DNA". Sequences can be derived from the biological raw material through a process called DNA sequencing.

Single DNA strands tend to form double helices with other single DNA strands. Thus, 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, being linked by a set of weak (hydrogen) bonds.

Because each of the strands 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, for example the character strings CATTGCCAGT and ACTGGCAATG can be alternatively used to describe the same DNA double strand, but they specify two different single strands which are complementary to each other. DNA strands that are complementary to themselves are called self-complementary, or *palindromes*. For example AATCTAGATT is a palindrome.

Most of the identified genomic data is publicly available over the Web at various places worldwide, one of which is the Entrez search and retrieval system of the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH). The NIH nucleotide sequence database is called GenBank and contains all publicly available DNA sequences. For example, one can go to http://www.ncbi.nlm.nih.gov/entrez and identify the DNA sequence with Accession Number AF 099922; choose Nucleotide under Search and then fill out the other entry by typing: AF 099922 [Accession] and pressing "Go." Clicking on the resulting accession number will show the annotation for the genes as well as the whole nucleotide sequence in the form of raw data.

Similarly, Entrez provides access to databases of protein sequences as well as 3-D macromolecular structures, among other options. As another example, a specialized repository for the processing and distribution of 3-D, macromolecular structures can be found in the Protein Data Bank at www.rcsb.org.

In a DNA sequence of length N, assume that we assign the numbers a, t, c, g to the characters A, T, C, G, respectively. A proper choice of the numbers a, t, c and g 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:

$$\tilde{x}[n] = x^*[-n + N - 1], \text{ n=0, 1,…, N-1} \qquad (1)$$

and, in this case, all palindromes will yield conjugate, symmetric numerical sequences which have interesting mathematical properties, including generalized linear phase.

One such assignment (the simplest out of many possible ones) is the following:

$$a = 1 + j, \;\; t = 1 - j, \;\; c = -1 - j, \;\; g = --1 + j. \; (2)$$

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 possible resulting numerical amino acid sequence (if $x[n]$ is within a coding region in the proper reading frame):

$$y[n] = h[0]x[n] + h[1]x[n-1] + h[2]x[n-2]. \; (3)$$

For example, if we set $h[0] = 1$, $h[1] = 1/2$, and $h[2] = 1/4$, and $x[n]$ is defined by the parameters in (2), then $y[n]$ can only take one out of 64 possible values.

Furthermore, if for example, $x[n]$ corresponds to a forward coding DNA sequence in the first reading frame, 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.

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.

In the frequency domain, because of (3), 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, or correlation analysis of nucleotide sequences, has already been recognized as an important tool in bioinformatics by authors outside the DSP community. In other words, certain useful frequency-domain properties of proteins can be evaluated from the corresponding frequency-domain properties of nucleic acids.

In the field of multirate signal processing there are several results and equations connecting the frequency spectra of polyphase components, which more accurately relate the frequency spectra of proteins with those of nucleic acids, providing a novel computational framework. Of course, it is possible that other choices of the parameters *a*, *t*, *c*, *g*, and of the FIR coefficients, may provide a better fit with actual data when solving such bioinformatics problems as alignment of nucleotide or amino acid sequences.

3. DNA ANALYSIS USING TIME-FREQUENCY REPREZETATIONS

Whe chosed for our experience a sequence dubbed M10051 from the GenBank database, corresponding to the human insulin receptor mRNA. First some statistical data processing was done using the Bioinformatics Toolbox from Matlab, which contains a number of useful programs for genomic data analysis. For the sequence we considered, a nucleotide count yields the following result:

A: 1068
C: 1298
G: 1311
T: 1046

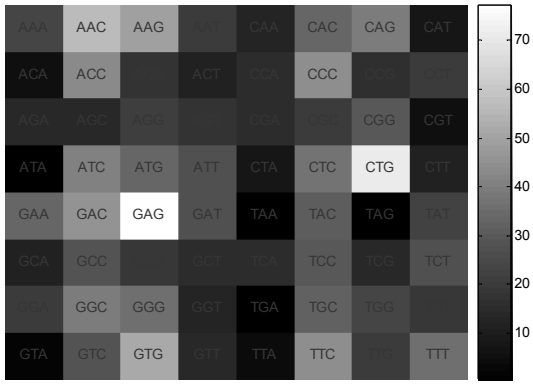

Fig.1. A heat map representation of the codon count for the analysed sequence.

Figure 1 shows a heat map for the occurrences of different codons within the analysed sequence

Another good statistical tool is the density of nucleotides in a sequence. In our case the density of the nucleotides A,T,C,G in the sequence is presented in figure 2.

By mapping the characters from the genomic sequence to numerical values a digital signal can be obtained. The choice of values can be adjusted obtaining different signals that can be processed to extract different properties of the genomic sequence.

Generally the signals that are obtained with this method are non-stationary signals, with parameters changing in time. Time-frequency representations are the best tools when nonstationary signals are analysed. For example, the short-time Fourier transform of a discrete-time signal is given by the following relation:

$$F_x(t,v;h) = \int_{-\infty}^{+\infty} x(u)h^*(u-t)e^{-j2\pi vu}du \quad (4)$$

where h(t) is a short time analysis window localized around t = 0 and $\upsilon = 0$.

It is well known that the appearance of time – frequency representations provides significant information about signals, to the extent that trained observers can figure out the words uttered in voice signals by simple visual inspection of their spectrograms. Similarly, it appears that these representations are powerful visual tools for biomolecular sequence analysis.

Here we present the short-time Fourier transform (STFT), of genomic sequences using the discrete Fourier transform (DFT) as a simple example of a frequency-domain analysis tool.

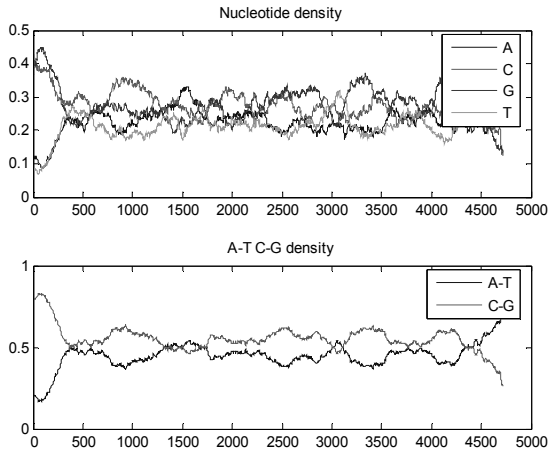

Fig.2. The density of the nucleotides A,T,C,G in the analyzed sequence.

For a numerical sequence $x[n]$ of length N, the DFT $X[k]$ is another sequence of the same length N, providing a measure of the frequency content at frequency k, which corresponds to an underlying period of N / k samples, where the maximum frequency (period 2) corresponds to $k = N / 2$, assuming that N is even.

The STFT representation results by applying the DFT over a sliding window of small width to a long sequence, thus providing a localized measure of the frequency content.

In the case of biomolecular sequences, we want these spectrograms to simultaneously provide local frequency information for all four bases; therefore, it is best to avoid using just one choice of assigned numbers *a*, *t*, *c*, *g* to the characters *A*, *T*, *C*, *G*, respectively.

4. RESULTS

To show how time-frequency-domain analysis of DNA sequences can be a powerful tool for specifically identifying protein coding regions in DNA sequences, we chosed to analyse two different genomic sequences.

The first one is a totally random sequence; the second one was taken from the Genbank database and contains a periodic pattern that is clearly visible in the time-frequency plane.

First we obtained the spectrum of the random signal, using the Fast Fourier algorithm. The spectrum obtained in this case is presented in figure 3.

In figure 4 is presented the results obtained in the time-frequency plane using the short-time Fourier transform.

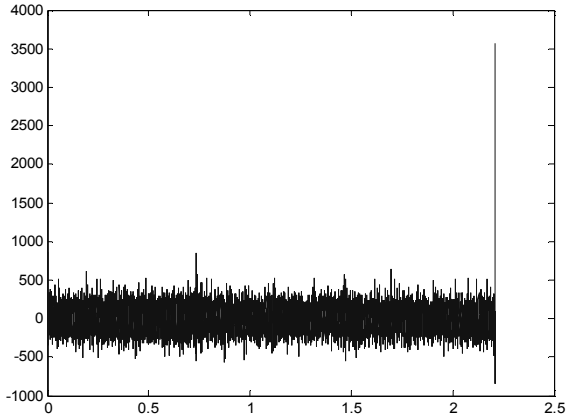

Fig.3. Plot of the spectrum of the random DNA sequence

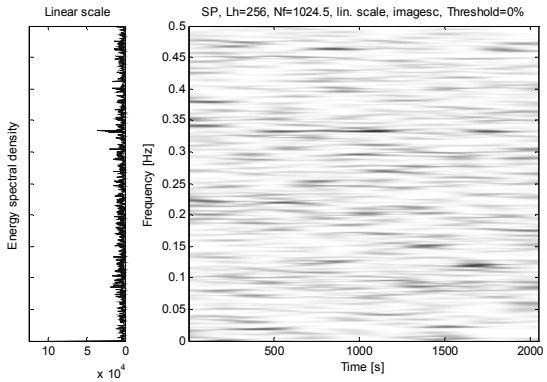

Fig. 4. Short-time Fourier representation of the random DNA sequence

As it can be noticed in figure 4 in the time-frequncy plane no obvious patterns are discernable, meaning that the analysed sequence in this case has a totally random nature.

In the case of a sequence that contains periodic patterns, i.e. a sequence of protein coding DNA, the spectrum shows peaks. This is the case of the second sequence we analyzed, which contains the human insulin receptor cDNA. This sequence has the M10051 Genbank accession number.

The plot of the spectrum obtained in this case is presented in figure 5. This plot clearly shows the presence of a peak that corresponds to the presence of a periodic pattern in the genomic sequence that was analysed.

However, this representation does not give us information about the position of this periodic pattern in the genomic sequence.

The results obtained in the time-frequency plane using the short-time Fourier transform show the presence of the DNA coding region with even more accuracy. This representation is presented in figure 6.

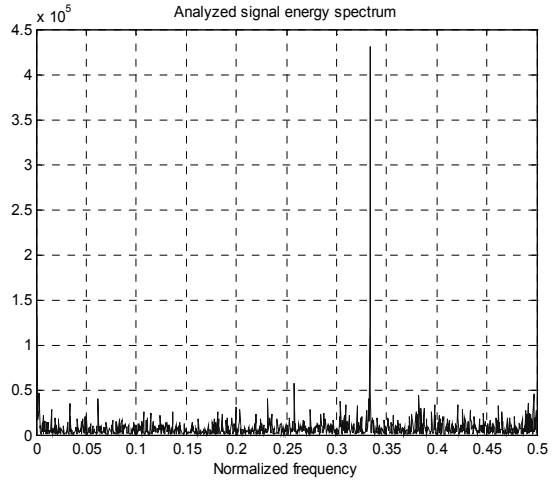

Fig.5. Plot of the spectrum of a sequence that contains a DNA coding region

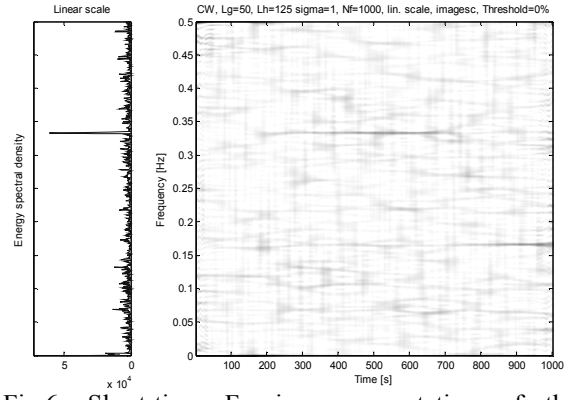

Fig.6. Short-time Fourier representation of the sequence that contains a DNA coding region

Using a 3D representation of the same sequence the presence of the periodic pattern is even more obvious, as presented in figure 7.

Of course the choice of the time-frequency representation influences the results. There is an increasing number of such representations and some of them (i.e. the wavelet based representations) are especially useful to this type of research.

Other factors that can influence the results are the parameters that were chosen for the representation, i.e. in the case of STFT the type and length of the windowing function. A Morlet scalogram of the same sequence is presented in figure 8.

The patterns that are observed in the time frequency plane show clearly the existence of the coding region in the analysed sequence. Because the representation in this case is an image, further information can be obtained using image processing and pattern recognition methods.

Of course, there are numerous other ways in which spectrograms can be defined. We may use tapered windows, and adjust their width and shape.

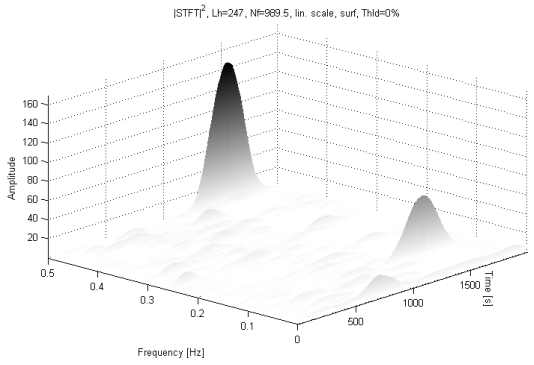

Fig.7. 3D Short-time Fourier representation of the sequence that contains a DNA coding region

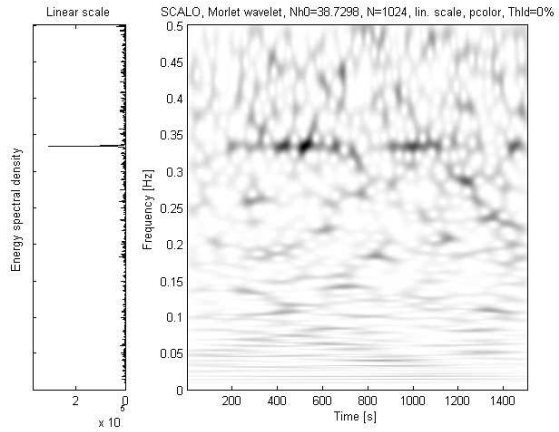

Fig.8. Morlet scalogram of the sequence that contains a DNA coding region

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.

5. CONCLUSIONS

Signal processing-based computational and visual tools are meant to synergistically complement character-string-domain tools that have successfully been used for many years by computer scientists. In this article, we illustrated one of several possible ways that signal processing can be used to directly address biomolecular sequences The assignment of optimized, complex numerical values to nucleotides and amino acids provides a new computational framework, which may also result in new techniques for the solution of useful problems in bioinformatics, including sequence alignment, macromolecular structure analysis, and phylogeny. The use of time-frequency representations extends the possibilities that are offered to researchers in the field of genomic signal processing.

REFERENCES

D. Anastassiou, ″Frequency-domain analysis of biomolecular sequences″, *Bioinformatics*, vol. 16, no. 12, pp. 1073-1082, Dec. 2000.

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

L. Cohen. Time-Frequency Distributions - A Review. *Proceedings of the IEEE*, 77(7):941{980, 1989.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, "*Biological Sequence Analysis*", Cambridge, U.K.: Cambridge Univ. Press, 1998

A. Stein and M. Bina, "A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment," *Nucleic Acids Res.*, vol. 27, pp. 848-853, 1999