Biomedical Signals Local Maxims Detection Using Time Frequency Transforms

Reiz Romulus and Gordan Cornelia
Department of Electronics,
University of Oradea, Faculty of Electrical Engineering and Information Technology,
1 Universității Str., 410087 Oradea, Romania, E-Mail: cgordan@uoradea.ro

Abstract **– *The wealth of genomic data currently available in online databases has caused a need for new algorithms and analysis techniques to interpret genomic data. In this paper we explore techniques for locating critical genomic data in protein sequences and for estimating the similarity between proteins. By converting genomic data into numeric sequences signal processing methods can be applied to process the resulting information. We demonstrate the use of the Short-Time Fourier Transform and the Continuous Wavelet Transform in locating important amino acid properties contained in protein sequences.***

Keywords: ***Genomic signal, time-frequency representations,***

I. INTRODUCTION

Genomic analysis is a highly cross-disciplinary field, which will offer many significant scientific and technological endeavors in the 21st century, because of the vast information that is revealed from sequencing the genomes of living organisms[1].

Genomic information is digital represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, as DNA and proteins, can be 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.

If we properly map a character string into one or more numerical sequences, then the digital signal processing provides a set of novel and useful tools for solving highly relevant problems from genetic field.

For example, both the magnitude and the phase of properly defined Fourier transforms can be used to predict important features like location and certain properties of protein coding regions in DNA. Time-frequency representations offer information concerning the spectral content of the non-stationary signal at every moment of time. In the case of bio molecular sequences, we want the spectrograms to simultaneously provide spectral information for all 4 basic characters: a, t, c, g. Using vertex spaces mathematics techniques can be reduced the number of variables from 4 to 3.

The three-dimensional structure of a protein is important because protein structure is linked to protein's function. In order to affect a function on a cell, it is not uncommon for a protein to have to dock a specially shaped section of its three-dimensional structure into a specifically shaped receptor on a target cell. In short, many proteins have a site on them that initiate, mediate or terminate a particular biological action.

Some international teams working on this subject have established physic mathematical models for protein analysis. The basis of this discovery is connected to the fact that there exists a significant correlation between the spectra of numerical representations of amino acids and their biological activity [2]. More specifically, the biological function of a protein is characterized by certain frequencies of its signal representation [3].

This type of analysis first involves converting the amino acids that constitute a protein into a "discrete time series." The position of an amino acid in the sequence can be thought of as the time. After the conversion of the amino acid sequence is made into the protein time (space) series signal (which we call a "protein signal") the signal is analyzed to locate the dominant frequencies.

In this study we use the Continuous Wavelet Transform (CWT) and the Short-Time Fourier Transform (STFT) to perform time-frequency analyses of proteins. These two transforms have the advantage of presenting information about space (time) that in our case is associated to a particular amino acids location in a protein and we are consequently able to identify the active amino acids contributing to the characteristic frequencies of the proteins.

II. GENERALITIES

1. Hemoglobin

Respiration in living cells requires oxygen. Oxygen enters the human body through the pulmonary system and has to be carried to cells all over the body. Hemoglobin is the protein found in red blood that carries oxygen to most of the cells the body.

In human adults hemoglobin is a protein with a quaternary structure composed of two sets of two identical subunits. The two alpha subunits are each made up of 141 amino acids and the beta subunits are made up of 146 subunits each. The most important part of each subunit is its hemmed group. The hemmed group is a cofactor (a non-protein compound required for the proper functioning of certain proteins) with a central iron atom. An oxygen molecule binds reversibly to each

subunit via its hemmed group; consequently, each molecule of hemoglobin can carry four molecules of oxygen. Upon reaching its destination, hemoglobin unloads the oxygen molecules.

2. Myoglobin

Myoglobin is similar in function to hemoglobin except that its function is to store oxygen in muscle cells. Diving animals like seals and whales often spend extended periods of time underwater and as such are not able to take in oxygen through their nostrils and into their pulmonary systems as they normally do on land. In times like these diving animals rely heavily on myoglobin for oxygen. Myoglobin is also found in the skeletal and cardiac muscle of non-diving animals. Human myoglobin consists of one peptide chain consisting of 154 amino acids. It only has one hemmed group and consequently can only store one molecule of oxygen.

3. Time-frequency representations used to analyze genomic sequences

To obtain our sequences, the proteins were converted into a "time series" of consecutive amino acids in the protein. In the application of signal processing techniques to the sequences, the sampling rate can be assumed to be 1 since the distance between amino acids is about 3.8Å.

To obtain the time-frequency representations of the DNA sequence the protein sequence needs to be converted into a numeric form that can be further processed using digital signal processing methods. 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 resulting numerical sequence $x[n]$. One of the simplest assignments is the following one [1]:

$$a = 1 + j, \quad t = 1 - j, \quad c = -1 - j, \quad g = --1 + j. \qquad (1)$$

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.

As it was noticed by many researchers coding regions present a period-3 behavior, that can be easily be detected using spectral analysis. This property can be used to detect coding regions in the genomic sequences, and even to make predictions on the role of the coding region. To perform gene prediction based on the period-3 property, some indicator sequences for the four bases are defined then the DFTs of short segments of these are computed. To obtain the DNA spectrograms we used the Short-Time Fourier Transform (STFT). For this representation is calculated the Discrete Fourier transform (DFT) of a sequence from the signal, using a narrow window that slides over the whole existing time domain. Thus, is provided a localized measure of the frequency content for some moments of time[10].

$$TF_x^{STFT}(t,\omega) = \int_{-\infty}^{+\infty} x(\tau)\cdot \mathbf{w}(\tau - t)\cdot e^{-j\omega\tau} d\tau \qquad (2)$$

where $x(t) \in L^2$® and w(t) is called "time window".

At instant t, TF_x^{STFT} is the Fourier transform of x(t) sequence multiplied with w(τ-t) window. Because this window will eliminate all the x(t) characteristics that are place outsides the extreme vicinity of t, TF_x^{STFT} will define the "local spectrum" concentrated around x(t). Taking into account all the t values between (-∝, ∝) will be found the entire x(t) spectral content.

If the window is "Gaussian' type then STFT will be called Gabor transform. Depending on the application one can chose different type of windows, as: Hamming, Hanning, Blackmann, rectangular, adaptive, etc.

We chose a 12 points length Hamming window as this is close to the average length of an alpha helix, one of the common secondary structure conformations in proteins. We had and overlap of 11 points between windows. Overlap is important as it dictates the amount of frequency information lost due to the splitting of the signal into windows.

Another useful signal processing tool similar to the Short-Time Fourier Transform is the wavelet transform. This representation is superior to the STFT because it uses a variable resolution to analyze the signal, according to the frequency content of that signal.

Continuous wavelet transform CWT, uses a window depending both on time and frequency. So, CWT analyses the time-frequency plane with variable dimensions cells, based on a window defined as follows[12]:

$$w(\tau) = \sqrt{s}\cdot\psi(s(\tau - t)) \qquad (3)$$

Continuous wavelet transform is introduced by the following expression:

$$CWT_x(s,t) = \sqrt{s}\int_{-\infty}^{+\infty} x(\tau)\cdot\psi(s(\tau - t))d\tau \qquad (4)$$

where ψ(τ) is the analyzing wavelet and s is a scale parameter depending on the x(t) sequence length.

By applying the CWT to the numerical analyzed sequence one has to look for the local energy maxim in the space-frequency representation of the protein signal as the amino acids being searched for are those that contribute the most frequency-wise.

Although the linear time-frequency representations are very useful, in some energetic distributions applications are preferred quadratic representations. So,

the $\left|TF_x^{STFT}(t,\omega)\right|^2$ function is known as being the spectrogram of x(t) and the $\left|TF_x^{CWT}(t,\omega)\right|^2$ is the scalogram. In these two cases the number of interference terms is much bigger then for the linear situations.

III. RESULTS

Whe chosed for our paper three sequences from the GenBank database, corresponding to the human alpha hemoglobin alpha-1 globin chain (HBA1), the hemoglobin beta (HBB) and the myoglobin chain. 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 analysed all of these sequences and obtained the corresponding spectrograms and scalograms. The obtained time-frequency representations are presented in figure 1 to 6. The dark spots that are present in the images are high-energy areas located in the time0frequency plane, that correspond to important features of the processed genomic sequence. Our results reveal that neither the Short-Time Fourier Transform nor the Continuous Wavelet Transform is significantly better than the other in localizing "hot spots" in the DNA sequences.

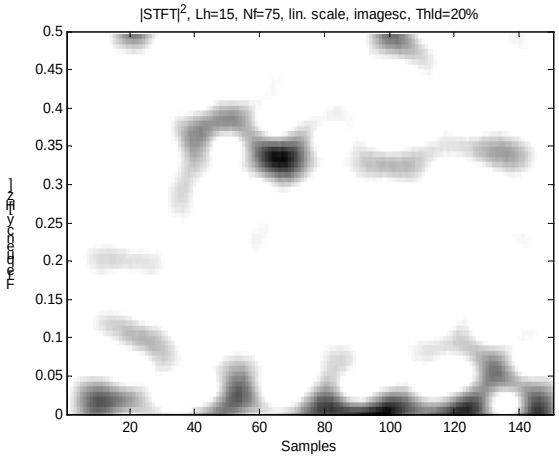

Fig. 1 STFT of human hemoglobin alpha chain

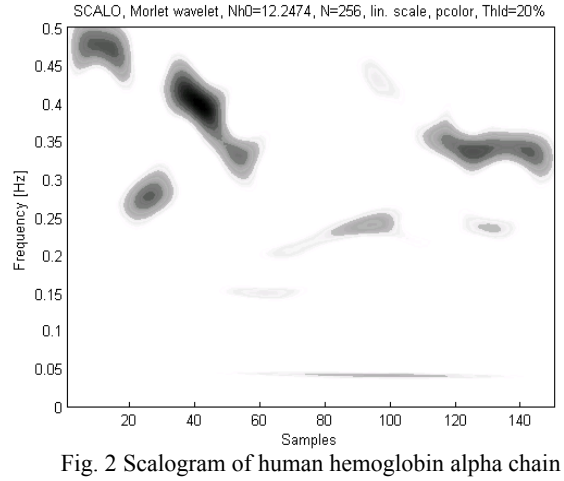

Fig. 2 Scalogram of human hemoglobin alpha chain obtained using the Morlet wavelet

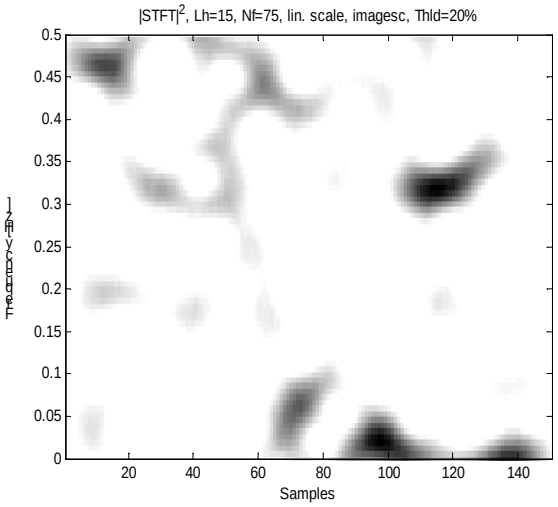

Fig. 3 STFT of human hemoglobin beta chain

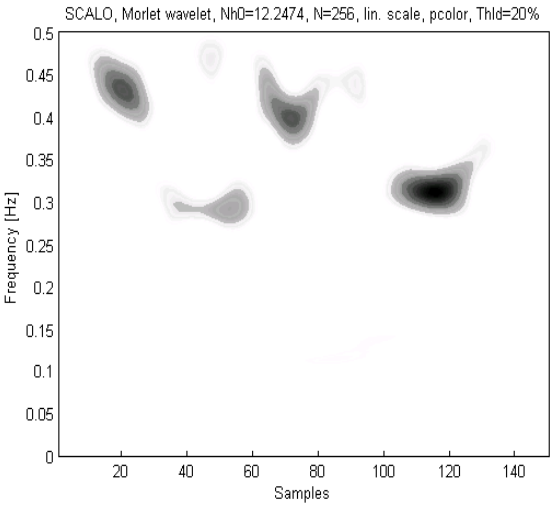

Fig. 4 Scalogram of human hemoglobin beta chain obtained using the Morlet wavelet

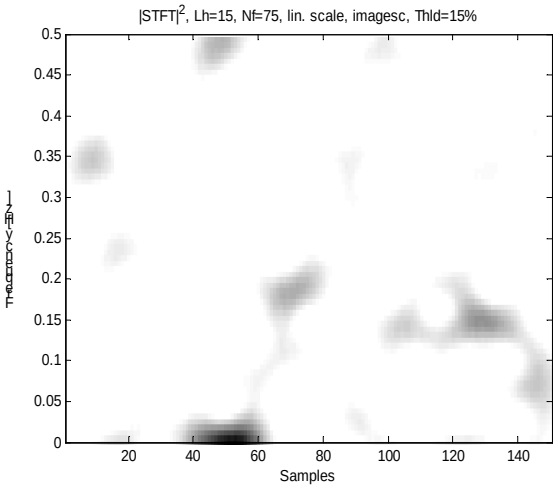

Fig. 5 STFT of human hemoglobin alpha chain

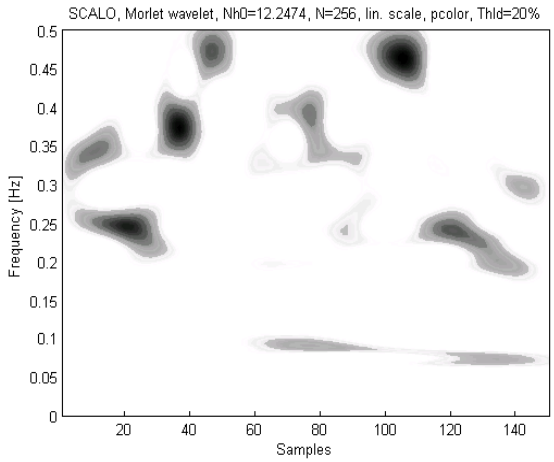

Fig. 6 Scalogram of human hemoglobin alpha chain obtained using the Morlet wavelet

IV. CONCLUSIONS

Our objective was to locate the amino acids ("hot spots") critical to the function of these DNA chains in the resulting spectrogram and scalogram. Our results show that both the STFT and the CWT are able to locate a genomic signal's "hot spots" given a signal based on amino acid chains. 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

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

[2] E. Pirogova and I. Cosic, "Examination of Amino Acid Indexes Within the Resonant Recognition Model," Proceedings of the IEEE Engineering in Medicine and Biology Society Conference: Biomedical Research in 2001

[3] C. de Trad, Q. Fang, I. Cosic, "The Resonant Recognition Model (RRM) Predicts Amino Acids in Highly Conserved Regions of the Hormone Prolactin (PRL)," Biophysical Chemistry 84 (2000)149-157

[4] E. 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

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

[6] E. Coward, "Equivalence of two Fourier methods for biological", J. Math. Biol., vol. 36, pp. 64-70, 1997

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

[8] 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

[9] 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

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

[11] L. Cohen, "Time-Frequency Analysis", Englewood Cliffs, NJ: Prentice-Hall, 1995

[12] P. Flandrin, "Time-Frequency/Time-Scale Analysis", San Diego, CA: Academic, 1999