Genomic Signal Analysis of HIV-1 Clade F Gene Variability

Paul Dan Cristea, *Senior Member, IEEE*

Abstract

Genomic Signal methodology is one of the approaches that can contribute to bridging the gap between the overwhelmingly large and continuously increasing amount of data produced by the high throughput experimental techniques available today in Genomics, Transcriptomics and Proteomics, on one hand, and the still unsatisfactory techniques for automatic processing and analysis of the information at molecular level, on the other hand. The extraction of biologically relevant features and the integration of knowledge over many orders of magnitude – running from molecular, cellular organelle, cellular, tissue, and organ levels, to whole organism level, promise to provide unprecedented insight and guidance for the medical decisions at clinical level.

In the largest understanding of the term, genomic signals are conceived as carriers of genomic information in all the processes taking place in an organism and essentially determining the dynamics of genomic events governing such processes in conjunction with environmental factors [1]. The research in this proposal is focused on a more restricted meaning, specifically the genomic signals used in sequence analysis [2]. Put simply, this is the conversion of nucleotide and amino acid sequences into digital genomic signals offering the possibility to apply signal processing methods for the analysis of genomic data [3-5]. The genomic signal conversion used in our work is a one-to-one numeric representation of symbolic genomic sequences, chosen to be as little biased as possible, treating the same the exons and the introns, i.e., coding and non-coding regions [3]. This methodology has already been used for data mining [6,7] the available large genomic databases [8], but the same approach shows a great potential in epidemiological (including pathogen variability), anthropological (including population genetics) and mezological (including environmental conditions impact) studies.

The paper presents results on the study of HIV-1, Clade F, variability using IT techniques for the quantitative analysis of nucleotide sequences [9]. Specifically, phase analysis and RNA secondary structure for the protease and reverse transcriptase genomic signals have been used to characterize the variability of HIV strains isolated in Romania. Variability signals with respect to average, median and maximum flat references, as well as digital derivative signals are used to describe HIV variability. Some data will also be presented on the current research we are currently conducting to study the variability of mycobacterium tuberculosis, another well known pathogen with significant societal impact.

References

1. E.G. Dougherty, I. Shmulevich, Jie Chen, Z. Jane Wang (Editors), "Genomic Signal Processing and Statistics", *EURASIP Book Series on Signal Processing and Communications*, Hidawi Publishing Corporation, 2005.
2. P. D. Cristea, "Conversion of Nitrogenous Base Sequences into Genomic Signals", *Journal of Cellular and Molecular Medicin*e, 6, 2, pp. 279-303, April – June 2002.
3. P. D. Cristea, "Representation and analysis of DNA sequences", in: *Genomic Signal Processing and Statistics*, Editors: E.G. Dougherty, I. Shmulevich, Jie Chen, Z. Jane Wang, *EURASIP Book Series on Signal Processing and Communications*, Hidawi Publishing Corporation, 2005, Chapter 1, pp. 15-65.
4. I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and application", *IEEE Trans Biomed Eng,* **41**(12): 1101-14, 1994.
5. D. Anastassiou, "Frequency-domain analysis of biomolecular sequences", *Bioinformatics*, vol.16, no. 12, pp. 1073-1081, 2000.
6. P. D. Cristea, "Large Scale Features in DNA Genomic Signals"*, ELSEVIER, Signal Processing, Special Issue on Genomic Signal Processing*, 83, pp. 871-888, 2003.
7. P. D. Cristea, "Genomic Signals of Re-Oriented ORFs"*, EURASIP – Journal on Applied Signal Processing, Special Issue on Genomic Signal Processing*, vol. 2004, no.1, pp. 132-137, January 1, 2004.
8. National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, GenBank, *http://www.ncbi.nlm.nih.gov/genoms*.
9. P. D. Cristea, D. Otelea, Rodica Tuduce, "Genomic signal analysis of HIV variability", in *Proc. SPIE - BIOS 2005 – Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues II Conf. ,* vol. 5699, January 22-27, paper 52.

Keywords – Genomic signals, Pathogen variability, HIV, MT, RNA secondary structure.