Genomics and Proteomics: *Information-Theoretic* Analysis in Frequency Domain

Sergey Edward Lyshevski
Department of Electrical Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA
E-mail: Sergey.Lyshevski@mail.rit.edu Web site: www.rit.edu/~seleee

ABSTRACT

We study the feasibility of analysis and the evaluation of large-scale genomic and proteomic data by applying frequency concepts. The frequency-domain estimates and measures lead to the *information-theoretic* analysis departing from statistical methods. The frequency methods and *spectrum analysis* may have significant advantages by guaranteeing robustness as well as ensuring quantitative and qualitative features. The symbolic and numeric approaches are applied providing overall coherency. The frequency-domain analysis necessitates the use of numeric mappings. Though additional efforts and consistent evaluations are needed to assess the proposed methodology, we demonstrate the genomic and proteomic correlations. It is also illustrated that the information in some or other form is coded by the nucleotides and amino acids sequences. The analysis of sequences and information complexity, in general, cannot be accomplished without a great number of assumptions and postulates. The results are illustrated for HIV, cancer and other sequences.

Keywords: frequency domain, genome, genomics, *information-theoretic* analysis, proteomics

1. INTRODUCTION

Different statistical methods have been used to analyze and evaluate large-scale data by performing data analysis and data mining. These attempts were partially successful due to overall complexity, sequences gaps, nonconding *low complexity* regions, inaccuracy, etc. The use of statistical methods in analysis of large-scale data, produced by high-throughput experiments, has limitations and drawbacks.

Meaningful databases have been developed. The SCOP, CATH and FSSP databases classify proteins based on structural similarity, Pfam and ProtoMap identify families of proteins based on sequence homology, while PartList and GeneCensus examine the occurrence of protein families in various genomes. Genome sequences for different organisms are available. In particular, (1) GenBank, DDBJ and EMBL provide nucleic acid sequences; (2) PIR and SWISS-PROT report protein sequences; (3) Protein Data Bank offers protein structures.

Statistical methods test *a priori* hypotheses against data with a great number of assumptions and simplifications under which the genome-genome comparison can be performed. The "learning" methods (clustering, Bayesian networks, decision trees, neural networks and other) were used to study trends and patterns in the large-scale data within moderate progress. We propose an *information-theoretic* approach. This concept promises to ensure robust, systematic and coherent analysis in the frequency domain [1, 2]. The proposed approach complies with conventional data formats and complements other methods ensuring assessment of complex large-scale data under uncertainties.

The large-scale genomics and proteomics are the forefront of medicine, life science and engineering. The qualitative and quantitative analyses are performed for various sequences, including HIV and cancer genomes.

2. FOURIER TRANSFORM AND ITS APPLICATION

Postulate. *The information is coded and the functionalities are defined by a finite sequence of nucleotides or amino acids in the genomic and proteomic sequences. These finite sequences are distinguishable and provide unique characteristics identifiable in the frequency domain.* ■

Let $\mathbf{A}$={A, C, G, T} is the *symbolic quaternary alphabet*. This *alphabet* can be mapped (represented) as

$\mathbf{M}$={0 1 2 3}, $\mathbf{M}$={j $-j$ 1 -1}, $\mathbf{M}$={1+j -1+j 1$-j$ $-1-j$}.

Other mappings can be used. The arbitrary pairs of quaternary N-sequences (words of length N) are $x=(x_1,x_2,\ldots,x_{N-1},x_N)$, $x_i\in\mathbf{A}$ and $y=(y_1,y_2,\ldots,y_{N-1},y_N)$, $y_i\in\mathbf{A}$. For a pair (x,y) of quaternary words, the statistical measures and similarity $S(x,y)=\sum_{i=1}^{N} s(x_i,y_i)$ can be found. We depart from these approaches which have a limited practicality.

Consider a finite sequence of nucleotides A, T, C and G. We assign the symbol or values a, t, c and g to the characters A, T, C and G. These a, t, c and g can be mapped by real and complex mappings. The numerical sequence, resulting from a character string of length N, is

$$x[n]=au_A[n]+tu_T[n]+cu_C[n]+gu_G[n],\ n=0,1,2,\ldots,N-1,$$

where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are the binary indicators which take the value of either 1 or 0 at location n depending on whether the corresponding character exists or not at location n; N is the length of the sequence.

The amino acid sequences are expressed as

$$x[n]=A_{la}u_{Ala}[n]+A_{rg}u_{Arg}[n]+\ldots+T_{yr}u_{Tyr}[n]+V_{al}u_{Val}[n].$$

Using the amino acids, the *symbolic alphabet* is $\mathbf{A}$={Ala,Arg,…,Tyr,Val} with the corresponding *alphabet mapping*. The amino acid sequence is

$$x[n]=au_a[n]+ru_r[n]+\ldots+tu_t[n]+vu_v[n].$$

We obtain the symbolic strings which map nucleotides and amino acids finite sequences. The discrete Fourier transform of a sequence $x[n]$ of length N is

$$X[k]=\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn},\ k=0,1,2,\ldots,N-1.$$

This Fourier transform provides a measure of the frequency content at frequency k which corresponds to a period of N/k samples. The sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ are the discrete Fourier transforms of the binary indicators $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$. The finite genomic and proteomic sequences are distinguishable and may provide unique characteristics identifiable and observable in the frequency domain. These characteristics and data may not be observable by statistical,"learning" or other methods.

Example. We apply the Fourier transform and examine the frequency components of a perfect nucleotide sequence and $x[n]$ under uncertainties. The *symbolic quaternary alphabet* $\mathbf{A}$={A,C,G,T} is mapped as $\mathbf{M}$={1 2 3 4}. Figure 1.a reports the mapping. The magnitude of Fourier transform $|X[k]|$ indicates that there are four distinguished frequencies. The uncertainties (gap, error, missing site, inconsistency, etc.) are mapped by $\mathbf{U}$. Let the uncertainties occur at 7, 24, 38 and 48 sites. We map the gaps and

inconsistencies as **U**={–1 –2}. The resulting sequence and $|X[k]|$ are given in Figure 1.b. Under very large uncertainties, the results are documented in Figure 1.c. Examining $|X[k]|$ we conclude that the proposed concept ensures robustness, detection, observability, data mining and other features under large uncertainties. The qualitative and quantitative estimates correspond to a perfect sequence. The statistical methods do not provide sound estimates.

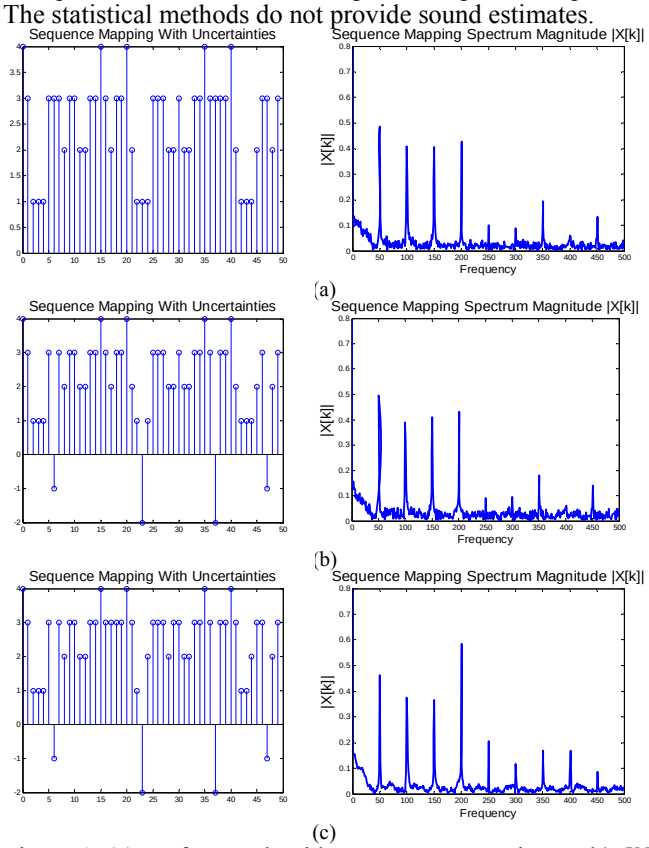


Figure 1. (a) Perfect nucleotide sequence: Mapping and $|X[k]|$; (b) $x[n]$ with missing sites and corresponding $|X[k]|$; (c) $x[n]$ with large uncertainties and resulting $|X[k]|$

3. Applications and Results

The frequency-domain analysis was performed for complete *E.coli* and *S.typhimurium* genomes with 4,639,221 and 4,937,381 base pair strains [3-5]. An interactive toolbox is developed in MATLAB to accomplish a robust frequency-domain analysis. We utilize the *spectrum analysis*, power spectral density (PSD) and other estimates. Various parametric (autocorrelation, covariance, etc.), non-parametric (periodogram, Welch, etc.) and space methods are applied and utilized to obtain PSD.

Figure 2 reports the distinguishable PSDs for bigA's (www.genome.jp/dbget-bin/www_bget?stm+STM3478) and ratA's (www.genome.jp/dbget-bin/www_bget?stm:STM2515) sequences. We perform the analysis for the putative surface-exposed virulence protein bigA (1953 AA and 5862 NT length) and putative outer membrane protein ratA (1865 AA and 5598 NT length). Our *spectrum analysis* indicates that these *S.typhimurium* genes are distinct and there is no correlation. These comply with the established findings.

The sequences may not be complete, there can be missed sites, etc. The HIV and cancer genes are typical examples [6]. It is virtually impossible to analyze patterns using statistical and "learning" methods. Furthermore, linear maps may not be found. In contrast, the reported concept is effectively applied providing meaningful results.

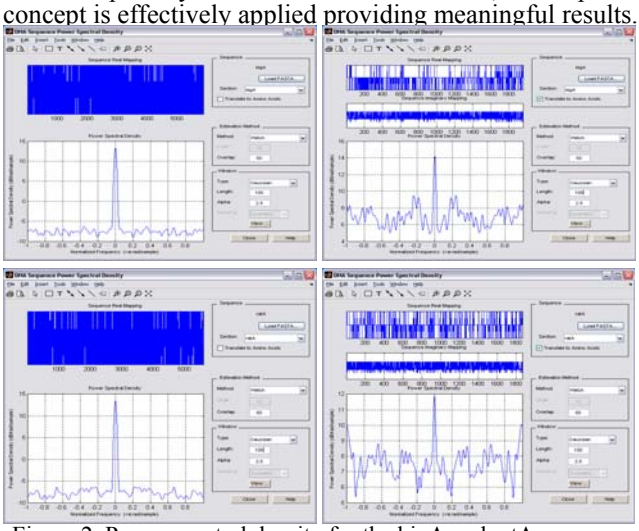


Figure 2. Power spectral density for the bigA and ratA sequences

Figures 3 and 4 illustrate the PSDs for the nucleotide and amino acid sequences for HIV. For a cancer gene, the results are documented in Figures 5 and 6.

The frequency analysis promises to solve a spectrum of problems such as: (1) Detect, identify and distinguish proteins and genes; (2) Examine and identify protein coding genes; (3) Potentially define structural and functional characteristics; (4) Analyze the data and perform data mining; (5) Identify patterns in gene sequences, etc.

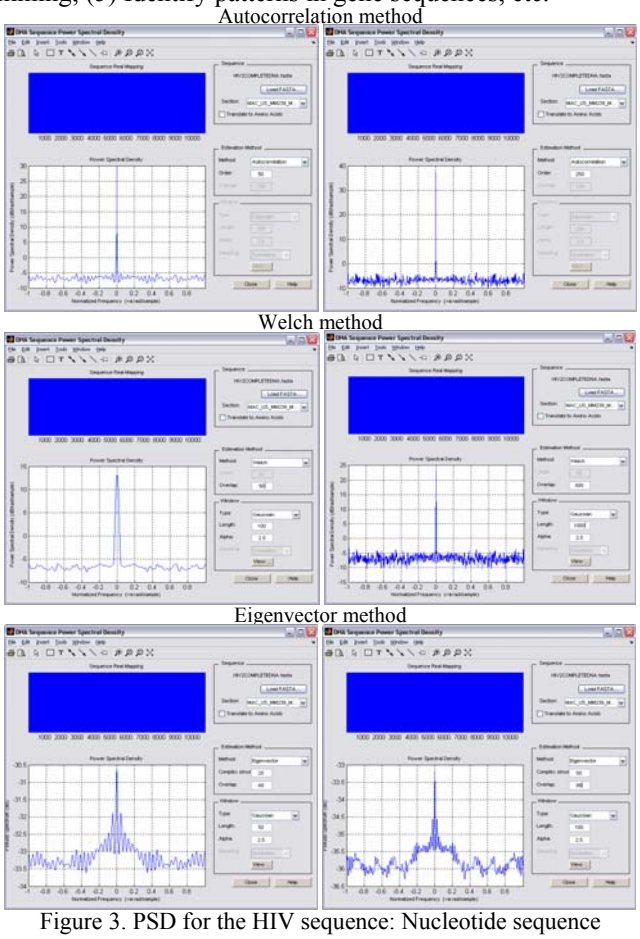
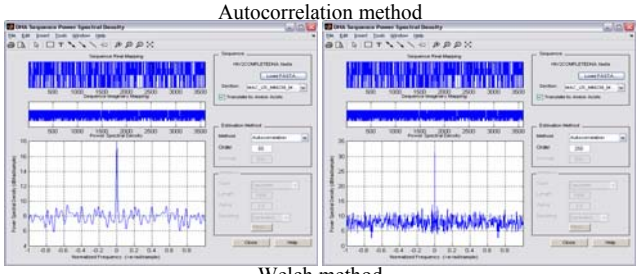


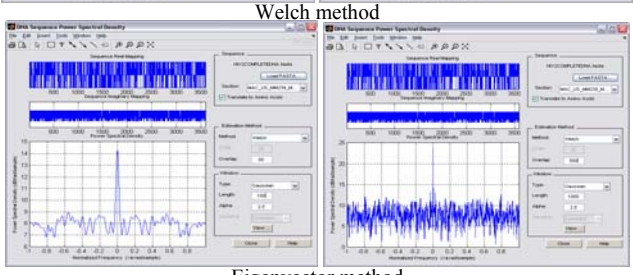
Figure 3. PSD for the HIV sequence: Nucleotide sequence

NSTI-Nanotech 2008, www.nsti.org, ISBN 978-1-4200-8504-4 Vol. 2

Autocorrelation method



Welch method



Eigenvector method

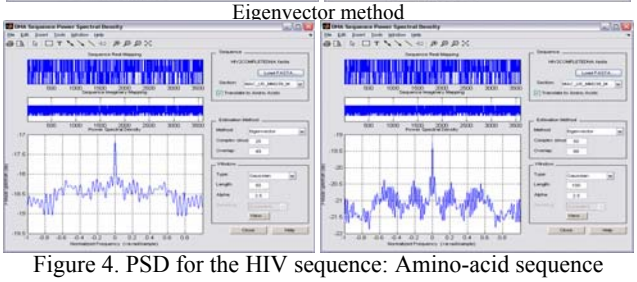
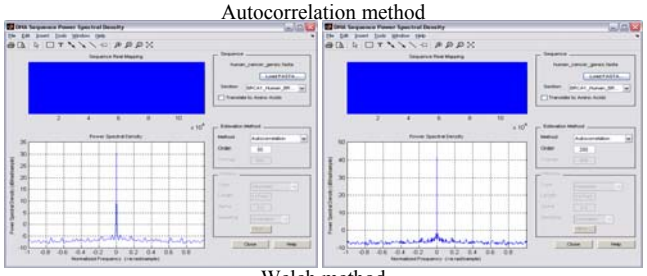


Figure 4. PSD for the HIV sequence: Amino-acid sequence

Autocorrelation method



Welch method

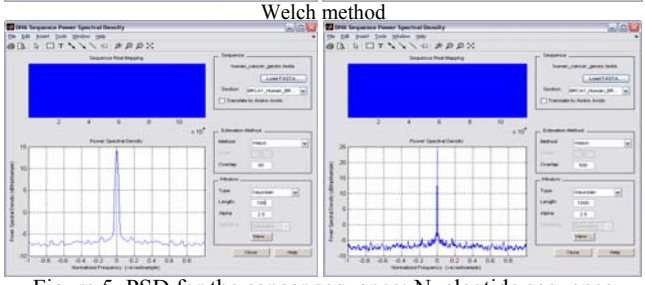
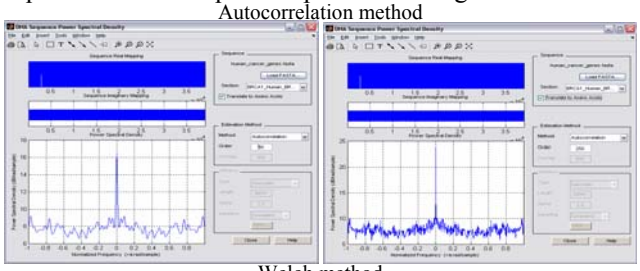


Figure 5. PSD for the cancer sequence: Nucleotide sequence

A data-intensive large-scale study of proteins, their structures and functionality potentially may be quantitatively and qualitatively examined utilizing the *information-theoretic* approach. The proteomic analysis is more complex due to protein diversity and protein-protein interactions [7, 8]. While a genome does not evolve, a proteome differs from cell to cell and undergo changes (modifications, degradations, etc.) through various transitions, interactions and events with the genome and the environment. The number of proteins is much higher than genes. The increased complexity motivates the development of alternative approaches. The solution of this problem will affect the discovery of biomarkers, disease treatments, diagnostics, etc. For example, the genome and proteome information can be utilized to identify or implicate proteins associated with a disease. Specific and customized drugs can be designed to interfere, refine or inactivate the protein functionality. Drugs were found to target and inactivate the HIV-1 protease (an enzyme that cleaves a very large HIV protein into smaller functional proteins). The *information-theoretic* concept promises to perform the homology and matching analyses, data mining, protein-protein "evolutionary" matching, profiling and other tasks using sequenced and unsequenced proteins from genomes.

Autocorrelation method



Welch method

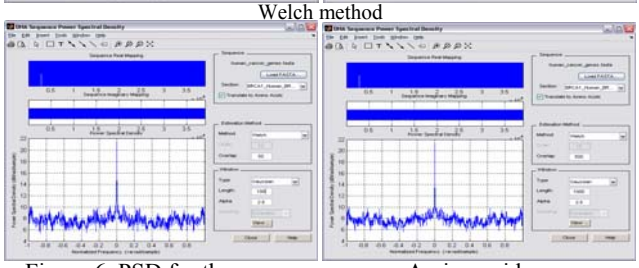


Figure 6. PSD for the cancer sequence: Amino-acid sequence

4. CONCLUSIONS

We proposed solutions to important problems in robust qualitative and quantitative genome and proteome analyses. The frequency analysis was performed to illustrate that complex sequences and patterns can be robustly examined under uncertainties. This analysis provides a viable concept in pattern recognition, identification, prototyping, etc. The proposed approach is valuable due to: (i) Robust homology search and gene detection with high accuracy under uncertainties; (ii) Accurate data-intensive analysis and evaluation; (iii) Analysis of multiagent pathways for multi-genes; (iv) Multifunctional analysis; (v) Computational efficiency and mathematical soundness; (vi) Information extraction and information retrieval; (vii) Large-scale capabilities using multiple databases; (viii) Correlation analysis; etc. The *information-theoretic* approach is found to be consistent, coherent, robust, compact and illustrative. Our approach contributes to *bioinformatics* by developing a sound fundamental concept. The needs for further studies and assessments were emphasized.

REFERENCES

1. S. E. Lyshevski, "Entropy-enhanced genome analysis in frequency domain," *Proc. NanoTech Conf.*, Boston, MA, vol. 2, pp.325-328, 2006.
2. S. E. Lyshevski and F. A. Krueger, "Robust entropy-enhanced frequency-domain genomic analysis under uncertainties," *Proc. IEEE Conf. Nanotech.*, Munich, Germany, pp. 556-558, 2004.
3. K. E. Rudd, "EcoGene: A genome sequence database for Escherichia coli K-12," *Nucleic Acids Res.*, pp. 60-64, 2000.
4. Genome Sequencing Center, Washington University in St. Louis, School of Medicine http://genome.wustl.edu/
5. *Access to Complete Genomes and Proteome Analysis*, European Bioinformatics Institute. http://www.ebi.ac.uk/proteome/ *FASTA – Proteomes Search* www.ebi.ac.uk/fasta33/proteomes.html
6. *HIV Databases*, Los Alamos National Laboratory www.hiv.lanl.gov
7. W. P. Blackstock and M. P. Weir, "Proteomics: quantitative and physical mapping of cellular proteins", *Trends Biotechnol.* , vol 17, no. 3, pp. 121-127, 1999.
8. R. M. Twyman, *Principles of proteomics*, BIOS Scientific Publishers, New York, 2004.