Entropy-Enhanced Genome Analysis in Frequency Domain

Sergey Edward Lyshevski
Department of Electrical Engineering, Rochester Institute of Technology, Rochester, NY 14623-5603, USA
E-mail: Sergey.Lyshevski@rit.edu Web site: www.rit.edu/~seleee

ABSTRACT

We perform an entropy-enhanced frequency-domain analysis to examine large-scale genomic data. This ensures superior qualitative and quantitative coherency. Different statistical methods are used to analyze and evaluate large-scale data performing data mining. These attempts have been partially successful due to sequences gaps, nonconding and *low complexity* regions, inaccuracy, etc. The proposed novel concept complies with the conventional data formats and complements other methods ensuring comprehension of complex large-scale genomic data under uncertainties. The analysis is performed and reported for various genomic sequences (*E.coli*, *S.typhimurium*, HIV, cancer and other).

Keywords: analysis, entropy, frequency domain, genome

1. INTRODUCTION

This paper examines complex genomic patterns with the ultimate objective to provide robust analysis and coherent evaluation. The use of statistical methods in attempt to analyze large-scale data produced by high-throughput experiments have some limitations. Despite the importance of application of bioinformatics to solve problems, efforts to date have been progressed with a limited progress. The mathematical foundations used for incomplete genomic data under uncertainties are obscure. Therefore, we develop and demonstrate the entropy-enhanced frequency-domain concept examining most complex genomic sequences including human genome.

2. FUNDAMENTALS AND STATISTICAL ANALYSIS

Meaningful databases have been developed. For example, 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. The large-scale genomics and proteomics are the forefront of not only biological and genomic research, but also engineering and technology developments. The learning methods (clustering, Bayesian networks, decision trees, neural networks) can be used to study trends and patterns in the large-scale data. 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.

In addition to sequence and structure databases, efforts have been directed focusing on functionality aspects. Integrated data-intensive large-scale analysis and heterogeneous intelligent data-mining are essential. There is a need to develop novel paradigms that will allow one to integrate genomic data from different databases in a common framework. A general problem is to integrate the large-scale diverse genomic information in the viable taxonomies or categories. Currently, the majority of methods are based on the statistical analysis employing unsupervised learning, self-organization, classification, hierarchical clustering, etc. For example, a clustering method ensures multitiered partitioning of the data sets.

Using the Pearson correlation coefficient $r_{ij} = \frac{1}{N-1} X_i \cdot X_j$, given as a dot product of two *profiles* X_i and X_j, the similarity between genes (or groups of genes) is obtained. The aggregation of proteomic data from multiple sources must be performed to identify and predict various protein properties, functionality and features. The DNA sequences of several human pathogens are reported. To achieve reasonable accuracy and high-quality continuous sequences, each base pair was sequenced many times. As a result, 90-93% of the euchromatin sequence has an error rate of less than 1 base per 10,000 bases. Different sequencing technologies, mathematical methods, procedures and measurement techniques have been used. However, it is very difficult to estimate the accuracy, and there are many gaps and unknown strings of bases in the large-scale genomic sequence data. There are differences even in the count of genes. For example, the public human genome database reports 31,780 genes (2,693 million bases sequenced). These include 15 thousands known genes and 17 thousands predicted genes. However, it is estimated that there can be less than 20 thousands actual genes. Some predicted genes can be "pseudogenes" (noncoding) or fragments of real genes leading to predictions that there could be only 7 thousand real genes. For example, Celera reported 39,114 genes (2,654 million bases sequenced) advising that 12 thousand genes are "weak" (http://www.celera.com/). Hence, it is very difficult to identify the disease-associated genes.

Different statistical techniques have been applied to attain global and local sequence comparisons. However, under even the simplest random models and scoring systems, the distribution of optimal global alignment scores is unknown. Monte Carlo experiments potentially can provide some promising results for specific scoring systems and sequence compositions, but these results cannot be generalized. In the BLAST program, the database search is performed utilizing high-scoring segment pairs (HSPs). To analyze the score probability, a model of random sequences is applied. For proteins, the simplest model chooses the amino acid residues in a sequence independently, with specific background probabilities for the various residues, and the expected score for aligning a random pair of amino acid is required

to be negative. For sequence (with lengths m and n), the HSP scores statistics are characterized by the scaling parameters K and λ. The expected number of HSPs with score at least S is given as $E=mnKe^{-\lambda S}$. One obtains the E-value for the score S. However, the length of sequence changes E, and sound methods to find the scaling positive parameters K and λ have not been reported. The number of random HSPs with score greater or equal to S is described by a Poisson distribution. In BLAST, the E-value is used to compare two proteins of lengths m and n. To assess the significance of an alignment that arises from the comparison of a protein of length m to a database containing many different proteins of varying lengths, one assumes that all proteins in the database are *a priori* equally likely to be related to the query. This implies that a low E-value for an alignment involving a short database sequence should carry the same weight as a low E-value for an alignment involving a long database sequence. To calculate a "database search" E-value, one multiplies the pair-wise-compared E-value by the number of sequences in the database using, for example, the FASTA protein comparison programs. The approaches applied to date have a sound theoretical foundation only for local accurate alignments that do not have gaps and short sequences estimating K and λ. Different amino acid substitution scores $S_{ij}=\frac{1}{\lambda}\ln\frac{q_{ij}}{p_i p_j}$ are used. Here, q_{ij} is the target frequency; p_i and p_j are the background frequencies for the various residues. The target frequencies and the corresponding substitution matrix may be calculated for any given distance. However, this method has serious deficiencies. Due to the application of vague mathematical methods, uncertainties and *low complexity* regions lead to unsolved difficulties in sequence similarity searches, e.g., high score results for sequences that are not related, existing matches cannot be found, etc.

3. Software for Statistical Analysis

For the distinct concepts reported in Section 2, specialized software is available. For example, a rudimentary window-based (size is 81) statistical analysis of the *E.coli* genome is performed in the MATLAB toolbox, and the results are summarized in Figure 1.

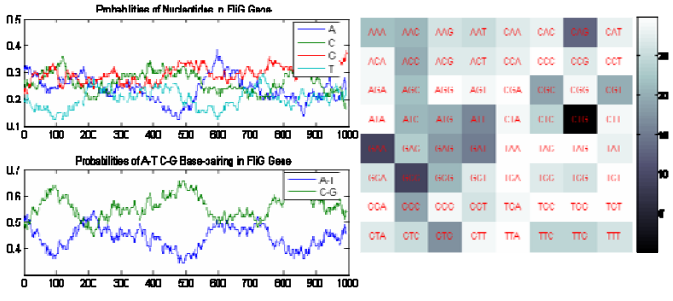

Figure 1. Nucleoties and codons statistics for FliG

4. Fourier Transform

To guarantee robust analysis, we propose to analyze the genomic sequences in the frequency domain using the Fourier transform [1]. Consider a sequence of nucleotides A, T, C and G. We assign the numbers a, t, c and g to the characters A, T, C and G. These a, t, c and g can be complex numbers. There exists a numerical sequence resulting from a character string of length N as
$x[n] = au_A[n] + tu_T[n] + cu_C[n] + gu_G[n]$, n=0,1,2,…, N–1,
where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are the binary indicator sequences.

For amino acids, we have the following expression for the amino acid sequence
$x[n] = A_{la}u_{Ala}[n] + A_{rg}u_{Arg}[n] + \ldots + T_{yr}u_{Tyr}[n] + V_{al}u_{Val}[n]$.

The DFT of a sequence $x[n]$ of length N is

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

$$U_A[k]=\sum_{n=0}^{N-1}u_A[n]e^{-j\frac{2\pi}{N}kn},\ U_T[k]=\sum_{n=0}^{N-1}u_T[n]e^{-j\frac{2\pi}{N}kn},$$

$$U_C[k]=\sum_{n=0}^{N-1}u_C[n]e^{-j\frac{2\pi}{N}kn},\ U_G[k]=\sum_{n=0}^{N-1}u_G[n]e^{-j\frac{2\pi}{N}kn}.$$

If we assign numerical values a, t, c and g, then
$X[k] = aU_A[k]+ tU_T[k] + cU_C[k] + gU_G[k]$, k=0,1,2,…,N–1.

In general, DNA character strings lead to the sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ resulting in four-dimensional representation of the frequency spectrum. The total power spectral content of the DNA character string is $S[k]=|U_A[k]|^2+|U_T[k]|^2+|U_C[k]|^2+|U_G[k]|^2$

For the amino acids, the frequency spectra and power analysis are identical to those reported for DNA. Twenty proteinogenic amino acids are represented as complex-valued functions $z[n]=x[n]+jy[n]$ mapping their structures and properties. One has

$$Z[k]=\sum\left(x[n]+jy[n]\right)e^{-j\frac{2\pi}{N}kn}=X[k]+jY[k].$$

5. Applications of Fourier Transform

The application of the Fourier transform is reported in Figure 2 for complete *E.coli* and *S.typhimurium* genomes with 4,639,221 and 4,937,381 base pair strains [2, 3]. The nucleotide pattern for these bacteria is completely distinct. It is virtually impossible to analyze patterns using statistical methods. In contrast, the DFT is effectively applied providing meaningful results.

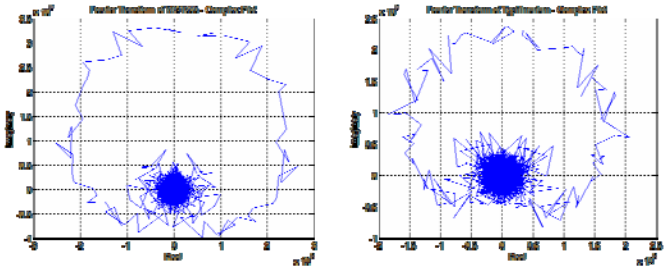

Figure 2. Fourier transforms for *E.coli* and *S.typhimurium* genomes

6. Frequency-Domain Analysis Under Uncertainties

In general, gene sequences may not be complete as there can be many missed sites. The HIV genes are typical examples [4, 5]. Correspondingly, statistical methods cannot be applied, and linear maps cannot be found. The frequency analysis of sequences promises to solve a spectrum of problems such as: examine and identify protein coding genes in genomic DNA, detect genes, define structural and functional characteristics, analyze the data, identify patterns in gene sequences, etc.

A high-performance interactive software has been developed in the MATLAB environment to support robust frequency-domain analysis. We utilize the power spectral density (PSD) analysis applying different methods of PSD estimation (covariance, multiplier, periodogram, etc.). For example, Welch method is based on dividing the sequence of data into (possibly overlapping) segments, computing a modified periodogram of each segment, and averaging the PSD estimates. That is, we consider

$$x_m[n] = x\left[\frac{N}{M}m - \frac{L}{2} + n\right], \quad n = 0,1,2,...,L-1$$

to be the mth segment of the sequence $x \in C^N$ divided into M segments of length L. The Welch PSD estimate is given as $R_x = \left\{ \left| X_m[k] \right|^2 \right\}_m$, where $\{\cdot\}_m$ denotes averaging across the data segments.

Figure 3 illustrates the power spectra of the DNA sequence of the human gene CISH [5]. Lung and kidney tumors frequently exhibit deletions of this gene.

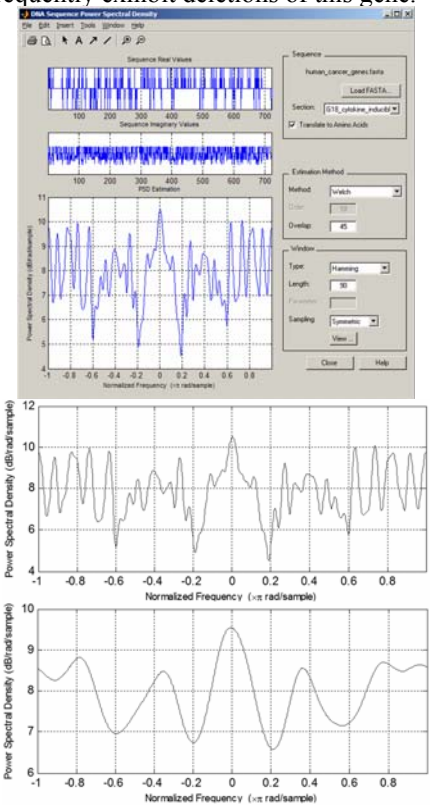

Figure 3. Interactive software: Power spectral density of the human gene CISH (3p21.3) and two estimated PSDs

Using distinct methods, the results of the application of the developed interactive software for *E.coli* (genome sequence O157:H7, AE005174-1, segment 1), HIV2 and human cancer genes are reported in Figures 4, 5 and 6.

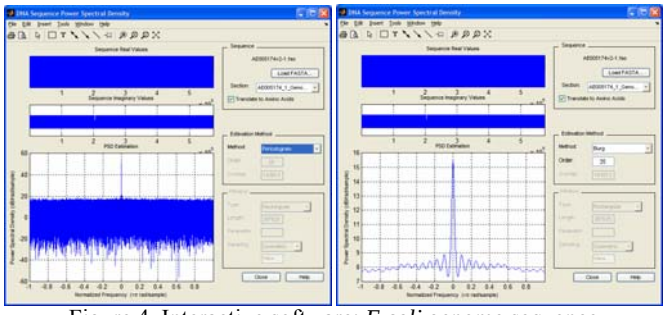

Figure 4. Interactive software: *E.coli* genome sequence

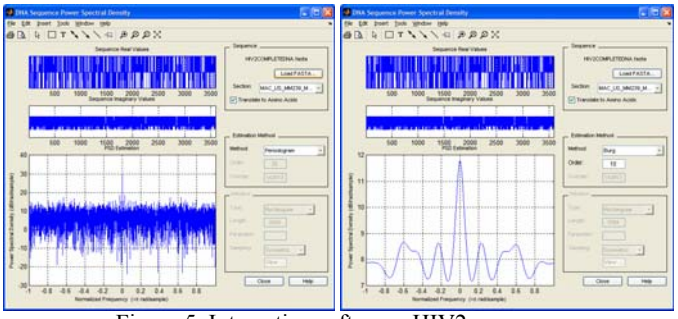

Figure 5. Interactive software: HIV2 sequence

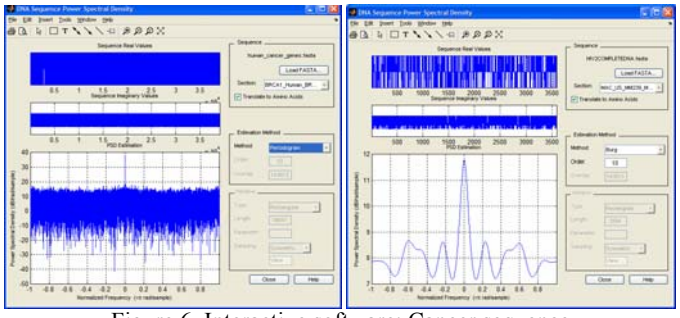

Figure 6. Interactive software: Cancer sequence

7. Fourier Transform and Genomic Pattern Analysis

We use the PSD estimation to distinguish genomic sequences versus non-genomic sequences. Figure 7 reports four plots. The first is the estimated PSD of the *E.coli* gene FliG as a standalone gene. The next three are the estimated PSDs for the FliG gene surrounded by other nucleotides. In particular, for the second PSD, we consider FliG surrounded by the FliM and FliN genes. In the third plot, FliG is surrounded by random nucleotides. The fourth plot reports PSD for the nucleotides from *E.coli* genome and FliG. The documented results demonstrate very distinct PSDs for a standalone gene, three genes, and gene-nucleotide sequences. Thus, the proposed concept allows one to distinguish genomic versus non-genomic sequences.

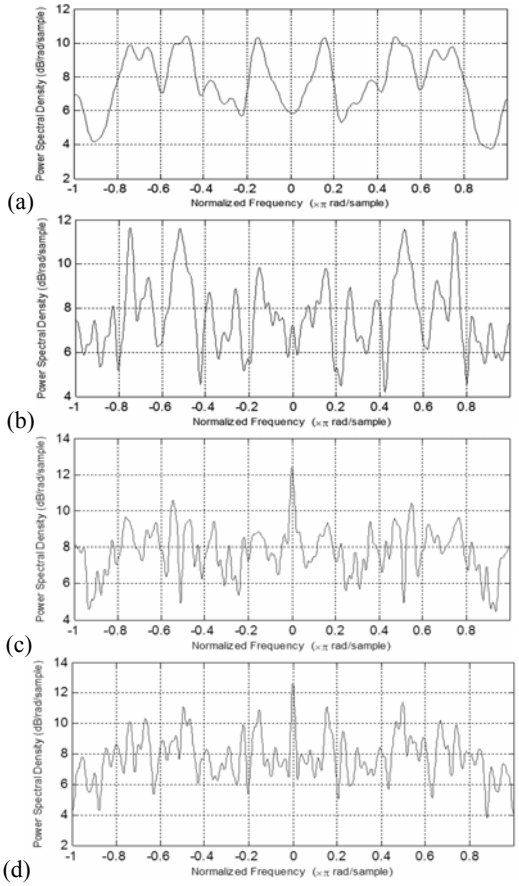

Figure 7. PSD plots of the sequences: (a) Standalone FliG; (b) FliM-FliG-FliN genes; (c) Random nucleotides-FliG-random nucleotides; (d) Nucleotides from genome-FliG-nucleotides

8. ENTROPY ANALYSIS

The entropy, which is the Shannon quantity of information, measures the complexity of the set. The uncertainty after binding for each site (Shannon entropy of position l) is $H(l) = -\sum_{b \in \mathbf{A}} f(b,l) \log_2 f(b,l)$, where **A** is the cardinality of the four-letter DNA alphabet, **A**={A,C,G,T}; $f(b,l)$ is the frequency of base b at position l. For DNA, the maximum uncertainty at any given position is $\log_2\mathbf{A}$=2 bits. For amino acids, the alphabet is **A**={Ala, Arg, …, Tyr, Val}. Therefore, for amino acids, the maximum entropy at any given position is $\log_2\mathbf{A}$=4.32 bits. Using the entropy $H(l)$, one derives the information at every position in the site. In particular,

$$R(l) = \log_2 \mathbf{A} - \left(-\sum_{b \in \mathbf{A}} f(b,l) \log_2 f(b,l) \right).$$

The total amount of pattern in ribosome binding sites is found by adding the information from each position, e.g., $R_\Sigma(l) = \sum_l R(l)$ bits per site. For *E.coli* and *Salmonella typhimurium* one finds 11.2 and 11.1 bits per site. We apply probability methods to study *E.coli* and *S.typhimurium* genomes [1]. Our ultimate goal is to apply fundamental mathematical methods to identify interesting sections of a genome including the *low complexity* regions. Examining DNA as a coding system, it is shown that distinct DNA segments have different entropy. In general, entropy depends on the probability model attributed to the source. Repetitions (*low complexity* segments) have low entropy. Figure 8 presents the entropy of verified gene sequences for *E.coli* EDL933 with 5476 genes [2, 3]. A similar entropy analysis is performed and reported for the *S.typhimurium* genome with 4596 genes in Figure 8. Low and high complexity regions in genomes are found. This entropy concept is applied to the entropy-enhanced frequency analysis.

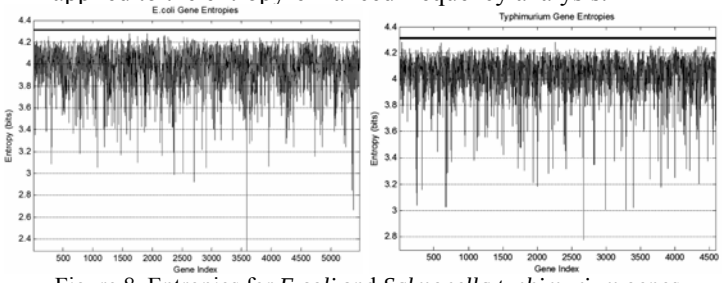

Figure 8. Entropies for *E.coli* and *Salmonella typhimurium* genes

9. CONCLUSIONS

We proposed the solutions to important problems in robust qualitative and quantitative genome analysis. The frequency and entropy analyses were performed to illustrate that the template patterns can be robustly examined in the frequency domain under uncertainties. This analysis provides a viable method in pattern recognition, prototyping, synthesis, etc. The proposed concept is valuable due to: (i) robust homology search and gene detection (identification) with high accuracy and robustness under uncertainties; (ii) accurate and robust data-intensive analysis and evaluation; (iii) analysis of multiagent pathways for multi-genes and multifunctional standpoints; (iv) superior computational efficiency and mathematical soundness; (v) coherent information extraction and information retrieval; (vi) correlation between large-scale multiple databases. These results demonstrated the utility of the application of the frequency- and entropy-domain analysis as compared with the conventional approaches. The frequency domain maps are shown to be robust, compact and illustrative. The concept was tested and software was developed.

ACKNOWLEDGEMENTS – The author acknowledges the contribution of F. Krueger in the software developments during his graduate studies at RIT (2003-2004).

REFERENCES

1. 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.
2. K. E. Rudd, "EcoGene: A genome sequence database for Escherichia coli K-12," *Nucleic Acids Res*., pp. 60-64, 2000.
3. Genome Sequencing Center, University of St. Louis, 2004. http://genome.wustl.edu/projects/bacterial/styphimurium/
4. *Proteome Analysis*, European Bioinformatics Institute, 2004. http://www.ebi.ac.uk/proteome/
5. *HIV Databases*, Los Alamos National Laboratory, 2004. http://www.hiv.lanl.gov