Nanoengineering Bioinformatics: Fourier Transform and Entropy Analysis

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

Abstract **– In this paper, we apply and enhance the cornerstone theoretical fundamentals of engineering bioinformatics to complement nanotechnology. In particular, nanoengineering bioinformatics is examined and formulated as a coherent abstraction in cognitive analysis of complex inorganic, organic and hybrid nanosystems. We report the application of an entropy-enhanced frequency-domain analysis concept to examine large-scale genomic data. This ensures superior coherency for qualitative and quantitative analysis. Conventionally, bioinformatics emphasizes the application of statistical methods attempting to analyze large-scale data produced by high-throughput experiments including complex gene sequencing. It is illustrated that bioinformatics can be expanded to a systems-based perspective by making use of novel concepts thereby positioning bioinformatics to play a significant role in engineering and technology. It is our goal to evolve the nanoengineering bioinformatics to coherently analyze the genomic data identifying, qualifying and quantifying complex genes in functional biological systems. The ultimate goal for the application of nanoengineering bioinformatics is in the development of system-level knowledge in order to devise novel paradigms for discovering entirely new systems with superior functionality and performance. In contrast, biomedical informatics examines the data from a more narrow-focused perspective and focuses on data and knowledge integration to analyze the biological processes. To enable the integration between computational, experimental, stochastic and deterministic modeling, novel information-theoretical methods should be applied to guarantee coherent representation and possible evaluation from organic to hybrid systems. These methods must be robust and utilize incomplete and inaccurate information, sequencing gaps, noncoding regions, unsolved interactions, multiple modeling hierarchies, unknown phenomena, lack of information, etc. It is demonstrated that the proposed entropy-enhanced frequency-domain concept promises to solve a number of long-standing problems. The reported paradigm complements a number of far-reaching perceptions of engineering bioinformatics.**

I. Introduction

Nano- and microscale biological systems exist in nature in enormous variety and sophistication. These complex biological patterns can be applied to devise, analyze, and examine distinct organic, inorganic, and hybrid molecular-scale systems. Bioinformatics emphasizes the application of statistical methods attempting to analyze large-scale data produced by high-throughput experiments as well as accomplish data mining [1-6].

One cannot blindly copy biosystems due to the fact that many complex phenomena and effects have not been comprehended, and system architectures and functionalities have not been fully examined. The protein geometry and functionality for even many simple proteins have not been identified. Typical examples include unsolved problems to comprehend the simplest single-cell *Helicobacter pylori* (*H.pylori* is the most prevalent bacterial pathogen), *Escherichia coli* (*E.coli*) and *Salmonella typhimurium* bacteria [7]. The genome size of *H.pylori* strain 26695 is 1,697,867 bp, and 1,643,831 bp for strain J99. The number of base pairs is different for all bacteria, and this number also depends on the strain (MG1655, CFT073, EDL933 and other *E.coli* strains). For *E.coli*, this paper examines 4,639,221 and 5,528,445 bp strains, as well as 4,857,432 bp strain of *Salmonella typhimurium* [8, 9]. Though the bacteria genomes are different, there are many common features. These bacteria (only up to 5 μm long) exhibit a remarkable level of sophistication and functionality. In particular, these bacteria integrate three-dimensional nanobiocircuitry, computing – processing – networking nanobioelectronics, nanobiomotors, nanabiosensors, etc. The above mentioned nanoscale subsystems and devices can be used as prototypes for devising organic three-dimensional nanoICs, inorganic nanomachines, nanosensors, etc.

Significant research efforts have been dedicated to examine biosystems, their subsystems and components, e.g., organic windings, three-dimensional networking – processing – computing bioelectronics, biosensors, etc. The fundamental, applied and experimental research has been progressed to understand and control molecular-scale processes to attain the controlled synthesis and direct self-assembly of functional structures and components into functional nano- and microscale devices and systems. These novel molecular-scale devices and systems can be used in nano-electronics, computing, information processing, wireless communication, sensing, actuation, propulsion, energy sources, etc. Despite the importance of applying bioinformatics to unsolved problems, efforts to date have met with limited progress. This paper, therefore, examines engineering bioinformatics to approach and solve different problems in systems analysis and design. Nanoengineering bioinformatics is a coherent abstraction in

devising, prototyping, design, optimization and analysis of complex systems that will allow one to effectively apply nanotechnology. Our attention is concentrated on devising novel paradigms in systematic analysis through bioinformatics with the ultimate objective to examine and possibly fabricate these systems. The proposed nanoengineering bioinformatics concept promises to provide one with the methodology to analyze complex genomes, derive new operating principles, examine complex hierarchies, comprehend functionality of different subsystems, research novel structures, study advanced architectures (topologies), characterize distinct systems / subsystems / devices and reach out other tasks, thereby defining the *nanoarchitectronics* horizon.

This paper examines complex genomic patterns in biosystems because superior inorganic and hybrid systems can be devised and analyzed through engineering bioinformatics. Our ultimate objective is to provide the focused study of nanoengineering bioinformatics and systematic analysis. These are far-reaching frontiers of modern nanoscience and nanoengineering. The synergetic paradigm reported is demonstrated researching biosystems and coherently examining their genomic content. We develop and demonstrate entropy-enhanced frequency domain analysis with applications.

II. Application of Fourier Transform in Nanoengineering Bioinformatics

2.1. Fourier Transform in Genome Analysis

To guarantee robust analysis, we propose to analyze the genomic sequences in the frequency domain. The Fourier transform is used. This concept offers superior computational advantages, accuracy, versatility and coherence in examining complex genomes composed from billions of experimentally obtained nucleotides (bases). It is demonstrated with this technique, the sequenced human genome can be examined. This frequency domain analysis of DNA and amino acid sequences is an important task in experimental studies. In fact, protein coding genes in genomic DNA can be identified apart from non-coding regions performing frequency-domain data-intensive analysis and data mining. Our goal is to develop a concept that will allow one to identify the protein coding genes, define structural and functional characteristics, analyze patterns in gene sequences, etc.

The genomic sequences are available with the satisfactory accuracy. In particular, the strings of nucleotides (bases) As, Cs, Gs and Ts or amino acids Ala, Arg, … , Tyr and Val have been found.

We consider a sequence of nucleotides A, T, C and G. Let us assign the number a to the character A, the number t to the character T, the number c to the character C, and the number g to the character G. These a, t, c and g can be complex numbers. There exists a numerical sequence resulting from a character string of length N. In particular,

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

where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are the binary indicator sequences (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.

For amino acids, we have the following expression for the amino acid sequence

$$x[n] = A_{la}u_{Ala}[n] + A_{rg}u_{Arg}[n] + \ldots + T_{yr}u_{Tyr}[n] + V_{al}u_{Val}[n],\quad n = 0, 1, 2, \ldots, N-1.$$

The discrete Fourier transform (DFT) of a sequence $x[n]$ of length N is

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

This DFT provides a measure of the frequency content at frequency k which corresponds to a period of N/k samples. The resulting sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ are the discrete Fourier transforms of the binary indicator sequences $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$. E.g.,

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

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

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

If we assign numerical values a, t, c and g, then

$$X[k] = aU_A[k] + tU_T[k] + cU_C[k] + gU_G[k],\ k=0,1,2,\ldots,N-1.$$

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 with

$$U_A[k] + U_T[k] + U_C[k] + U_G[k] = \begin{cases} 0, k \neq 0 \\ N, k = 0 \end{cases}.$$

The total power spectral content of the DNA character string at the frequency k 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.

Consider amino acids. Each amino acid has a hydrogen atom, a carboxyl group, and an amino acid group bonded to the α-carbon. It is important to note that all 20 amino acids that make proteins differ only by what is attached by the fourth bond to the α-carbon. The amino acids are grouped according to the properties of the side chains (R-group). Physical and chemical properties of the side chain define

the unique characteristics of amino acids. Amino acids are classified based on the chemical and structural properties of their side chains and polarity, e.g., nonpolar (hydrophobic), polar (hydrophilic), electrically charged, etc. Twenty proteinogenic amino acids, as well as their structures and properties are represented as complex-valued functions, e.g.,

$$z[n] = x[n] + jy[n].$$

Therefore, we have

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

Correspondingly, the complex Fourier transform is used.

2.2. Fourier Transform in Analysis of Escherichia Coli and Salmonella Typhimurium Genomes

First, to illustrate the proposed method, we examine a gene from a complete 4,639,221 bp *E.coli* genome. This genome is comprised of 4,293 genes. Figure 1 reports the frequency spectrum of amino acids of the gene FliM.

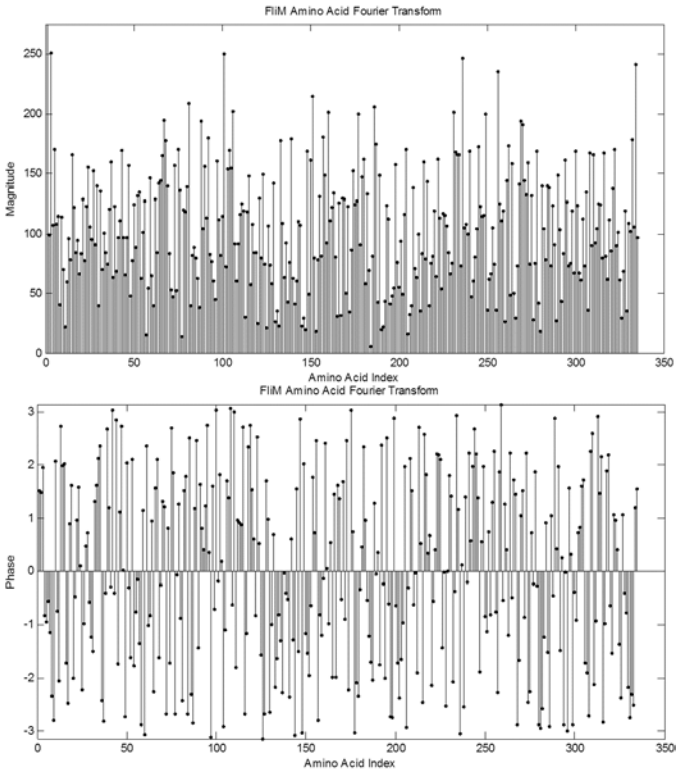

Figure 1. Right-hand-side Fourier transform of FliM gene for the *E.coli* MG1655 genome: magnitude and phase plots

The application of the Fourier transform is reported in Figure 2 for complete *E.coli* (strain EDL933) and *Salmonella typhimurium* genomes. These genomes are 5,528,445 and 4,857,432 bp in size, respectively [8, 9].

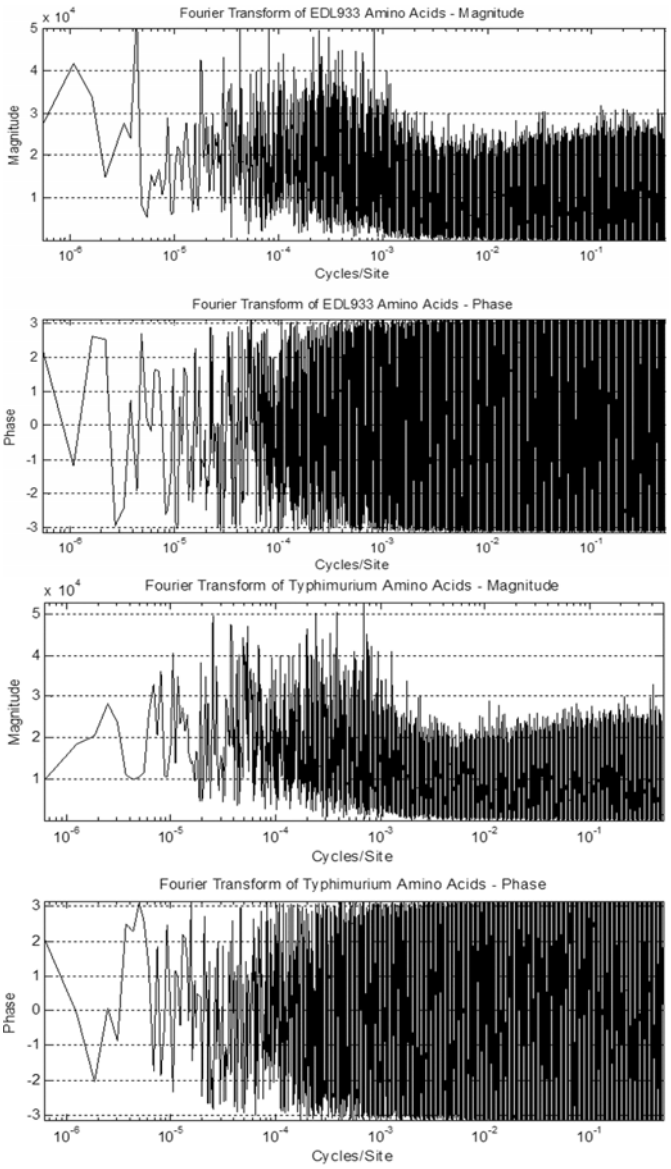

Figure 2. Fourier transforms for complete *E.coli* and *Salmonella typhimurium* genomes

The nucleotide (and amino acid) patterns for these bacteria are distinct, and it is difficult to analyze patterns using statistical methods. In contrast, the Fourier transform can be effectively applied as it does not depends on the sequence's size. The resulting frequency domain maps are compact compared with the linear sequences.

In general, gene sequences may not be complete as there are many missed sites. The HIV genes are typical examples [10]. Correspondingly, statistical methods cannot be applied, and linear maps cannot be found.

The frequency analysis of sequences promises to solve a spectrum of problems, e.g., examine and identify protein coding genes in genomic DNA, identify the protein coding genes, detect genes, define structural and functional characteristics, analyze the data, identify patterns in gene

sequences, etc. To demonstrate the proposed approach, we apply the mathematical basics reported in Section 2.1. The components of the high-performance interactive software have been developed to support robust frequency-domain analysis. In particular, to examine the power spectral density (PSD), we apply different methods (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 then averaging the PSD estimates.

That is, we consider

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

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 segments of data.

Figure 3 illustrates the power spectra of the DNA sequence of the human gene CISH [11]. Lung and kidney tumors frequently exhibit deletions of this gene. In addition, to perform the spectral analysis, two distinct methods (covariance and Welch) are utilized in Figure 3.

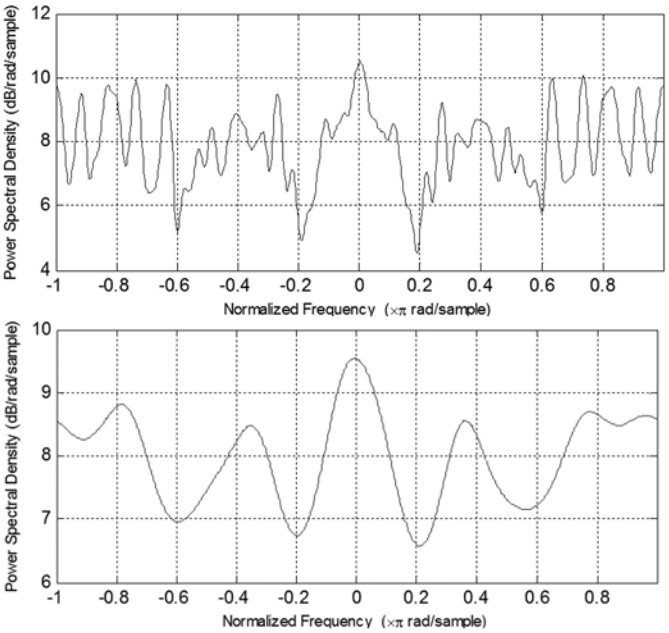

Figure 3. Power spectral density of human gene CISH (3p21.3) estimated using the covariance and Welch methods

We examine the ability of the Welch method, as power spectral density estimation, to distinguish genomic sequences versus non-genomic sequences. In Figure 4, four plots are given. 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 the second PSD plot, consider FliG as 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. It is important that the proposed concept allows one to distinguish genomic versus non-genomic sequences.

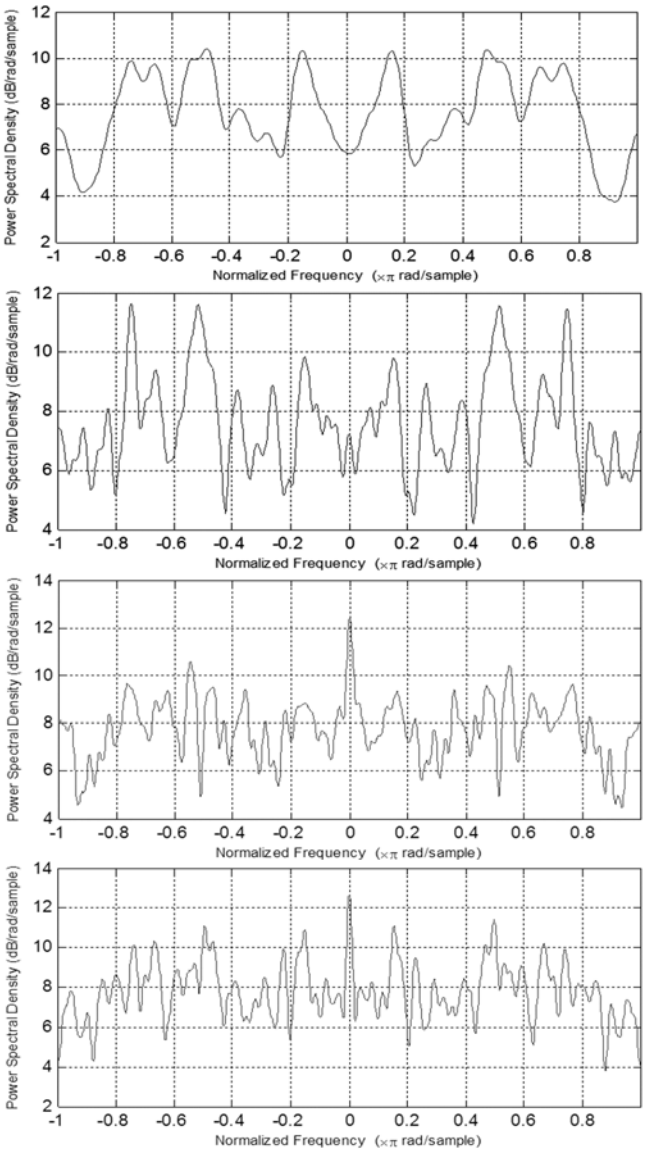

Figure 4. PSD plots of the sequences: standalone FliG; FliM - FliG – FliN genes; random nucleotides - FliG - random nucleotides; nucleotides from genome - FliG - nucleotides from genome

2.3. Autocorrelation Analysis

The deterministic autocorrelation sequence $r_{xx}[n]$ of a sequence $x[n]$ is given as

$$r_{xx}[n] = \sum_{k=-\infty}^{\infty} x[k]x[n+k],\ n = 0,\ 1,\ 2,\ \ldots,\ N-1,$$

where $x[n]$ is a sequence of either nucleotides or amino acids.

The autocorrelation sequence measures the dependence of values of the sequence upon values at different positions

in the sequence. A finite random sequence has an autocorrelation sequence of all zeros with the exception of a single large value at zero. We examine the "randomness" of the studied protein sequence applying the autocorrelation analysis. Figures 5 and 6 report the autocorrelation sequences of the amino acid sequences for the FimB and TerA genes.

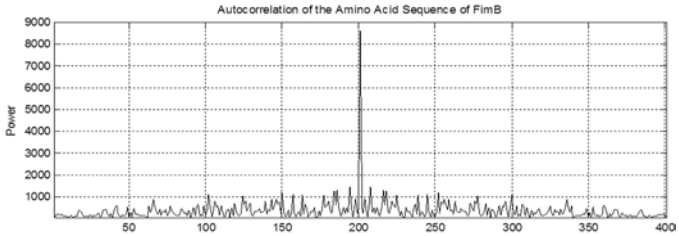

Figure 5. Autocorrelation of the amino acid sequence for FimB gene from *E.coli* EDL933

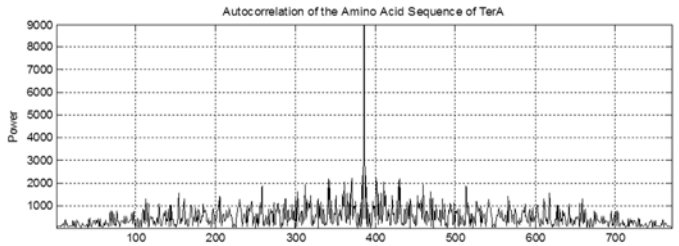

Figure 6. Autocorrelation of the amino acid sequence for TerA gene from *E.coli* EDL933

The cross-correlation analysis is performed using the equation

$$r_{xy}[n] = \sum_{k=-\infty}^{\infty} x[k]y[n+k].$$

This analysis is of our particular interest for its potential to identify similar genes. Figure 7 reports the cross-correlation of FimB and FimE proteins. The correlation coefficients $\rho_{xy}[n] = r_{xy}[n] \Big/ \sqrt{r_x[0] r_y[0]}$ provide a normalized measure of correlation. If $\rho_{xy}[n] > 0$, the sequences are positively correlated, while correlation coefficients of 0 denote uncorrelated sequences. Thus, the genes patterns can be identified and recognized using $\rho_{xy}[n]$.

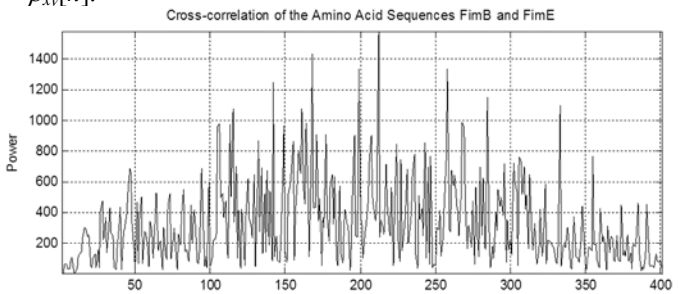

Figure 7. Cross-correlation of the FimB and FimE genes from *E.coli* EDL933

III. Entropy Analysis

During DNA translation, individual ribosomes locate specific sequences of DNA through pattern matching to form the amino acid sequence. The fundamental question is how much information is needed to describe the patterns. Before binding, the ribosome's sites have four possibilities and do not distinguish between them. Thus, each site is uncertain by $\log_2 4=2$ bits. After binding, the uncertainty at each site is lower, e.g., $\log_2 1=0$ bits. 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} and $f(b, l)$ is the frequency of base b at position l.

It should be emphasized that, 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 Shannon entropy $H(l)$, one derives the information at every position in the site as

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

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, respectively. There is enough pattern at ribosome binding sites for them to be found in the genetic material of the cell (there is no excess and no shortage of patterns).

We apply probability methods to study *E.coli* and *Salmonella typhimurium*. Our ultimate goal is to apply fundamental mathematical methods to identify interesting sections of a genome including the so-called 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. For example, if the DNA or amino acid patterns are repeatable, the sequences have low entropies.

Consider how the frequencies $f(b,l)$ shall be determined. The most direct means of doing this is to calculate the frequency at which each of the nucleotides (or amino acids) occur in a given *window*. Such *windows* can be smaller than the entire genome but large enough to capture representative statistics for a given segment. To identify the different and important sections of a genome, one can slide this window over the genome and estimate the entropy within it. The choice of the size of the *window* and the amount by which it should slide remains. There is no mathematical basis for determining the window and step

sizes. It was shown that a *window* size directly influences the derived entropies [12].

We use genes in the complete genome to define the *window* sizes and positions. The EcoGene database [8] provides the required boundaries for *E.coli* while other databases are available for other genomes [9]. These databases define specific start and stop positions for genes within the entire genomes. We use the sizes of various genes as variable *window* sizes, and step the *window* one gene at a time. Figure 8 presents the entropy of verified gene sequences for *E.coli* EDL933 – 5476 genes total. A similar analysis is performed and reported for the *Salmonella typhimurium* genome (4596 genes) in Figure 9.

By making use the entropy analysis performed, one can determine low and high complexity regions in genomes. This entropy concept can be applied in the entropy-enhanced frequency analysis.

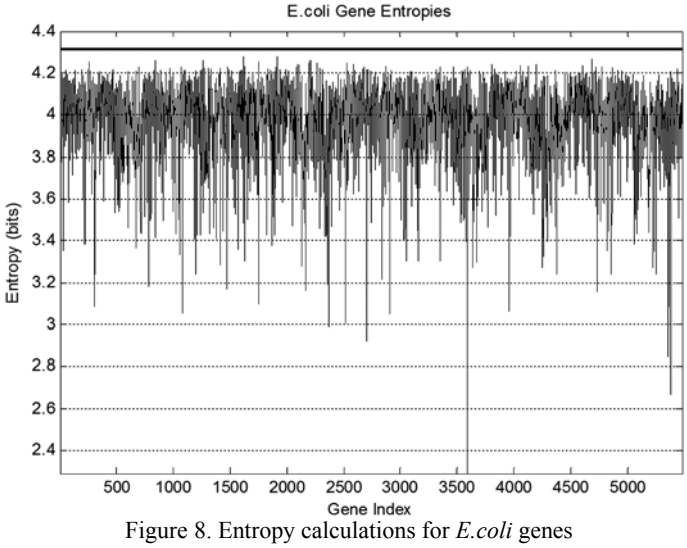

Figure 8. Entropy calculations for *E.coli* genes (maximum entropy is 4.32)

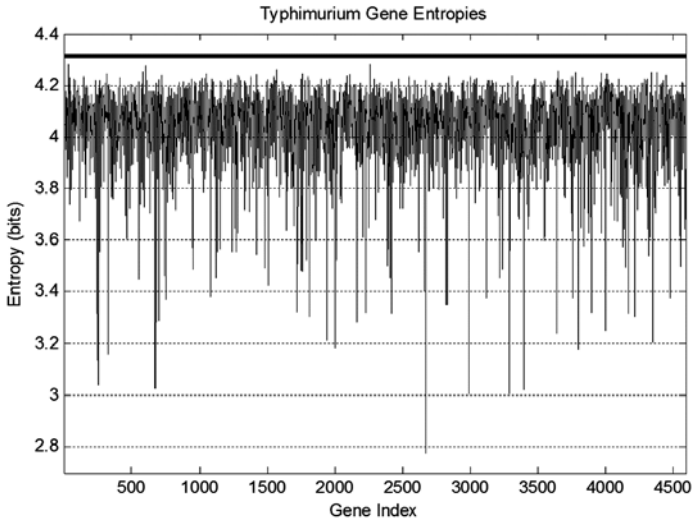

Figure 9. Entropy calculations for *Salmonella typhimurium* genes

IV. Conclusions

This paper approaches the very important problems in analysis of nanobiosystems applying the nanoengineering bioinformatics paradigm. Distinct biosystems and their subsystems have been examined to show unique patterns. The frequency and entropy analyses were performed to illustrate that the template patterns should be robustly examined in the frequency domain. This analysis promises to provide a viable method in pattern recognition, functionality analysis, optimization, prototyping, synthesis, etc. The results reported document that the proposed paradigm is valuable due to: (i) robust homology search and gene detection with superior accuracy and robustness under uncertainties; (ii) accurate and robust data-intensive analysis and decision-making; (iii) analysis of multiagent pathways for multi-genes and multifunctional analysis; (iv) superior computational efficiency and mathematical soundness; (v) coherent information extraction and information retrieval; (vi) correlation between large-scale multiple databases.

References

[1] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases", *Comput. Chem*., vol. 17, pp. 149-163, 1993.

[2] P. Bertone and M. Gerstein, "Integrative data mining: The new direction in bioinformatics," *IEEE Engineering in Medicine and Biology*, no. 4, pp. 33-40, 2001.

[3] J. M. Claverie and D. J. States, "Information enhancement methods for large-scale sequence-analysis", *Comput. Chem*., vol. 17, pp. 191-201, 1993.

[4] G. Lusman and D. Lancet, "Visualizing large-scale genomic sequences," *IEEE Engineering in Medicine and Biology*, no. 4, pp. 49-54, 2001.

[5] G. H. Gonnet, M. A. Cohen and S. A. Benner, "Exhaustive matching of the entire protein sequence database", *Science*, vol. 256, pp. 1443-1445, 1992.

[6] S. Slaab (editor), "Mining information for functional genomics," *IEEE Intelligent Systems*, no. 3, pp. 66-80, 2002.

[7] H. C. Berg, "The rotary motor of bacterial flagella," J. Annual Rev. Biochemistry, vol. 72, pp. 19-54, 2003.

[8] K. E. Rudd, "EcoGene: A genome sequence database for Escherichia coli K-12," Nucleic Acids Res., vol. 28, pp. 60-64, 2000. http://bmb.med.miami.edu/EcoGene/EcoWeb/

[9] Genome Sequencing Center, University of St. Louis, 2004. http://genome.wustl.edu/projects/bacterial/styphimurium/

[10] HIV Databases, Los Alamos National Laboratory, 2004. http://www.hiv.lanl.gov

[11] Proteome Analysis, European Bioinformatics Institute, 2004. http://www.ebi.ac.uk/proteome/

[12] S. E. Lyshevski, Frank A. Krueger and Elias Theodorou, "Nanoengineering bioinformatics: Nanotechnology paradigm and its applications," Proc. IEEE Conference on Nanotechnology, San Francisco, CA, pp. 896-899, 2003.