Title: Digital signal processing post-hybridization of DNA Biosensor

Samuel A. Afuwape, Ph.D., SOET, National University, San Diego, CA 92026

Abstract

The non-label DNA biosensor detects hybridization due to enhanced biomolecules surface-binding specifically on glass or plastic platform, now broadly extended to active surface such as field-effect transistors with scalability of the platform gate. The latter platform provides opportunity to compute the transfer functions of the activities on the functionalized surface pre and post DNA hybridization. In this paper, the discrete transfer functions are derived and exploited to compute and display the codon regions while the computed Fourier response provided the power spectral density at $2\pi/3$ frequency domain. The DNA sequences are digitally filtered to narrow bandpass center at the coding region. Furthermore digital signal processing DSP applications are propounded for exploration into customized pharmaceutical, medical, environmental, agriculture, and defense, based on oligonucleotide DNA hybridization.

Key words: non-label DNA biosensor, field-effect transistors, discrete transfer functions, power spectral density, coding region, DSP applications.

Background

All living cells go through unique biological processes; different subsets of their genomics are expressed in different stages of the processes, specificity of the gene expressed their relative abundance are crucial to a single cell's proper function. Human Genome Project has thus far presented unprecedented enormous challenging data yet to be analyzed or functionally consensus to phenotypes. Therefore, the need arises to develop concepts and methodologies to integrate this huge genomic data to compact analytical tools, such as evolving non-labeled oligonucleotides DNA biosensors [1], see Figure 1. For example, identifying short sequence of critical genes expression involved in diseases, diagnoses, treatments, vaccines, and prognoses. One can real-time track and identify instructional coding, for logical processing and applications association. The goal is short sequence identification, analysis, comparison and application recognition see Figure 2:

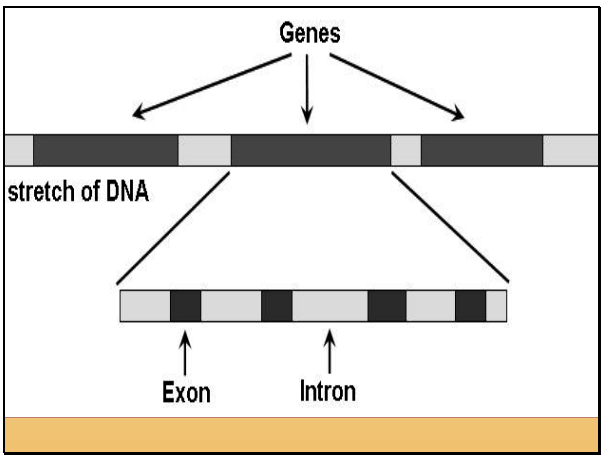

Figure 2: Conceptual Schematic of Extraction of Active Genomic sequence exons.

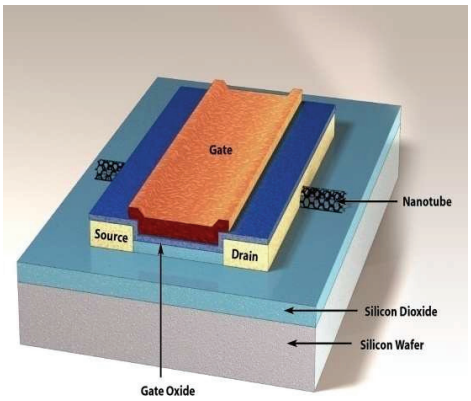

Figure1. Schematic Active Biosensor [Acc. Chem. Res. 35, 1026(2002)]

DNA genome of short sequence of bases (A, G, T, and C) labeled motif, in a repeated sequence and in variable number of times, open up genetically differential indicator of individual differences. These chromosome motif sequences can vary in repeats from 7 bases to 40 bases. Those repeats, the loci, (Figure 2) are termed variable number tandem repeats (VNTR). In addition, shorter loci variable motif (sequence) labeled short tandem repeats (STR), are highly discernable specific fragmental individual genomic. The applications of VNTR/STR are prolific; explore by investigative institutions such as FBI/CIA and major law enforcement organizations around the planet.

Statistical analysis of sequential data is abundant in the literature. For example, clustering algorithms have been used in partitioning genomic data [2]. Furthermore, Pearson correlation coefficient $r = (1/N - 1)X_i.X_j$ computes similarity of clustered gene groups. Fortunately, common DNA of human pathogen is accurately known, due to high-quality continuous sequence at a low error rate of 1: 10,000 bases [3]. Nevertheless, alternatively, one can develop compact real-time data-intensive robust analytical tool using frequency-domain analysis [4].

Method

The interfacial equivalent circuit model of DNA hybridization on modified diamond thin film simulation provides an illustration for analysis of dynamics interfacial system to extract dynamic parameters, such as settling time, rise time, and established system stability during design process of device like DNA biosensor. The dynamics and/or design optimization of the interfacial DNA hybridization can be studied in the time and frequency domains, respectively []. Application of optimization provides a path to real-time detection of DNA hybridization. The essential parameters were extracted from frequency

analyses of the equivalent impedance circuit models, illustrated in the following:
Computed Simulation of pre-hybridization on denatured DNA on diamond surface base data derived [1]: Rdl = 160 k Ω, double layer resistance; Rct = 0.42 M Ω, charge transfer resistance; Rsol = 68.0 Ω, bulk resistance; Cd1 = 80 nF, double–layer capacitance; $Z(s) = T(s)p$, CPE: constant phase element, T = 0.32E–6, p = 0.68.:

$$\text{Transfer function: } H(s) = \frac{3.656e^{005} + 6.723e^{010}}{5376s + 16.1}$$

zero = –1.8390e+005; pole = –0.0030; open loop gain (k) = 68.0.
Computed Simulation for post-hybridization of DNA on diamond surface base data derived [1]: Rdl = 104 kΩ; Rct = 0.41 MΩ; Rsol = 27.0Ω ; C1 = 115.0 nF; $Z(s) = T(s)^p$, T = 0.19E–6; p = 0.729;

$$\text{Transfer function: } H(s) = \frac{1.324e^{005}s + 4.265e^{010}}{4904s + 5.164}$$

A sequence of similarity for symbolic quaternary N-sequences can be explored:
s = {A, C, G, T}; or s = {1, 2, 3, 4}, or s = {1+j, 1-j,-1+j,-1-j}
A pair of cross-similarity can be computed: S(A, G), S(T, C), S(A, C), or S(G,T) for natively uncomplimentary pair, otherwise compute self-similarity S(A, T), and S(G, C) for complementary pair of N sequences []:

$$S(X_i,Y_i) = \sum_{i=0}^{N} s(x_i,y_i)$$

Given: n= 16; x_i =(A A C G T G T A C C A T G C G T) ; y_i= [G C T A A G T A C T G A C C G A];

Cross-similarity: $S(X_i, Y_i)$ = 0+0 +0+0 + 0+3+4+1 + 2+0+0+0 + 0+2+3+0] = 15 for s(1,2,3,4),

Self-similarity: $S(X_i, Y_i)$ = 0+0+0+0 + 5+0+0+0 + 0+0+0+5+ 5+0+0+5] = 20 for s{1,2,3,4}.

The frequency domain analysis can further be elucidated by a numerical sequence resulting from a character string of length N above:

$$x[n] = AX_A[n] + CX_C[n] + GX_G[n] + TX_T[n], \qquad n = 0, 1, 2, \ldots N-1,$$

Where symbol or numerical assignment of A, C, G, and T can be real or imaginary as initially specify above. We can thus write discrete Fourier Transform of the x[n] sequence:

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

The Fourier transform provides a measure of the frequency content at discrete quantization of frequency k, equivalent to a period of N/k samples. Four discrete Fourier Transform sequences can be computed with presence or absent, otherwise binary indicator in the sequence:

$$Y_i = \sum_{n=0}^{N-1} X_i e^{-j(2\pi/N)kn}, \quad i = A,C,G, \text{ or } T; \quad k = 0, 1, 2, \ldots\ldots N-1$$

For S = (1, 2, 3, 4):

$$X(k) = 1Y_A(k) + 2Y_C(k) + 3Y_G(k) + 4Y_T(k), \quad k = 1, 2, 3, \ldots\ldots N-1.$$

Given a 4-dimensional frequency spectrum that sum to zero when k is nonzero but N when k=0, i.e.:

$$Y_A(k) + Y_C(k) + Y_G(k) + Y_T(k) = \begin{cases} 0, & k \text{ un= } 0; \\ N, & k = 0; \end{cases}$$

Finally, the total power spectrum (PSD) content of the DNA character strings at the discrete frequency k is given:

$$S(k) = |Y_A(k)|^2 + |Y_C(k)|^2 + |Y_G(k)|^2 + |Y_T(k)|^2$$

Fourier Transform provides computational efficiency, robust, versatile coherency in exploring huge sequence as well as short sequence of DNA nucleotides. It will offer structural and functional identification of any definitive genomic sequence. This has been explored largely on high throughput micro-array gene expression, however will be versatile for compact non-labeled DNA biosensor, bearing along its digital processing power.

Computational Tools: Matlab toolbox for 3-Component periodicity will work for oligonucleotide or not? For a specific a DNA sequence, the indicator sequence for the base *A* is a binary sequence, e.g.:,

$$X_A(n) = 000110111000101010 \ldots$$

where 1 indicates the presence of an *A* and 0 indicates its absence. The indicator sequences for the other bases are defined similarly. It is clear that the sequence 111111 . . . is obtained by adding the four indicator sequences. The DFT of a length-N block of $X_A(n)$ is defined as

$$X_A[k] = \sum_{n=0}^{N-1} x_A(n) e^{-j2\pi kn/N}, \; 0 \le k \le N-1,$$

where the assigned number n = 0 to the beginning of the block. The DFTs $X_T[k]$, $X_C[k]$, and $X_G[k]$ are defined similarly. The period-3 property of a DNA sequence implies that the DFT coefficients corresponding to $k = N/3$ are large. Thus if we take N to be a multiple of 3 and plot

$$S[k] \Delta/= |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 \quad (1)$$

Then one should see a peak at the sample value $k = N/3$ as demonstrated in many papers [5]

Result:

Figure 2 a-d show the frequency simulation hardcopy outputs based the postulated theory presented in the paper. Figure 3a and b presented the spectral analyses based on 3-component analysis of pre and post-hybridization.

Case sample in biosystem:

ACTTAGCTACAGA…

The binary indicator sequences X for each base A, T, C and G are respectively coded [6]:

$X_A[k]$ = 1000100010101…

$X_T[k]$ = 0011000100000…

$X_C[k]$ = 0100001001000…

$X_G[k]$ = 0000010000010…

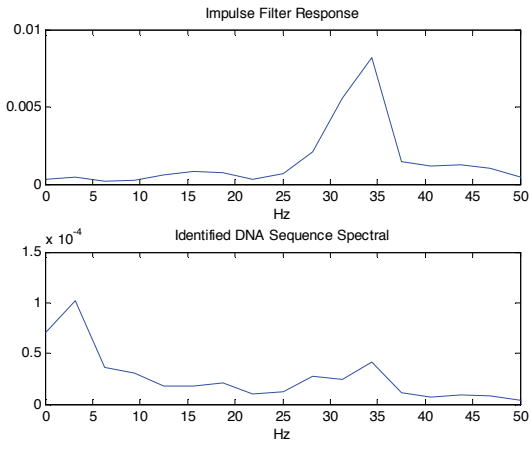

Figure 2a.

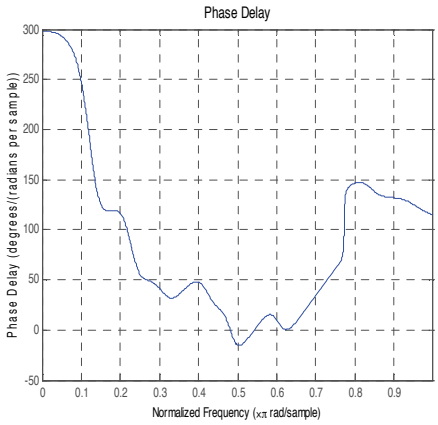

Figure 2b.

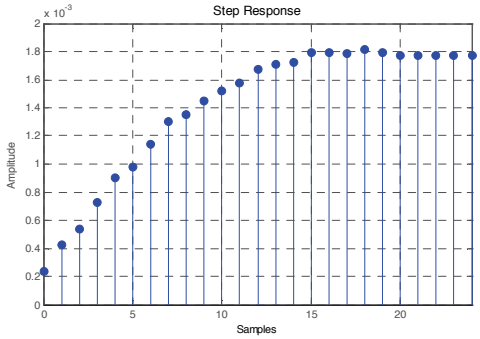

Figure 2c.

Figure 2: Frequency Analysis of DNA Hybridization

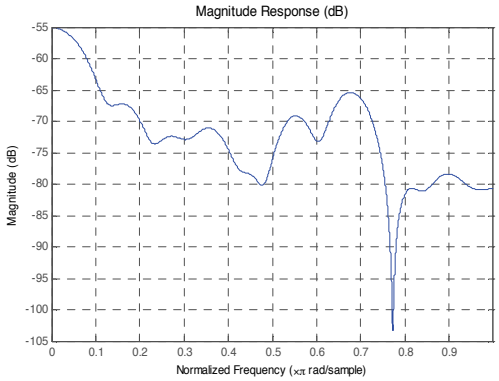

Figure 2d.

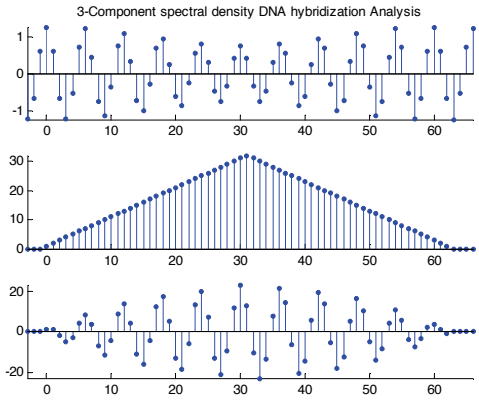

Figure 3a.

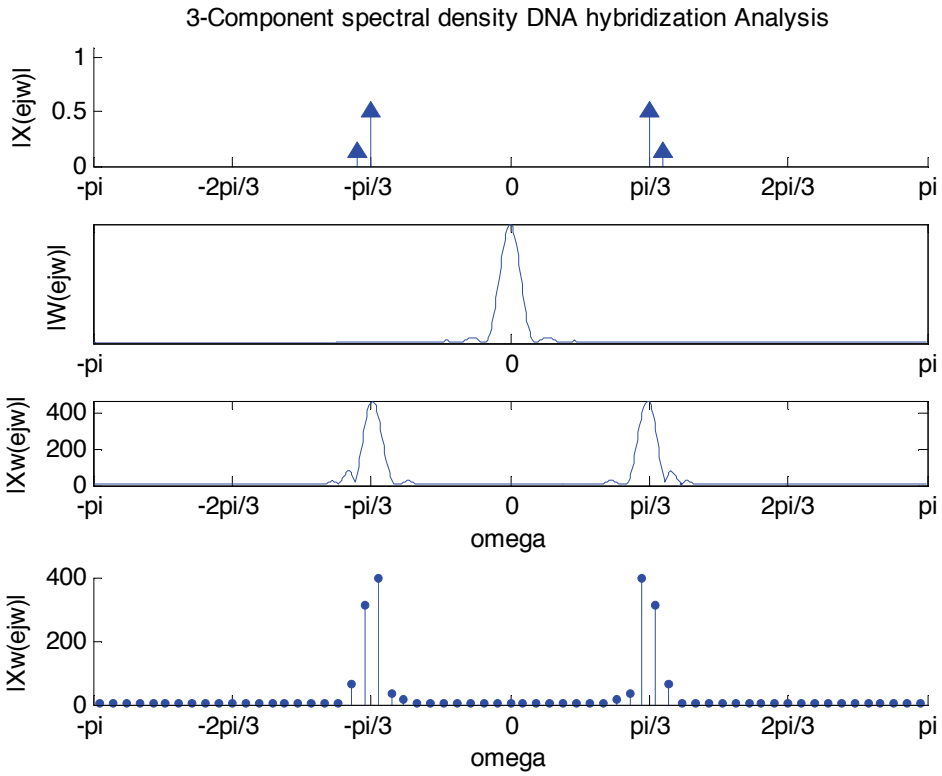

Figure 3b. Spectral Analysis Pre and Post Hybridization

Discussion and Conclusion:

There are 4^N possible sequences of length N made of 4 kinds of letters: a huge subset of typical sequences and a tiny subset of atypical sequences. Biological sequences resulted from billions years of evolution and natural selection, should have selected the atypical subset that fits individual genomic rationalized above expositions. The identification of gene prediction based on the period-3 property is based on computation of DFT of short indicator sequences of the four bases segments.

The exon of DNA active gene segment exhibits strong power spectra component at 2pi/3 period three components. The 3-component spectra power can be used to identify specific segment of the hybridized DNA. Fourier Transform provides computational efficiency, robust, versatile coherency in exploring huge sequence as well as short sequence of DNA nucleotides. It will offer structural and functional identification of any definitive genomic sequence. This has been explored largely on high throughput micro-array gene expression, however will be versatile for compact non-labeled DNA biosensor, bearing along its digital processing power. Furthermore,

given sequence of gene with discrete power spectrum distribution PSD, there exits correlation between similar sequence and cross-correlation among non-similar gene sequences as simplistically initially illustrated in this brief presentation.

Reference:

1 . Afuwape, S. A., Int. J. Nanotechnology., Vol. 5, Nos. 4/5, 2008.

2. Li, H., et al., 'Minimum Entropy Clustering and Applications to Gene Expression 'In Proceedings of IEEE Computational Systems Bioinformatics…' 2004.

3. Altschul, S. F. et al., Nuclei. Acids Res., vol. 25, pp. 3389-3402, 1997.

4. http://pnylab.com/pny/pny/papers/cdna/cdna/index.html

5. Tuqan, J. and Rushdi, A. A DSP Approach for Finding the Codon Bias in DNA Sequences Journal of Selected Topics in Signal Processing, Volume: 2, Issue: 3, Pages: 343-356, (2008).

6. Mohapatra, A. et al. 'International Journal of Computer Science and Network Security', vol.9 No.1, Jan. 2009.