I.J. Information Technology and Computer Science, **2012, 8, 22-36**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2012.08.03

Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques

Mohammed Abo-Zahhad, Sabah M. Ahmed, Shimaa A. Abd-Elrahman
Electrical and Electronics Engineering Department, Faculty of Engineering, Assiut University, Assiut, Egypt
(zahhad@yahoo.com, sabahma@yahoo.com, and shimaa.adly@gmail.com)

Abstract— Using digital signal processing in genomic field is a key of solving most problems in this area such as prediction of gene locations in a genomic sequence and identifying the defect regions in DNA sequence. It is found that, using DSP is possible only if the symbol sequences are mapped into numbers. In literature many techniques have been developed for numerical representation of DNA sequences. They can be classified into two types, Fixed Mapping (FM) and Physico Chemical Property Based Mapping (PCPBM) . The open question is that, which one of these numerical representation techniques is to be used? The answer to this question needs understanding these numerical representations considering the fact that each mapping depends on a particular application. This paper explains this answer and introduces comparison between these techniques in terms of their precision in exon and intron classification. Simulations are carried out using short sequences of the human genome (GRch37/hg19). The final results indicate that the classification performance is a function of the numerical representation method.

Index Terms—Genomic Signal Processing; DNA and Proteins Sequences; Numerical Mapping; Codon, Exons and Introns; Short Time Fourier Transform

I. Introduction

Genomic Signal Processing (GSP) is defined as the analysis, and use of genomic signals to gain biological knowledge, and the translation of that knowledge into systems-based applications. Genomic information is digital in a very real sense. It's It is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA and proteins, have been represented by character strings, in which each character is a letter of an alphabet. In case of DNA, the alphabet is of size 4 (for proteins it's 20) and consists of the letters A, T, C and G (e.g. ….ATCGCTGA ...). If numerical values are assigned to these characters, the resulting numerical sequences are readily amenable to DSP applications such as gene prediction which refers to locate the protein-coding regions (exons) of genes in a long DNA sequence [1]. Therefore, it is necessary to map the symbols into numerical sequences. An ideal mapping should be such that the period-3 component of the DNA sequence should be independent of the nucleotides mapping, which is possible only through symmetric mapping [2]-[3]. Once the mapping is done, signal processing techniques can be used to identify period-3 regions in the DNA sequence. The average length of a chromosome is of the order of millions of bases so it needs vast number of computations for identifying the protein coding regions. The computational complexity can be reduced either at mapping process or at implementation time. In recent years [4]-[5], a number of schemes have been introduced to map DNA nucleotides into numerical values. Some possible desirable properties of a DNA numerical representation include:

1. Each nucleotide has equal weight (e.g., magnitude), since there is no biological evidence to suggest that one is more important than another;

2. Distances between all pairs of nucleotides should be equal, since there is no biological evidence to suggest that any pair is closer than another;

3. Representations should be compact, in particular, redundancy should be minimized;

4. Representations should allow access to a range of mathematical analysis tools.

The paper is organized as follows. Section 1 presents this introduction. Sections 2 and 3 review the recently published mapping techniques of DNA sequences into numerical representations which are broadly classified into two major groups: fixed mapping techniques and physico-chemical property based mapping techniques and their applications. Section 4 introduces a comparison between different mapping techniques in terms of their merits and demerits. Section 5 introduces the simulation and results of using some of the existing mapping techniques in exon and intron classification. It also includes the comparison between mapping approaches relative to accuracy in exon and intron classification. Finally, section 6 concludes the paper.

II. Fixed Mapping of DNA Sequence

In FM techniques, the nucleotides of DNA data are transformed into a series of arbitrary numerical sequences [5]. These techniques are represented by two choices [6]. In the first choice four binary sequences are created, one for each character (base), which specify whether a character is present (1) or absent (0) at a specific location. These binary sequences are known as indicator sequences. The second choice is based on the geometric representations where meaningful real or complex numbers are assigned to the four characters A, T, G, and C. In this way a single numerical sequence representing the entire character string is obtained. In general FM techniques [5] include the Voss [7], the tetrahedron [2], the complex [8]-[11], the integer [9], the real [12], and the quaternion [13] - [14] representations. Each of the DNA numerical representations offers different properties as will be show in details in the following sections.

2.1 Voss Mapping Technique

One of the most popularly used mapping techniques is the Voss mapping which maps the nucleotides A, C, G, and T into four binary indicator sequences $x_A(n), x_C(n), x_G(n)$, and $x_T(n)$ [7]. Consequently it is a four dimensional mapping [3], because each base in the DNA sequence is represented by a four dimensional vector composed of either '0' or '1'. The number of 1's in any vector is exactly one. For example in the indicator sequence $x_A(n)$, '1' indicates the presence of base A and '0' indicates its absence as shown in the following example.

DNA Sequence	…T	T	G	T	C	A	C	T	C	G	G…
$x_A(n)$:	…0	0	0	0	0	1	0	0	0	0	0…
$x_C(n)$:	…0	0	0	0	1	0	1	0	1	0	0…
$x_G(n)$:	…0	0	1	0	0	0	0	0	0	1	1…
$x_T(n)$:	…1	1	0	1	0	0	0	1	0	0	0…

Voss mapping method does not predefine any mathematical relationship among the bases, but only indicates the frequencies of the bases. This method is an efficient representation among fixed mapping methods for spectral analysis of DNA sequences. The Fourier power spectrum of the Voss's binary indicator sequences, reveals a peak at frequency 1/3 (period-3) for coding regions and shows no peak for non-coding regions. The main application of Voss technique is the efficient identification of the coding and non-coding regions in a DNA sequences [15]. DFT of a sequence $x(n)$ of length L, is itself another sequence X[K], of the same length L.

$$\mathrm{X[K]} = \sum_{n=0}^{L-1} x(n)\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{\mathrm{L}}\mathrm{Kn}}, K = 0, 1, \ldots\ldots, L-1 \qquad (1)$$

The sequence X[K] provides a measure of the frequency content at frequency K, which corresponds to an underlying period of L/K samples. Using the above definition the $\mathrm{X_A[K]}$, $\mathrm{X_T[K]}$, $\mathrm{X_C[K]}$, and $\mathrm{X_G[K]}$ are the DFTs of the binary indicator sequences $x_A(n)$, $x_T(n)$, $x_C(n)$, and $x_G(n)$, respectively. As a result, the total power spectral content P[K] of the DNA character string, at frequency K is given by;

$$\mathrm{P[K]} = |\mathrm{X_A[K]}|^2 + |\mathrm{X_T[K]}|^2 + |\mathrm{X_C[K]}|^2 + |\mathrm{X_G[K]}|^2, \qquad (2)$$
$$\mathrm{K} = 0, 1, \ldots\ldots, \mathrm{L} - 1$$

The period-3 property of a DNA sequence implies that the DFT coefficients corresponding to K = L/3 is large, where L is a multiple of 3. The period-3 property is related to the different statistical distributions of codons between protein-coding and non-coding DNA sections. This property is used as a basis for identifying the coding and non-coding regions in a DNA sequence. Instead of evaluating the DFT of a full-length sequence (L), Short Time Discrete Fourier Transform (STDFT) is computed over N samples for each binary indicator to get a better time domain resolution by sliding a window by one entry in the sequence.

$$\mathrm{X_\alpha[K]} = \sum_{n=0}^{N-1} \mathrm{w(n)x_\alpha(n)e^{-j2\pi Kn/N}}$$
$$\text{for } 0 \leq \mathrm{K} \leq \mathrm{N} - 1$$

Where, $\alpha =$A, T, C, G and $\mathrm{w(n)}$ is a rectangular window given by;

$$\mathrm{w(n)} = \begin{cases} 1 & \text{for } 0 \leq \mathrm{n} \leq \mathrm{N} - 1 \\ 0 & \text{elsewhere} \end{cases} \qquad (4)$$

Fig 1 shows the result of applying STDFT on nucleotide sequence of the gene F56F11.5 of C. elegans [GenBank website [16], Accession number: AF099922] over the bases 7021 to 15120 by using rectangular window with length 351. It is known that this sequence has five known coding regions, but the use of Voss mapping technique shows four discernible peaks for coding regions 2, 3, 4, 5 and the first coding region is unrecognizable.

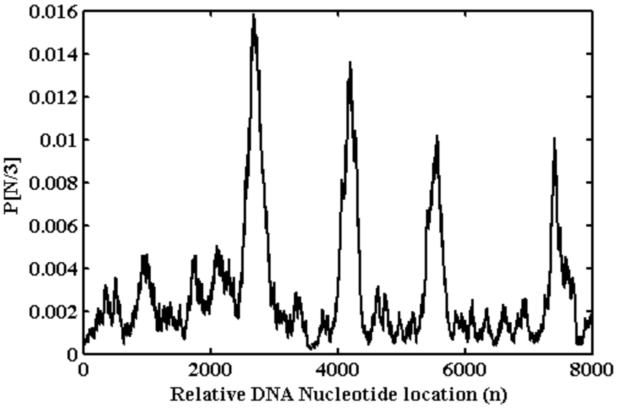


Fig 1: The power spectrum P[N/3] vs. the base location n for the gene F56F11.4 in the C.elegans chromosome III using Voss mapping.

2.2 Tetrahedron Representation

This representation [2] reduces the number of indicator sequences from four to three but in a manner symmetric to all the four sequences, where the four sequences $x_A(n)$, $x_C(n)$, $x_G(n)$, and $x_T(n)$ are mapped to the four 3- dimensional vectors pointing from the center to the vertices of a regular tetrahedron as shown in Fig 2. Voss and the tetrahedron mapping methods are two equivalent representations when being used in power spectrum analysis [4]-[6]. The resolving of the four three-dimensional vectors, results the following:

$$(a_r, a_g, a_b) = (0, 0, 1) \tag{5}$$

$$(t_r, t_g, t_b) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right) \tag{6}$$

$$(g_r, g_g, g_b) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right) \tag{7}$$

$$(c_r, c_g, c_b) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right) \tag{8}$$

which give rise to the following three numerical sequences

$$x_r(n) = \frac{\sqrt{2}}{3}\left(2x_T(n) - x_C(n) - x_G(n)\right) \tag{9}$$

$$x_g(n) = \frac{\sqrt{6}}{3}\left(x_C(n) - x_G(n)\right) \tag{10}$$

$$x_b(n) = \frac{1}{3}\left(3x_A(n) - x_T(n) - x_C(n) - x_G(n)\right) \tag{11}$$

Where r, g, b are red, green, and blue indicators, respectively.

Obtaining DNA spectrograms of biomolecular sequences is the main application of tetrahedron representation. These simultaneously provide local frequency information for all four bases by displaying the resulting three magnitudes by superposition of the corresponding three primary colors, red for $|X[K]|_r$, green for $|X[K]|_g$, and blue for $|X[K]|_b$. Thus color conveys real information, as opposed to pseudo color spectrograms, in which color is used for contrast enhancement. For example, Fig 3 shows a spectrogram using DFTs[8] of length 60 of a DNA stretch of 4,000 nucleotides from chromosome III of C. elegans (GenBank[16], Accession number: NC000967). The vertical axis corresponds to the frequencies K from 1 to 30, while the horizontal axis shows the relative nucleotide locations starting from nucleotide 858,001; only frequencies up to K =30 are shown due to conjugate symmetry as x_r, x_g and x_b are real sequences. The resulting DNA color spectrogram can be used to locate repeating DNA sections as shown in Fig 3.

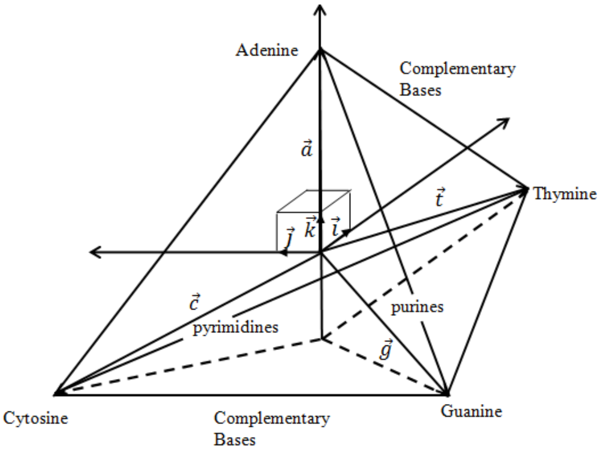


Fig 2: The Tetrahedron representation

Spectrograms also can be used to locate CG rich regions in DNA called CpG islands [6]. The `p' in CpG simply denotes that C and G are linked by a phosphordiester bond as shown in Fig 4.

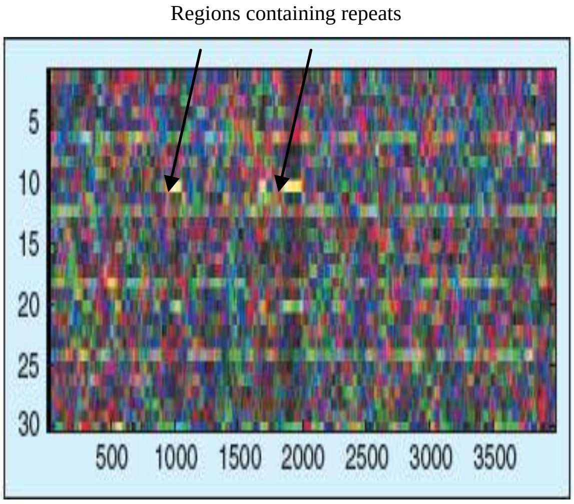


Fig 3: Real DNA section

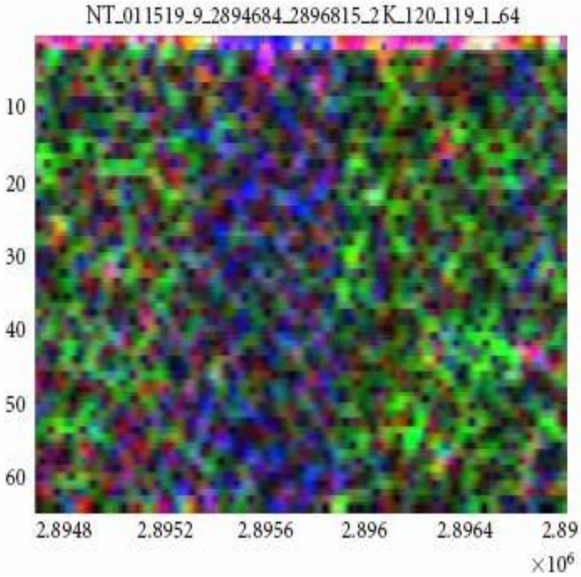


Fig 4: CpG islands (green) separated by regions rich in A (blue)

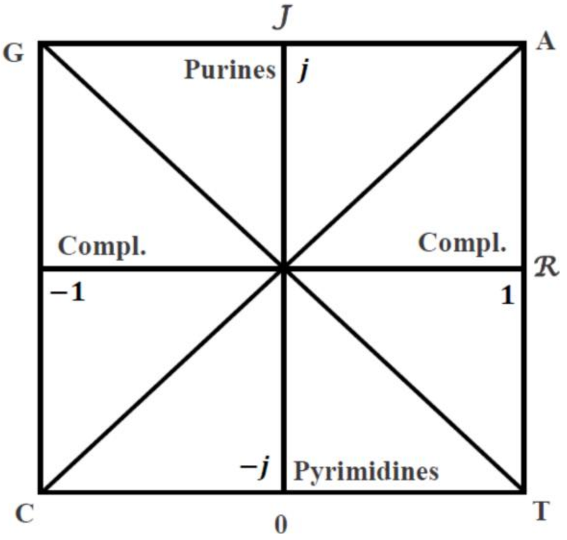
2.3 Complex Representation

The dimensionality of the tetrahedral representation can be reduced to two by projecting the basic tetrahedron on different planes [8]-[11]. Such planes can be chosen in various ways that conserve the symmetry of the representation and reflect biological properties in corresponding mathematical properties. For instance, the planes can be defined by a pair of the coordinate axes. On the other hand, these planes can be put in correspondence with a complex plane, so that a complex representation of the bases is obtained. For example, the four bases are placed in a quadrantal symmetry as shown in Fig 5-a and the complex representation of the bases is given by:

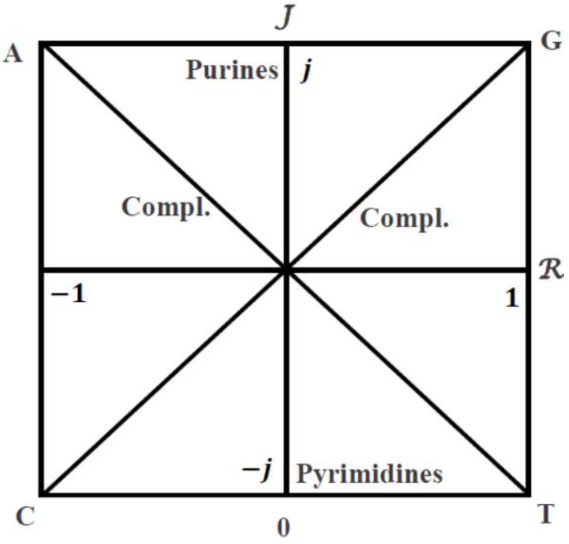
$$x(n)=ax_A(n)+cx_C(n)+tx_T(n)+gx_G(n) \tag{12}$$

Where, $a = 1 + j$, $t = 1 - j$, $c = -1 - j$, and $g = -1 + j$ are complex numbers. This representation would be more advantageous from signal processing perspective depending upon the chosen values of the variables (a, t, c, g) [1].

The complex representation has the advantage of better translating some of the features of the bases into mathematical properties. For instance, in the representation of Fig 5-a, the complementarily of the pairs of bases A-T and G-C, respectively, is expressed by the fact that their representations are complex conjugates, while purines and pyrimidines have the equal imaginary parts and real parts of opposite sign. Permuting the bases A-G, i.e., choosing the left projection plane, the representation shown in Fig 5-b is obtained, for which $a = -1 + j$, $t = 1 - j$, $c = -1 - j$, and $g = 1 + j$. This representation has the advantage that the two complementary strands of a DNA molecule correspond to digital signals of equal absolute values, but opposite signs, so that their sum is always zero.



(a)



(b)

Fig 5: The projections of the tetrahedral representation of the nucleotides on the (a) Right plane, (b) Left plane.

The main application of the complex representation is the detection of exon regions [17], where the P-3 property can be identified by calculating the STDFT $X_\alpha[K]$ for each binary indicator $x_\alpha(n)$; (α =A, T, C, G) over N samples given by equations (3) and (4). The power spectrum of the DNA sequence P[K] is calculated by:

$$P[K]=\left|\frac{1}{N}\ [aX_A[K]+tX_T[K]+cX_C[K]+gX_G[K]]\right|^2 \tag{13}$$

Where, a, t, c, and g are complex mapping constants given respectively by [17];

$$a = 0.10 + 0.12j, \tag{14}$$
$$t = -0.30 - 0.20j, \tag{15}$$
$$c = 0, \text{and} \tag{16}$$
$$g = 0.45 - 0.19j \tag{17}$$

The above constants are calculated such that discriminatory capability between protein coding regions (with corresponding random variables A, T, C, and G) and random DNA regions is maximized [8]. If P-3 property is present, the STDFT coefficient $X_\alpha[N/3]$ is significantly larger than the surrounding STDFT coefficient. Consequently, $P[N/3]$ is large in a coding region. As an illustrative example, the STDFT method was used to analyze coding regions in the gene F56F11.4 in C.elegans chromosome III over the bases 7021 to 15120, using rectangular window with length 351. This 8100-length DNA sequence, which obtained using the Nucleotide Accession Number ‘AF099922’ from Genbank website [16], yields to the power spectrum $P[N/3]$ shown in Fig 6. From this figure, all the five known coding regions are recognizable. Thus, the complex mapping has a good feature in gene prediction.

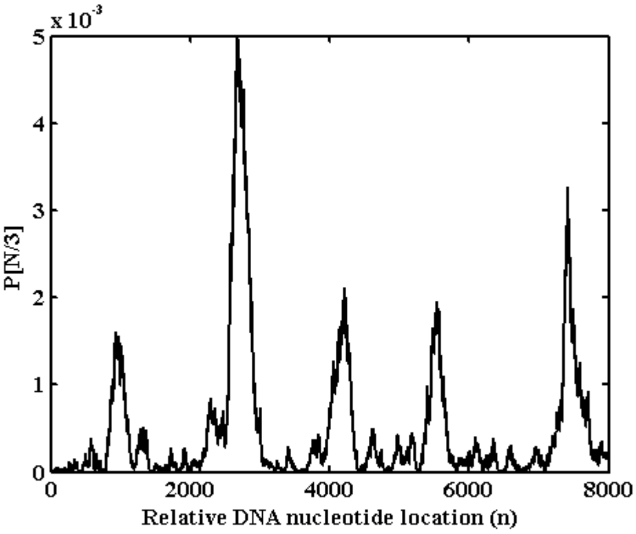


Fig 6: The power spectrum P[N/3] vs. the base location n for the gene F56F11.4 in the C.elegans chromosome III using complex mapping.

2.4 Integer Representation

The integer representation is a one-dimensional (1-D) mapping of DNA bases [5], [9]. This mapping can be obtained by mapping numerals {0, 1, 2, 3} to the four nucleotides as: T=0, C=1, A=2, and G=3. However, this method implies a structure on the nucleotides such as purine (A, G) > pyrimidine (C, T). Similarly [4], the representation A=0, C=1, T=2 and, G=3 suggests that T>A and G>C. Arbitrarily assigned integer representation may introduce some mathematical property which does not exist in a base sequence. For measuring the accuracy of this mapping technique in gene prediction, the STDFT method was used to analyze coding regions in the gene F56F11.4 in C.elegans chromosome III over the bases 7021 to 15120, using rectangular window with length 351. The power spectrum P[N/3] of 8100-length DNA sequence (Nucleotide Accession Number 'AF099922' from Genbank website [16]), using first possible representation (T=0, C=1, A=2, and G=3) and second possible representation (A=0, C=1, T=2 and, G=3) are shown in Fig 7 and Fig 8 respectively.

From these figures, the first possible mapping is not accurate in prediction of the first and forth exons, but it gives approximately the same results with the complex mapping in prediction of the second, third and fifth exons. This mapping gives better results than the second possible mapping in gene prediction. Fig 8 shows that, the second possible mapping is not accurate in gene prediction. Hence the DSP applications of integer mapping are limited suggesting that these integer mappings need to be used carefully for a given application [5].

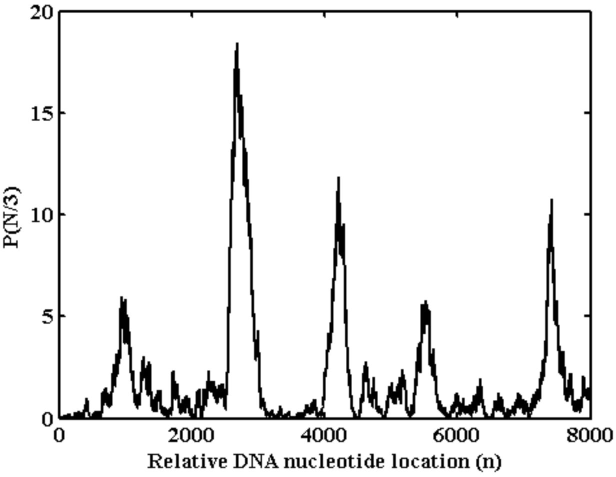


Fig 7: The power spectrum P[N/3] vs. the base location n for the gene F56F11.4 in the C.elegans chromosome III using integer mapping (T=0, C=1, A=2, and G=3).

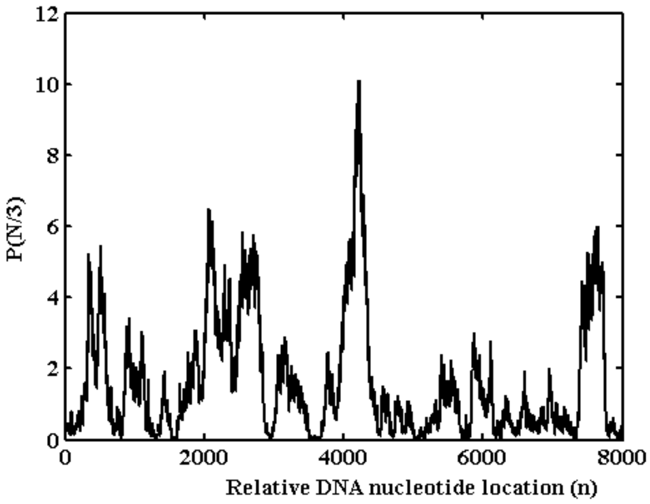


Fig 8: The power spectrum P[N/3] vs. the base location n for the gene F56F11.4 in the C.elegans chromosome III using integer mapping (T=2, C=1, A=0, and G=3).

2.5 Real Number Representation

The real number representation A = -1.5, T = 1.5, C =0.5, and G = -0.5, which bears complementary property, is efficient in finding the complimentary strand of a DNA sequence [5], [12]. However, the assignment of a real number to each of the four bases does not necessarily reflect the structure present in a DNA sequence. For example in the sequence CTGAA represented by 0.5; 1.5; −0.5; −1.5; −1.5., the change sign and reverse of the representation 1.5; 1.5; 0.5; −1.5; −0.5 yields to the sequence TTCAG. In the computation of correlations, real representations are preferred over complex representations. Furthermore, it is interesting to note that the complex, real, and integer representations can also be viewed as constellation diagrams, which are widely used in digital communications.

In the aforementioned analysis, the mapping rule used played an important role in identifying matches [12]. The real and integer-number mapping rules yielded different string matches. This is due to the inherent complementary property of the real mapping rule and the non-complementary property of the integer mapping rule. For example, the occurrences of the template 5”-TACGTGC-3” need to be found in a long DNA string. The corresponding numerical sequence obtained through real mapping would be 5”-1.5,−1.5, 0.5,−0.5, 1.5,−0.5, 0.5-3”. The following numerical sequences will have the same autoregressive (AR) parameters as the above template:

(i) 5”- −1.5, 1.5,−0.5, 0.5,−1.5, 0.5,−0.5-3” = 5”-ATGCACG-3”: (reversed complement of the template);
(ii) 5”-0.5,−0.5, 1.5,−0.5, 0.5,−1.5, 1.5-3” = 5”-CGTGCAT-3”: (reversed template);
(iii) 5”- −0.5, 0.5,−1.5, 0.5,−0.5, 1.5,−1.5-3” = 5”-GCACGTA-3”: (complement of the template).

This is due to the fact that: (a) the sign-reversed numerical sequence and the actual numerical sequence have the same linear dependence and hence the same AR parameters, and (b) minimizing the forward or the backward linear prediction error would theoretically yield the same AR model. Table 1 shows the detection of repeats of DNA segments via AR modeling. Real mapping rule and second-order AR model features are used; the template is 8 bp long. There are 5 repeats in the whole sequence. Identification of complementary and reversed sequences is obtained as well.

Table 1 Detection of repeats of DNA segments via AR modeling

Position with the same features	DNA segment
210–217 (template)	CTCACATT
5174–5181	CTCACATT
12572–12579	CTCACATT
19278–19285	AATGTGAG
29624–29631	CTCACATT
36387–36394	AATGTGAG
55805–55812	AATGTGAG
63106–63113	CTCACATT

2.6 Quaternion Representation

In the quaternion representation [13] of DNA bases, pure quaternions are assigned to each base through geometric representations [6]. One possibility is to assign to the 4 bases the 4 vectors from the center to the vertices of a regular tetrahedron as show in Fig 9, where the vertices of a regular tetrahedron form a subset of the vertices of a cube. Hence, the 4 vectors point towards alternate cube vertices. Assuming that the cube side length is 2 units and the origin is located at the cube center; the co-ordinates of the vertices of the cube are $(\pm 1,\ \pm 1,\ \pm 1)$. Consequently, the vectors a, c, g, and t are described by $a = i + j + k$, $c = i - j - k$, $g = -i - j + k$, and $t = -i + j - k$. It has been conjectured that the quaternion approach can improve DNA pattern detection in the spectral domain through the use of the quaternionic Fourier transform [14] .

This method of mapping can be used to compute the periodicity transform of DNA sequences [14]. It has been often observed that the occurrence of repetitive structures (or tandem repeats) in genomic data is symptomatic of biological phenomena. Perhaps the best known example of this association is the 3-base repetition of codons, which is characteristic of protein coding regions in DNA sequences of eukaryotic cells. The 3-base repeat is considered large-scale, as it occurs throughout the genome, in contrast to small-scale repeats, typically restricted to individual genes or gene subsets. Small-scale genomic repeats and repeat changes in the human genome have been associated, among others, with genetic diseases. Applications of repetitive structures include prediction of gene and exon locations and identification of diseases. The quaternionic approach can be utilized (via the quaternionic Fourier transform) to improve the spectral methods.

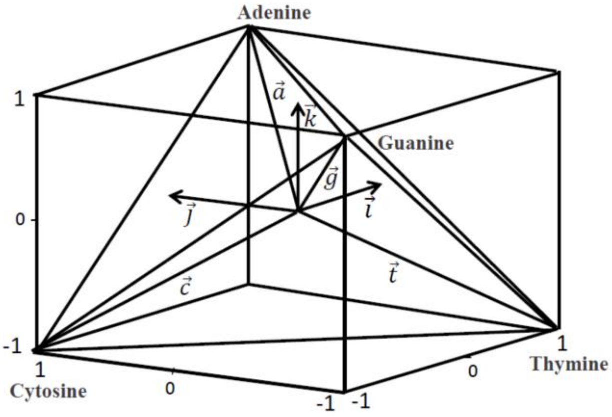


Fig 9: The quaternion representation

III. Physico Chemical Property Based Mapping

In physic chemical type of mapping, biophysical and biochemical properties of DNA biomolecules are used for DNA sequence mapping, which is robust and used to search for biological principles and structures in biomolecules [5]. In the following, the PCPBM techniques include, the Electron-Ion Interaction Potential (EIIP) [18]-[19], the atomic number [20], the paired numeric [21]-[22], the DNA walk [23] - [24], and the Z-curve representations [25]-[26], are reviewed.

3.1 EIIP Representation

In the EIIP representation the quasi-valence number associated with each nucleotide is used to map DNA character strings into numerical sequences [18]-[19]. It is just one example of a real number representation, which maps the distribution of the free electrons' energies along a DNA sequence. A single EIIP indicator sequence is formed by substituting the EIIP of the nucleotides A=0.1260, C=0.1340, G=0.0806, and T=0.1335 in a DNA sequence. If we substitute the EIIP values for A, G, C, and T in a DNA string $x(n)$, we get a numerical sequence which represents the distribution of the free electrons' energies along the DNA sequence.

This sequence is named as the EIIP indicator sequence, $x_e(n)$. For example, if $x(n) = [A\ A\ T\ G\ C\ A\ T\ C\ A]$, then using the values for each nucleotide, $x_e(n) = [0.1260\ 0.1260\ 0.1335\ 0.0806\ 0.1340\ 0.1260\ 0.1335\ 0.1340\ 0.1260]$.

This method may be used as a coding measure to detect probable coding regions in DNA sequences [18]. Let $X_e[K]$ be the STDFT of the sequence $x_e(n)$ which is given by;

$$X_e[K]=\sum_{n=0}^{N-1} w(n)x_e(n)e^{\left(\frac{j2\pi Kn}{N}\right)}, K=0,1,\ldots..N-1 \quad (18)$$

Where w(n) is a rectangular window described by equation (4). The corresponding absolute value of the power spectrum is calculated as;

$$S_e[K]=|X_e[K]|^2 \quad (19)$$

When $S_e[K]$ is plotted against K, [15] it reveals a peak at N/3 for a coding region and no such peak is observable for a noncoding region. In rectangular windows with length 240 were used for evaluating the STDFT of the gene F56F11.4 in C.elegans chromosome III over the bases 7021 to 15120 as show in Fig 10. This 8100-length DNA sequence, which obtained using the Nucleotide Accession Number ‘AF099922’ at Genbank website [16], contains five known coding regions.

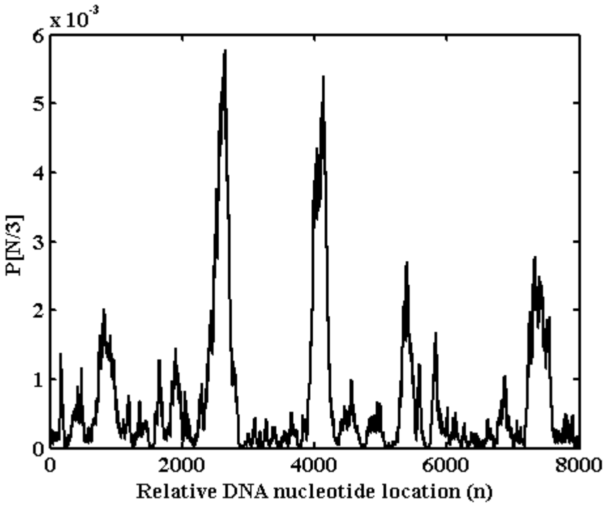


Fig 10: The power spectrum P[N/3] vs. the base location n for the gene F56F11.4 in the C.elegans chromosome III using EIIP mapping

3.2 Single Atomic Number Representation

A single atomic number indicator sequence is formed by assigning the atomic number to each nucleotide as: A=70, C=58, G=78 and T=66 in a DNA sequence [20].

3.3 Paired Numeric Representation

In the paired numeric representation nucleotides (A-T, C-G) are to be paired in a complementary manner and values of +1 and -1 are to be used respectively to denote A-T and C-G nucleotide pairs. It can be represented as one or two indicator sequences. This representation incorporates DNA structural property with reduced complexity, joining opposite strands of double helix DNA through hydrogen bonds [4] - [5]. However, this is not the reason for their pairing here, as only one strand is used for the computational analysis of DNA. This representation incorporates a very useful DNA structural property, in addition to reducing complexity. In general, a nucleotide sequence { n_i }, where i =(1, 2, ………, L) of length L is comprised of base pairs A, C, T, and G [21]. In order to apply numerical methods to a nucleotide sequence, there were seven rules for mapping the nucleotide sequence onto a one dimensional numerical sequence. They can be summarized in the following.

1. Purine-pyrimidine (RY) rule. If n_i is a purine (A or G) then u_i = 1; otherwise if n_i is pyrimidine (C or T) then u_i= -1.
2. $A\bar{A}$ rule. If n_i=A then u_i=1 ; in all other cases u_i=-1.
3. $T\bar{T}$ rule. If n_i=T then u_i=1 ; in all other cases u_i=-1.
4. $G\bar{G}$ rule. If n_i=G then u_i=1 ; in all other cases u_i=-1.
5. $C\bar{C}$ rule. If n_i=C then u_i=1 ; in all other cases u_i=-1
6. Hydrogen bond energy rule (called the SW rule). u_i= 1 for “ strongly bonded” pairs (G or C); u_i= -1 for “ weakly bonded “ pairs (A or T).
7. Hybrid rule (called the KM rule). u_i =1 for A or C ; u_i= -1 for T or G.

The RY rule has been perhaps the most widely used rule, but the other rules have also been applied. Moreover there were also other additional mapping rules (e.g., each base pair can be weighted by any characteristic of those base pairs), so u_i can be any number such as molecular mass, hydrophobicity, ect.

The main application of the paired numeric mapping is the gene and exon prediction [21]. The paired numeric representation for gene and exon prediction exploits one of the differential properties of exons and introns, according to which introns are rich in nucleotides “A” and “T” whereas exons are rich in nucleotides “C” and “G”. Fig 11 indicates that the genomic DNA sequences are first converted into single numerical sequences using the paired-numeric representation. This representation is based on complementary statistics of the occurrence of DNA nucleotides in exons and introns. In order to exploit this property, the nucleotides are paired (i.e., A-T, C-G). The values of +1 and –1 are assigned to show the presence and absence of A-T and C-G nucleotides respectively. The resultant single or two sequence DNA representations offer reductions in the cost of DFT processing compared with the four sequence binary representation. According to the paired-numeric representation, for a DNA sequence ‘ATGCTATT…’, the single sequence representation would look like x(n) = {+1, +1, –1, –1, +1, +1, +1, +1, ….}, where n represents the base index.

This representation incorporates a useful DNA structural property in subsequent processing. In a sliding window DFT, we normally calculate the DFT for a window centred upon a single base index (i.e., K = N/3 where N is the window size), which suggests that different DFT results will be obtained at a particular location of DNA sequence when moving the window in the forward and reverse directions for the same sequence. The following expression gives the Paired Spectral Content (PSC) [22].

$$PSC[K]=|X_F[K]|^2+|X_R[K]|^2 \quad (20)$$

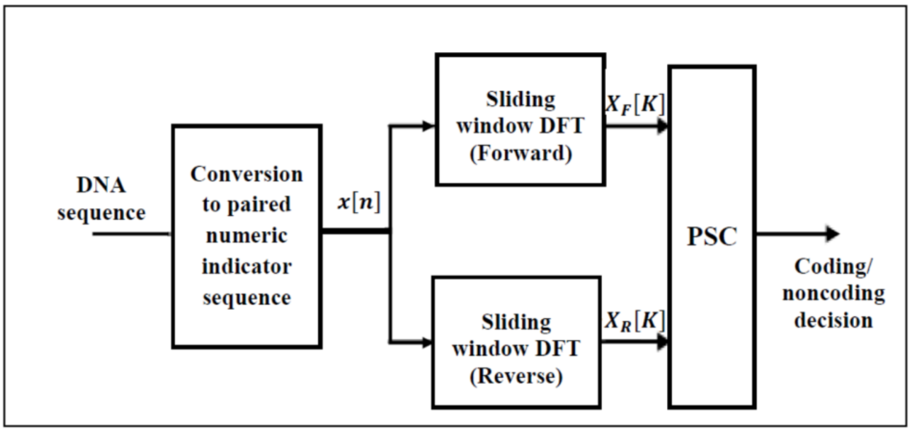

DNA sequence
Conversion to paired numeric indicator sequence
$x[n]$
Sliding window DFT (Forward)
Sliding window DFT (Reverse)
$X_F[K]$
$X_R[K]$
PSC
Coding/ noncoding decision

Fig 11: The diagram for the paired spectral content measure

Where, $X_F[K]$ and $X_R[K]$ are STDFTs for the indicator sequence x(n) in the forward and reverse directions respectively. Note that due to paired indicators, a DFT in the reverse direction of the same DNA strand is equivalent to a DFT on its complementary strand. Fig 12 shows five coding regions by using STFT of the gene F56F11.4 in C.elegans chromosome III from GenBank website [16] over the bases 7021 to 15120 when paired numerical mapping is adopted.

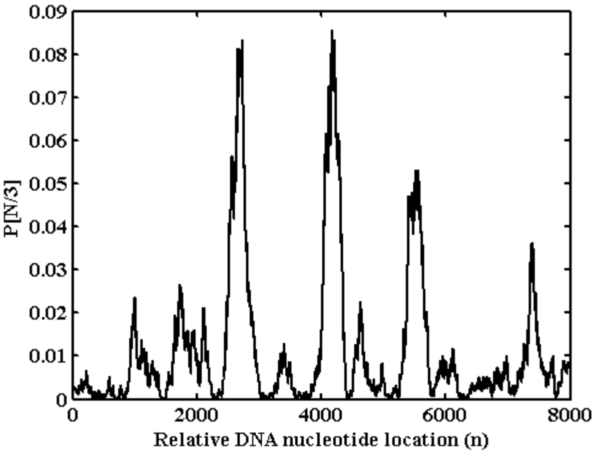


Fig 12: The power spectrum content P(N/3) vs. the base location n for the gene F56F11.4 in the C.elegans chromosome III using paired mapping

3.4 DNA-Walk Representation

In order to study the scale-invariant long-range correlations of a DNA sequence [23][24], a graphical representation of DNA sequences, which called fractal land-scape or DNA walk has been adopted. For the conventional one-dimensional random walk model, a walker moves either up $[u(i) = +1]$ or down $[u(i) = -1]$ one unit length for each step i of the walk. For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated random walk, the direction of each step depends on the history (memory) of the walker. One definition of the DNA walk is that the walker steps up $= +1$ if a pyrimidine (C or T) occurs at position i along the DNA chain, while the walker steps down$= -1$ if a purine (A or G) occurs at position (i). The graph continues to move upwards and downwards as the sequence progresses in a cumulative manner [5], with its base number represented along the x-axis .

The DNA walk allows one to visualize directly the fluctuations of the purine - pyrimidine content in DNA sequences. The positive slopes correspond to high concentration of pyrimidines, while the negative slopes correspond to high concentration of purines. Visual observation of DNA walks suggests that the coding sequences and intron-containing noncoding sequences have quite different landscapes. It can be used as a tool to visualize changes in nucleotide composition [24], as shown in

Fig 13,where it provides a graphical representation for each gene and permits the degree of correlation in the base pair (GC versus AT) sequence to be directly visualized. In the case of one dimensional walks representation a DNA sequence of length L is considered as follows:

$$X=\{x(i) \quad , \qquad i=1, 2, \ldots\ldots\ldots\ldots, L\ \} \qquad (21)$$

For a position i in the sequence, x(i) = -1 if a pyrimidine is present and x(i) = 1 if a purine is present. The corresponding DNA walk of the sequence is defined as:

$$S = \{ s[i],\ i = 1, 2, \ldots\ldots, L\} \qquad (22)$$

where,
$$s[i] = \sum_{K=0}^{i} x(K) \qquad (23)$$

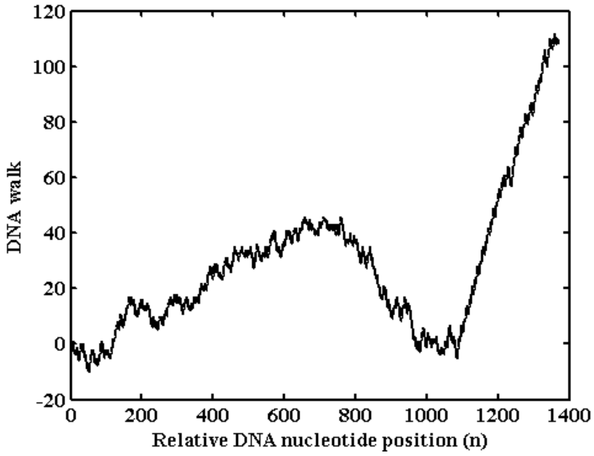


Fig 13: One-dimensional DNA walk. This illustrates the relative content of purine and pyrimidine residues within the non coding region of Helicobacter pylori strain J99 bacteria .Accession [AE001439] (4066 to 5435)

Locating periodicities or nucleotide structures may be found using the complex walk representation, which is shown in Fig 14, where S defined by equation (22), (23) with the values of x for apposition i in the DNA sequence defined as:

$$x(i) = 1 \quad \text{if } A \qquad (24)$$

$$x(i) = -1 \quad \text{if } G \qquad (25)$$

$$x(i) = j \quad \text{if } T \qquad (26)$$

$$x(i) = -j \quad \text{if } C \qquad (27)$$

For instance, in both

Fig 13 and Fig 14, a staircase-like behavior of the walk sequences occurring near (bp ≈ 5150 …5290). The repeat sequence is 5’-AAAGAATTTAAAAAC-3’ and it occurs eight times over the region.

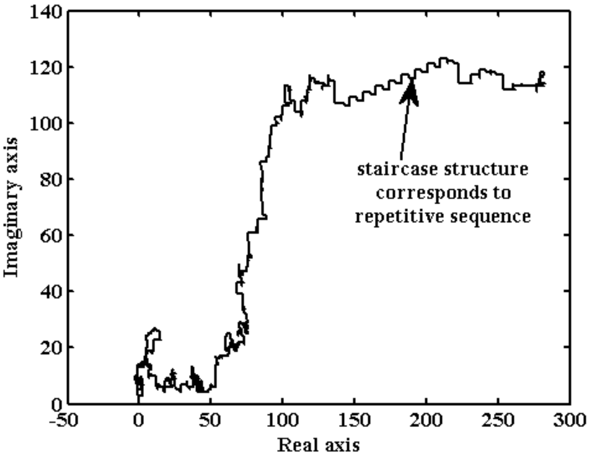


Fig 14: Two-dimensional DNA walk projection. This illustrates the repetitive sequence structure in the noncoding region of Helicobacter pylori strain J99 bacteria Accession [AE001439] (4066 to 5435)

3.5 Z-Curve Mapping

The Z-curve is a three-dimensional curve that provides a unique representation of a DNA sequence in that the DNA sequence and the Z-curve can each be uniquely reconstructed from the other [25]. Therefore, the Z-curve contains all the information that the corresponding DNA sequence carries. The resulting curve has a zigzag shape, hence the name Z-curve. A DNA sequence can be analyzed by studying the corresponding Z-curve. One of the advantages of the Z-curve is its intuitiveness; the entire Z-curve of a genome can be viewed on a computer screen or on paper, regardless of genome length, thus allowing both global and local compositional features of genomes to be easily grasped. By combining use of the Z-curve with statistical analysis, better results may be obtained. The Z-curve is composed of a series of nodes, P_0, P_1, P_2, …., P_L with coordinates x_n, y_n, z_n, where n = 0, 1, 2, …, L, and L is the length of the DNA sequence.

$$x_n = (A_n+G_n) - (C_n+ T_n) \equiv R_n\text{-}Y_n \qquad (28)$$

$$y_n = (A_n+C_n) - (G_n+T_n) \equiv M_n\text{-}K_n \qquad (29)$$

$$z_n = (A_n+T_n) - (C_n+G_n) \equiv W_n\text{-}S_n \qquad (30)$$

Here, A_n , C_n , G_n and T_n are the cumulative occurrence numbers of A, C, G and T, respectively, in the subsequence from the first base to the nth base in the sequence. Consider that $A_0 = C_0 = G_0 = T_0 = 0$, and therefore, $x_o = y_0 = z_0 = 0$. Here R, Y, M, K,W and S represent the purine, pyrimidine, amino, keto, weak hydrogen (H) bond and strong H bond bases, respectively. The three components of the Z-curve x_n, y_n and z_n, represent three independent distributions that completely describe the DNA sequence being studied. The components x_n, y_n and z_n, display the distributions of purine versus pyrimidine (R vs.Y), amino versus keto (M vs. K) and strong H-bond versus weak H-bond (S vs. W) bases along the sequence, respectively. In the subsequence constituted from the first base to the nth base of the sequence, when purine bases (A and G) are in excess of pyrimidine bases (C and T), $x_n > 0$, otherwise, $x_n < 0$, and when the numbers of purine and pyrimidine bases are identical, $x_n = 0$. Similarly, when amino bases (A and C) are in excess of keto bases (G and T), $y_n > 0$, otherwise, $y_n < 0$, and when the numbers of amino and keto bases are identical, $y_n = 0$. Finally, when weak H-bond bases (A and T) are in excess of strong H-bond bases (G and C), $z_n > 0$, otherwise, $z_n < 0$, and when the numbers of weak and strong H-bond bases are identical, $z_n = 0$. Fig 15 shows an example of the Z-curves for the M. mazei genome. The Z-curve for a genome is a three-dimensional (3-D) curve. Arrow indicates the position of cdc6 genes, and also the putative replication origin.

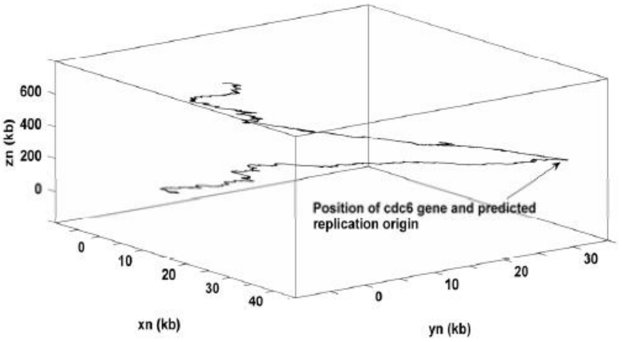


Fig 15: The Z-curve for M.mazei genome

To clarify the applications of Z-curve mapping, it should be noted that the periodicity assumption is different between the FT approach and the Z-curve approach [26]. In the Z-curve approach, the periodicity assumption applies with regards to the biological properties and the nucleotide positions induced by the different base combination. In contrast, the periodicity assumption in the FT approach is made regardless of the biological properties. It simply sums up the spectra of different nucleotide indicator sequences independently. For demonstration, let's consider an artificial sequence {T, A, G, C, G, A}. In the FT approach, this gives rise to four binary indicator sequences, {0, 1, 0, 0, 0, 1} (A), {0, 0, 1, 0, 1, 0} (G), {1, 0, 0, 0, 0, 0} (T) and {0, 0, 0, 1, 0, 0} (C). Periodicity cannot be observed in any sequence. In the Z-curve approach, the modified sequence $S_0[n]$ is {-1, 1, 1, -1, 1, 1}, which shows strong 3-periodicity. Finally, three modified sequences which characterize different biological properties are considered in the Z-curve approach whereas only the original sequence is considered in the FT approach. Hence, the FT approach considers only one spike at $2\pi/3$ for classification whereas the Z-curve approach considers both the DC value and the value at $2\pi/3$ of three modified sequences. In view of the above analysis, STDFT can be applied to the modified sequences $S_0[n]$, $S_1[n]$and $S_2[n]$. Then the power spectra of the modified sequences are first formed. The three DC values and the three values at $2\pi/3$ can then be used for sequence classification. They carry similar biological interpretation as the Z-curve features.

IV. Comparison Between Mapping Approaches

Table 2 summarizes the DNA numerical representations investigated in sections (2) and (3). One example and the dimension for each mapping method are given in the third and fourth columns respectively. Table 3 summarizes the merits and demerits of these representations.

Some of the above mentioned methods are effective in gene prediction and the other methods have their own applications. Table 4 shows the comparison between the most promising gene prediction methods; namely Voss, complex and EIIP mapping methods relative to measure accuracy in detecting the five coding regions for Gene F56F11.4. It indicates that the region of exons and length of each exon detected by each method compared with that obtained by National Center for Biotechnology Information (NCBI) [16].

Table 2: DNA Numerical Representations Methods.

Method	**Representations**	**Example : S(n)= [CGAT]**	**Dimension**
Voss [7]	X_n= 1 for $S[n] = X$; X_n= 0 for $S[n] \neq X$; X_n applies to any C_n, G_n, A_n, T_n	C_n= [1, 0, 0, 0]; G_n = [0, 1, 0, 0]; A_n= [0, 0, 1, 0]; T_n = [0, 0, 0, 1].	4
Tetrahedron [2]	$x_r(n) = \frac{\sqrt{2}}{3}(2T_n - C_n - G_n)$ $x_g(n) = \frac{\sqrt{6}}{3}(C_n - G_n)$ $x_b(n)= \frac{1}{3}(3A_n - T_n - C_n - G_n)$	$x_r(n)= \frac{\sqrt{2}}{3}[-1, -1, 0, 2]$ $x_g(n) = \frac{\sqrt{6}}{3}[1, -1, 0, 0]$ $x_b(n)= \frac{1}{3}[-1, -1, 3, -1]$	3
Integer [9]	$A = 2, C = 1, G = 3, T = 0$	[1, 3, 2, 0]	1
Real [12]	$A = -1.5, C = 0.5, G = -0.5, T = 1.5$	[0.5, −0.5, −1.5, 1.5]	1
Complex [8] - [11]	$A = 1 + j,\ C = -1 + j,$ $G = -1 - j,\ T = 1 - j$	$[-1 + j,\ -1 - j, 1 + j,\ 1 - j]$	1, 4
Quaternion [13] - [14]	$A = i + j + k,\ C = i - j - k,$ $G = -i - j + k,\ T = -i + j - k$	$[i - j - k,\ - i - j + k,$ $i + j + k,\ -i + j - k]$	1, 4

EIIP [18] - [19]	A=0.1260, C=0.1340,
	G=0.0806, T=0.1335
		[0.1340, 0.0806,0.1260, 0.1335]	1, 4

Atomic number [20]	A=70, C=58, G=78, T=66	[58, 78, 70, 66]	1, 4
Paired Numeric [21] - [22]	$A\ or\ T = 1, C\ or\ G = -1$	$P_{1n}=[-1,\ -1,\ 1,\ 1]$	1
		$P_{2n} = [-1, -1, 0, 0]$ & $[0, 0, 1, 1]$	2
DNA walk [23] - [24]	$C\ or\ T = 1, A\ or\ G = -1$	$[1, 0, -1, 0]$	1
Z-curve [25] - [26]	$x_n = (A_n+G_n) - (C_n+ T_n))\equiv R_n-Y_n$ $y_n= (A_n+C_n) - (G_n+T_n)\equiv M_n-K_n$ $z_n= (A_n+T_n) - (C_n+G_n)\equiv W_n-S_n$	x_n = [-1, 0, 1, 0] y_n= [1, 0, 1, 0] z_n= [-1, -2, -1, 0]	3

Table 3 Merits and demerits of DNA numerical representations.

Method	Merits	Demerits
Voss [7]	Efficient spectral detector of base distribution and periodicity features; offering numerical and graphical visualization.	Linearly dependent set of representation; redundancy.
Tetrahedron [2]	Periodicity detection.	Reduced redundancy.
Integer [9]	Simple integer representation. So it is computationally efficient.	(A,G) > (C,T); introducing mathematical properties not present in DNA sequence.
Real [12]	A-T and C-G are complement.	Introducing mathematical properties not present in DNA sequence.
Complex [8] - [11]	A-T and C-G are complex conjugate; reflecting complementary feature of nucleotides, more accurate in gene prediction.	Introducing base bias in time domain analysis.
Quaternion [13] - [14]	Overcoming base bias.	Working with DQFT only.
EIIP [18] - [19]	Reflecting DNA physico chemical property; reducing computational overhead; improving gene discrimination capability.	Failing to detect coding region in some genomes.
Atomic number [20]	Reflecting DNA physico chemical property.	Requiring further exploration.
Paired Numeric [21] - [22]	Reflecting DNA structural property; reduced complexity and DFT processing; improved coding region identification accuracy over other methods.	Requiring further exploration.
DNA walk [23] - [24]	Providing long range correlation information; sequence periodicities; changes in nucleotide composition; offering numerical and graphical visualization.	Not suitable for lengthy sequences (> 1000 bases).
Z-curve [25] - [26]	Clear biological interpretation; reduced computation: independent its x_n, y_n, z_n components; superior to sliding window technique; offering numerical and graphical visualization.	Not suitable for long lengthy sequences

Table 4: The five coding regions in Gene F56F11.4 and length of each exon using Voss, complex and EIIP mapping methods compared with that obtained in NCBI.(Le: length of sequence)

Method	Exon1 region (Le)	Exon2 region (Le)	Exon3 region (Le)	Exon4 region (Le)	Exon5 region (Le)
NCBI ranges	7947-8059 (112)	9548-9879 (331)	11134-11397 (263)	12485-12664 (179)	14275-14625 (350)
Voss mapping	7921-8021 (100)	9521-9821 (300)	11021-11221 (200)	12321-12521 (200)	14281-14621 (340)
Complex mapping	7950-8156 (207)	9549-9878 (330)	11135-11398 (264)	12486-12665 (180)	14276-14626 (351)
EIIP mapping	7821-8021 (200)	9521-9821 (300)	11021-11221 (200)	12421-12621 (200)	14221-14621 (400)

The above results show that the three numeric representations all give approximately equivalent DFT-based gene and exon prediction accuracy. In [4] improvements in detection accuracy was gained through the use of forward and backward sliding window DFTs, at the cost of increased complexity (paired number method). By comparison with the paired numeric reveal improved DFT-based gene and exon prediction with 75% less downstream processing. Paired numeric is the most accurate representation for this application, due to the fact that the approach exploits a key statistical property (according to which introns are rich in nucleotides “A ” and “T ” whereas exons are rich in nucleotides “ C“ and “ G“) for discriminating between structures of the genomic protein coding and non-coding regions.

V. Simulations and Results

The base sequences of Eucaryotes (cells with nucleus) have a period-3 property. Such this periodicity is found in the protein-coding regions of DNA (exons) and not found in the introns. This periodicity because of the codon structure involved in the translation of base sequences into amino acids. As indicated in the above section to perform gene prediction based on the period-3 property using STFT, the DNA sequence must be mapped to numerical sequence as introduced in previous sections. Some of these methods are effective in classifying exons and introns sequences and the other methods have their own applications. In order to measure the precision of each method in classification of intron and exons sequences. Discrete Fourier transform (DFT) based approach is adopted to extract the period-3 value of DNA sequences. Let a numerically represented DNA sequence of one-dimension methods is $x(n)$ for n =1to N, so its finite-length DFT sequence, $X[k]$ for k = 1 to N, is defined by [27].

$$X[k]= \frac{1}{\sqrt{N}} \sum_{n=1}^{N} x(n) W_N^{(k-1)(n-1)}, \qquad (31)$$

$$for \quad 1 \leq k \leq N \text{ and } W_N = e^{\frac{-j2\pi}{N}}$$

Using the windowing approach with a rectangular window length of L bases and an overlap width of L-3 bases between two adjacent windows, the normalized sum ($X_T[k]$) of the DFT spectrum ($X_m[k]$) of each of the windowed sequences ($x_m(n)$for m = 1 to N_w) gives.

$$X_T[k]= \frac{1}{N_w} \sum_{m=1}^{N_w} X_m[k] \qquad (32)$$

The spectral content measure can be obtained by taking the power spectrum of equation (32) as

$$S[k] = |X_T[k]|^2 \qquad (33)$$

The period-3 spectral component (P_3) can be obtained from the spectral content measure of a numerically represented sequence as

$$P_3 = S[N/3 \ +1] \qquad (34)$$

The statistics of the period-3 values determined from a training set of exon sequences and intron sequences can be used to classify an untrained sequence to be either an exon sequence or an intron sequence. Let meanP_{3e} and sdP_{3e} represent respectively the mean and standard deviation of the period-3 values obtained from the exon sequences of a training set; and meanP_{3i} and sdP_{3i} represent respectively the mean and standard deviation of the period-3 values obtained from the intron sequences of the same training set, consequently, the threshold value for classification is defined as [27];

$$T_3 = \frac{sdP_{3e}*meanP_{3i} + sdP_{3i}*meanP_{3e}}{sdP_{3e} + sdP_{3i}} \qquad (35)$$

Using (35), if a test sequence has a period-3 value P_{3t} greater than or equal T_3, the test sequence is classified as an exon sequence; otherwise it is classified as an intron sequence. By applying equations (31) to (35) on the training and testing of all the one-dimension numerical representation methods on a database of 50 to 300 base sequences downloaded from USCS Assembly: Feb. 2009 (GRCh37/hg19) (Clade: Mammal, Genome: Human, Assembly: Feb.2009 (GRCh37/hg19), Group: Genes and Gene prediction Tracks, Track: UCSC Genes, Table: knownGene) [28] - [31]. It consists of (a) Exon: total 8000 sequences; sequences 1 to 6000 are used for training and sequences 6001 to 8000 are used for testing; and (b) Intron: total 8000 sequences; sequences 1 to 6000 are used for training and sequences 6001 to 8000 are used for testing. The Exon Classification (EXCLASS) and Intron Classification (INCLASS) performances [32] of all the eight numerical representation methods which given by equations (36) to (38) with a window length of 15 bases and an overlap window length of 12 bases are summarized in Table 5.

$$\text{EXCLASS} = \frac{\text{NCEC}}{\text{Exon number}} \times 100\ \% \qquad (36)$$

$$\text{INCLASS} = \frac{\text{NCIC}}{\text{Intron number}} \times 100\ \% \qquad (37)$$

$$\text{Precision} \ = \frac{\text{NCEC+ NCIC}}{\text{TEIN}} \times 100\% \qquad (38)$$

Where, NCEC, NCIC and TEIN are the number of correct exons and the number of correct introns classifications and the total number of exons and introns respectively.

From the above results, it can be deduced that the paired numeric method achieves the highest precision 82.7624283 %, followed next by EIIP method 80.3867416 % , but the atomic number method achieves the lowest precision 72.4309387% in classifying numerically represented exon and intron sequences that have not been trained.

VI. Conclusion

Choosing one of the numerical representation techniques to be used in association with DSP depends on a particular application. Primarily, fixed mapping representation methods, such as the Voss or the tetrahedron maps a DNA sequence onto four or three

numerical sequence, potentially introducing different redundancy in each individual representation. The arbitrary assignment of integer and real number to DNA nucleotides does not necessarily reflect the structure present in the original DNA sequence. The

quaternion approach requires further exploration. The physico chemical property based mapping techniques such as the EIIP mappings, the atomic number, the paired numeric, the DNA walk, and the Z-curve, in which each method exploits the structural difference of protein coding and non coding regions, facilitates the DSP-based gene and exon predictions.

These methods contain less redundant and carry biological interpretations. In particular, there are some studies indicate that the Z-curve representation is robust and computationally efficient for DNA sequence analysis in which each of its x_n, y_n, z_n components is independent and generates a discrete signal that reflects biological properties. Further improvements in gene and exon prediction can be achieved by incorporating more DNA structural properties in existing or new DNA symbolic-to-numeric representations.

Table 5: Exon and intron classification performance by different methods

Method	Representation	Threshold Value	Classification Performance (%)		Precision
			Exon	Intron	
Integer [9]	$A = 2, C = 1, G = 3, T = 0$	0.109372	77.003486	82.718650	80.0000000
	$A = 0, C = 1, G = 3, T = 2$	0.090457	66.550521	84.615387	76.0221024
Real [12]	A=-1.5, C=0.5, G=-0.5, T=1.5	0.077127	75.377464	71.022125	73.0939255
Complex [8] - [11]	$A = 1 + j,\ C = -1 + j,$ $G = -1 - j,\ T = 1 - j$	0.140288	57.607433	88.198104	73.6464081
EIIP [18] - [19]	A=0.1260, C=0.1340, G=0.0806, T=0.1335	0.000033	81.300811	79.557426	80.3867416
Atomic number [20]	A=70, C=58, G=78, T=66	4.470244	59.581882	84.088516	72.4309387
Paired Numeric [21] - [22]	$A\ or\ T = 1,\ C\ or\ G = -1$	0.104576	72.357727	92.202316	82.7624283
DNA walk [23] - [24]	$C\ or\ T = 1,\ Aor\ G = -1$	0.0741001	68.989547	76.290832	72.8176804

References

[1] Vikrant Tomar, Dipesh Gandhi, and Vijaykumar Chakka, Advanced Filters for Genomic signal processing [J]. Int. J. Adapt. Control Signal Process.

[2] B. D. Silverman and R. Linker, A measure of DNA periodicity [J].Theor. Biol.,118:295-300.

[3] D.G. Grandhi and C. Vijaykumar, 2-Simplex Mapping for Identifying the Protein Coding Regions in DNA [C]. TENCON-2007, Taiwan, Oct. 2007, 530.

[4] M. Akhtar, Julien Epps, and E. Ambikairajah, Signal Processing in Sequence Analysis, Advances in Eukaryotic Gene Prediction [J]. IEEE Journal of selected topics in signal processing, June 2008, 2(3):310-321.

[5] H. K. Kwan and S. B. Arniker, numerical representation of DNA sequences [C]. IEEE Inter, Conf. on Electro/Information Technology, EIT '09, Windsor, 2009:307-310.

[6] P. Ramachandran and A. Antoniou, Genomic Digital Signal Processing. Lecture notes, www.ece.uvic.ca/~andreas.

[7] R. F. Voss, Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical Review Letters, 1992, 68(25):3805–3808.

[8] D. Anastassiou, Genomic signal processing [M]. IEEE Signal Processing Magazine, 2001, 18(4): 8–20.

[9] P. D. Cristea, Genetic signal representation and analysis [C]. in Proc. SPIE Inter. Conf. on Biomedical Optics, 2002, 4623:77–84.

[10] P. D. Cristea, Conversion of nucleotides sequences into genomic signals [J]. Cell. Mol. Med, April-June 2002, 6, 279-303.

[11] P. D. Cristea, Representation and analysis of DNA sequences. in Genomic signal processing and statistics: EURASIP Book Series in Signal Processing and Communications, (Eds) Edward R. Dougherty et al Hindawi Pub. Corp, 2005, 2 :15-66.

[12] N. Chakravarthy, A. Spanias, L. D. Lasemidis, and K. Tsakalis, Autoregressive modeling and feature analysis of DNA sequences [J]. EURASIP Journal of Genomic Signal Processing, January 2004, 1:13-28.

[13] M. Akhtar, J. Epps, and E. Ambikairajah, On DNA numerical representations for period-3 based exon prediction. in Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Tuusula, June 2007:1-4.

[14] A. K. Brodzik and O. Peters, Symbol-Balanced Quaternionic Periodicity Transform for Latent Pattern Detection in Dna Sequences [C]

Proceedings of IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing, ICASSP '05, 2005, 5: 373-376.

[15] T. P. George and T. Thomas, Discrete wavelet transform de-noising in eukaryotic gene splicing. BMC Bioinformatics 2010

[16] NCBI GenBank database, online access: http://www.ncbi.nlm.nih.gov/Genbank/.

[17] Stuart W. A. B and A. Antoniou, Application of Parametric Window Functions to the STDFT For

Gene Prediction [C]. IEEE Pacific Rim Conf. communications, computers and signal Processing, PACRIM’05, 2005:324-327.

[18] Achuthsankar S. Nair and Sreenadhan S. Pillai, A coding measure scheme employing electron-ion interaction pseudo potential (EIIP), Bio-information, Oct. 2006, 1: 197-202.

[19] I. Cosic, Macromolecular Bioactivity: Is it resonant interaction between macromolecules? Theory and Applications. IEEE Transactions on Biomedical Eng., Dec. 1994, 41:1101-1114.

[20] Todd Holden, R. Subramaniam, R. Sullivan, E. Cheng, C. Sneider, G.Tremberger, Jr. A. Flamholz, D. H. Leiberman, and T. D. Cheung, ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes. in Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE), August 2007, 6694:. 669417-1 to 669417-10.

[21] S. V. Buldyrev, A. L. Goilberger, S. Havlin, R. N. Mantegna, M. E. Mastsa, C.-K. Peng, M. Simons, and H. E. Stanley, Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phy. Rev. E, May 1995, 51(5): 5084-5091.

[22] M. Akhtar, J. Epps, and E. Ambikairajah, “Paired Spectral Content Measure for Gene and Exon Prediction in Eukaryotes [C]. Inter. Conf. on Information and Emerging Technologies, ICIET’07, Karachi, July 2007:1- 4.

[23] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, Analysis of DNA sequences using methods of statistical physics, Physica A, Elsevier Science B.V, 1998, 249: 430-438.

[24] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, Visualization and analysis of DNA sequences using DNA walks [J]. Journal of the Franklin Institute, January-March 2004, 341:37-53.

[25] R. Zhang and Chun-Ting Zhang, Identification of replication origins in archaeal genomes based on the Z-curve method. 2005 Heron Publishing-Victoria, Canada, Archaea 1, Nov. 2004:335–346

[26] N. F. Law, K. Cheng and W. Siu, On relationship of Z-curve and Fourier approaches for DNA coding sequence classification, Bioinformation, 2006, 1(7) : 242-246.

[27] J. Y. Y. Kwan, B. Y. M. Kwan and H. K. Kwan, Spectral analysis of numerical exon and intron sequences [C]. Proceedings of IEEE Inter. Conf. on Bioinformatics and Biomedicine Workshops, Hong Kong, 2010 : 876-877.

[28] D Karolchik, AS Hinrichs, TS Furey, KM Roskin, CW Sugnet, D Haussler, WJ Kent, The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32(Database issue), 2004, D493-496.

[29] J Goecks, A Nekrutenko, J Taylor, The Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 2010, 11(8).

[30] D Blankenberg, G Von Kuster, N Coraor, G Ananda, R Lazarus, M Mangan, A Nekrutenko, J Taylor, Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. Chapter 19, 2010, Unit 19.10.1-21.

[31] B Giardine, C Riemer, RC Hardison, R Burhans, L Elnitski, P Shah, Y Zhang, D Blankenberg, I Albert, J Taylor, W Miller, WJ Kent, A Nekrutenko, Galaxy.

[32] J. Y. Y. Kwan, B. Y. M. Kwan and H. K. Kwan, Novel methodologies for spectral classification of exon and intron sequences [J]. EURASIP Journal on Advances in Signal Processing, 2012.

Prof. Mohammed Abo-Zahhad (SIEEEM’00) received his B.S.E.E. and M.S.E.E degrees in electrical engineering in 1979 and 1983 respectively, both from Assiut University, Egypt. In 1988, he received Ph. D. degree from the University of Kent at Canterbury, UK and Assiut University (channel system). His research interests include switched-capacitor, optical and digital filters, biomedical and genomic signal processing, speech processing, data compression, wavelet-transforms, genetic algorithms, immune algorithms, and electronic systems. He has published more than 96 papers in national and international journals and conferences in the above fields. Professor Abo-Zahhad is currently a Professor of Electronics and Communication Engineering, since Jan.1999. Also, he is the director of AU Management Information System (MIS) center and a vice-dean for graduated studies, Faculty of Engineering, Assiut University, since August 2006. He is a member of the European Society of Circuit Theory and Applications, 1998 and a senior IEEE member, 2000**.**

Prof. Sabah M. Ahmed received her B.S.E.E. and M.S.E.E degrees in electrical engineering in 1979 (excellent with honors) and 1983 respectively, both from Assiut University, Egypt. In 1992, she received Ph. D. degree from the Technical University of Budapest, Hungary. Her research interests include speech processing, biomedical and genomic signal processing, data compression, wavelet-transforms, genetic algorithms, and immune algorithms. She has published more than 50 papers in national and international journals and conferences in the above fields. Professor Sabah is currently a Professor of Electronics and Communication Engineering, since Feb. 2009. Also, she is the director of Faculty of Engineering ICDL center, Assiut University and the manager of Assiut University communication and information technology training center.

Shimaa A. Abd-Elrahman received her B.Sc. (honors) degree in Electrical and Electronics Engineering Department, Faculty of Engineering, Assiut University, Assiut, Egypt, in 2008. She is currently a demonstrator of electrical engineering at Egypt at Assiut Univeristy and pursuing the M.S. degree in Prediction of gene locations in DNA sequence. Her main research interest is in genomic signal processing with specific focus on DNA representations, exons and introns classification, and gene prediction.