ONE-DIMENSIONAL MODELING OF DNA SEQUENCES

[1]Alain B. Tchagang and [2]Ahmed H. Tewfik

[1]Department of Biomedical Engineering, [2]Electrical and Computer Engineering University of Minnesota
200 Union Street S.E., 55455, Minneapolis, USA
phone: + (1) 612 625 6024, fax: + (1) 612 625 4583, email: {tcha0003, tewfik}@umn.edu
web: http://www.ece.umn.edu/~tewfik

ABSTRACT

Signal processing tools are increasingly being used for DNA analysis, including sequence structure prediction, sequence matching and classification and sequence function identification. Almost all of the proposed techniques perform the DNA analysis in a higher dimensional space, typically of dimension 4 to match the four-letter alphabet {A, C, T, G} representation of DNA sequences. In this study, we show that it is possible to obtain similar results to those reported in the literature by processing DNA data in a 1-D space with lower complexity. We illustrate our approach by focusing on the gene identification problem and comparing the 1-D processing results with prior work.

1. GENERAL INFORMATION

Signal processing techniques have recently been extensively used in the field of genomics to analyze, visualize, classify, and identify *DNA* sequences. Most of the approaches reported in the literature rely on a higher dimensional analysis of the *DNA* data. The need for an analysis in a higher dimensional space appears to be a direct result of the fact that *DNA* sequences are represented by strings of characters that belong to a four-letter alphabet *{A, C, T, G}* [1]. Hence, most researchers have used analysis in a *4-D* space, e.g., [2,3,4,5,6]. While these pioneering approaches have proven to be successful in addressing several challenges in genomics, such as sequence structure prediction, sequence matching and classification and sequence function identification, they appear to be needlessly complex requiring a very high complexity. In particular, we show in this study, that we can reproduce the results of prior research by processing DNA data in a *1-D* space with much lower complexity.

We illustrate the fact that 1-D analysis can yield results similar to those obtained in the past in higher dimensional space by considering the gene identification problem. The gene identification problem is the problem of interpreting nucleotide sequences by computer, in order to provide tentative annotation on the location, structure, and functional class of protein-coding genes. In other words, the purpose is to identify the genes and determine the functions of the proteins they encode. This process is essential, since without it a sequenced genome is merely a meaningless jumble of *A*'s, *C*'s, *T*'s, and *G*'s. Genes can be identified by methods confined to a single genome or by comparative methods that use information about one organism to understand another related one. As mentioned above, signal processing tools such as the Short Time Fourier Transform (*STFT*) and digital filters exploit the period-3 behavior observed within the protein-coding region of some *DNA* sequence to solve the problem of gene identification. Previous analyses apply those theories on a *4-D* sequence that is used to model the *DNA*. In this paper, we show that we can perform these same analyses in a *1-D* space using a *1-D* representation called the *DNA* Reproduction Sequence. The results obtained in each case are the same as those obtained by previous researchers in [2,3,4,5,6] but with much lower complexity. This establishes the usefulness of the *DNA* Reproduction Sequence. We report the results of using the *1-D* model of the *DNA* to study and analyze many other properties exhibited by the *DNA* in [7].

This paper is organized as follows. First, we develop a mathematical representation of the *DNA* strand. Secondly, we derive a similitude between the *DNA* representation and some color spaces. Then, color space transformations are applied to switch from one color space to another. In each color space, we perform the spectrum analysis of the given *DNA*. The spectrum obtained allows us to tell whether or not the properties of the given *DNA* are modified. We then derive from that construction a *1-D* sequence which is used to analyze and predict some interesting properties exhibited by *DNA* sequences. The biological properties of the *DNA* can be found in [1].

2. DNA: MATHEMATICAL REPRESENTATION

2.1. Mathematical Model

A strand of *DNA* can be viewed as a string of *N* characters *x*, where *x* belongs to *{A, C, T, G}*. Each *x* is called a nucleotide. Prior work assigned numbers *a*, *c*, *t*, and *g* to the characters *A*, *C*, *T*, and *G* respectively. Here we will use equation (1) to model a given DNA sequence.

$$u[n] = u_A[n]e_A + u_C[n]e_C + u[n]_T e_T + u_G[n]e_G \quad (1)$$

with $e_A=[a\ 0\ 0\ 0]^T$, $e_C=[0\ c\ 0\ 0]^T$, $e_T=[0\ 0\ t\ 0]^T$, $e_G=[0\ 0\ 0\ g]^T$ and $u_A[n]$, $u_C[n]$, $u_T[n]$, $u_G[n]$, are the Binary Indicator Sequence [3,4]. They are defined as in (2), with $n = 0$ to $N-1$.

$$u_X[n] = \begin{cases} 1 & Nucleotide\ \ x\ \ \text{Pr}\,esent \\ 0 & Nucleotide\ \ x\ \ Absent \end{cases} \quad (2)$$

In our study, we will consider a pure *DNA* sequence, that is *a = c = t = g = 1*, thus: $e_A=[1\ 0\ 0\ 0]^T$, $e_C=[0\ 1\ 0\ 0]^T$, $e_T=[0\ 0\ 1\ 0]^T$, and $e_G=[0\ 0\ 0\ 1]^T$.
Now define the *4 x N* Indicator Sequence Matrix *M* as follows. The columns of *M* are either e_A, e_C, e_T, or e_G. They show the content of the *DNA* sequence at each position *n*. The rows of *M* are respectively $u_A[n]$, $u_C[n]$, $u_T[n]$, and $u_G[n]$, with *n = 0* to *N-1*. They each show the evolution of the respective nucleotide in a *DNA* sequence. For example, the Indicator Sequence Matrix *M* of the *DNA* sequence: *AATCGGCCTG*, with *N = 10*, is:

$$M = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

2.2. *DNA* and Digital Color: from *ACTG* → *CMYK*

The *DNA* sequence: *u[n]* of length *N* with *n = 0* to *N-1* can be mapped into the color space: *CMYK* as follows: Adenine (*A*) → Cyan (C_y), Cytosine (*C*) → Magenta (*M*), Thymine (*T*) → Yellow (*Y*), and Guanine (*G*) → Black (*K*). Thus we can also define: $u_A[n] = u_{Cy}[n]$, $u_C[n] = u_M[n]$, $u_T[n] = u_Y[n]$, and $u_G[n] = u_K[n]$. Where $u_{Cy}[n]$, $u_M[n]$, $u_Y[n]$, and $u_K[n]$ are the rows of the matrix $M_{CMYK} = M$ in *CMYK* color space. We also have: $e_A = e_{Cy}=[1\ 0\ 0\ 0]^T$, $e_C = e_M = [0\ 1\ 0\ 0]^T$, $e_T = e_Y = [0\ 0\ 1\ 0]^T$, $e_G = e_K = [0\ 0\ 0\ 1]^T$. With e_{Cy}, e_M, e_Y, and e_K in *CMYK* color space. We assume that those colors are binary, thus they can either be *"0"* or *"1"*. We observe that, each vector can be seen as belonging to the color space *CMYK* or describing a given *DNA* sequence at a specific position *[7]*.

2.3. Color Space Transformation

We have studied several color space transformations to either reduce the dimension or to change the space of analysis [7]. Here we report the results that we have obtained for the gene identification problem. After each color space transformation, Fourier analysis is performed to study the spectrum of the given *DNA*. The method used to perform the spectrum analysis is the same defined in [2,3,5]. That is, the Fourier transform of each row of *M* is taken. Equation (3) is used to measure the frequency content in the range *k = 0* to *N/2*.

$$S_{ACTG}[k] = |U_A[k]|^2 + |U_C[k]|^2 + |U_T[k]|^2 + |U_G[k]|^2 \qquad (3)$$

For illustration purposes, the following *DNA* sequence has been used: *PHIX174* (accession number *V01128*), with *N = 5,386* base pairs (bp). It is known to have a period-3 behavior. In the *CMYK* space (*4-D* space), the spectrum is given by equation (4): With *k = 0* to *N/2*.

$$S_{CMYK}[k] = |U_{Cy}[k]|^2 + |U_M[k]|^2 + |U_Y[k]|^2 + |U_K[k]|^2 \qquad (4)$$

Fig.1 shows the spectrum. We observe a peak at frequency *N/3* corresponding to period-*3*.

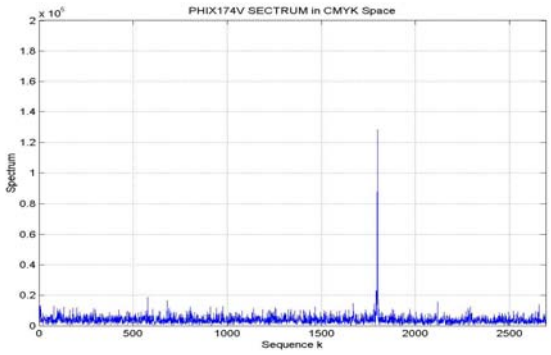

Fig.1. Spectrum analysis in *CMYK* space

2.3.1 Analysis into cmy color Space and other 3-D Color Spaces

We can use equation (5) to convert an element of *CMYK* into an element of *cmy* color space, thereby reducing the dimension of the analysis space from four to three.

$$\begin{cases} c = C_y(1-K) + K \\ m = M(1-K) + K \\ y = Y(1-K) + K \end{cases} \qquad (5)$$

The Indicator Sequence Matrix of the *DNA* sequence: *AATCGGCCTG* becomes:

$$M_{cmy} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The rows of M_{cmy} are now $u_c[n]$, $u_m[n]$, and $u_y[n]$. The spectrum analysis is performed using equation (6) with *k = 0* to *N/2*.

$$S_{cmy}[k] = |U_c[k]|^2 + |U_m[k]|^2 + |U_y[k]|^2 \qquad (6)$$

Fig.2 shows the spectrum of the *DNA* sequence *PHIX174*. Note that the result is identical to those in the *CMYK* space and [5]. The peak at frequency *N/3*, and thus the period-3 behavior, is conserved.

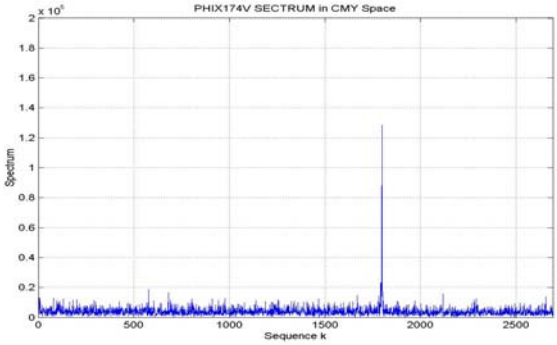

Fig.2. Spectral analysis in *cmy* space

We have performed similar spectral analyses in many other *3-D* color spaces model, including *RGB, CIEXYZ, CIELuv, CIELa*b**, etc. In each case, the appropriate transformation [8] is used to switch from one color space to another and the spectrum is computed. As reported in [7], these transformations do not modify the properties of a given *DNA* captured by the techniques discussed in [5] and similar references.

2.3.2 The DNA Reproduction Sequence E[n]

A *DNA* reproduction sequence is any *1-D* sequence that can be used to model a strand of *DNA* without changing or modifying its properties. We will use *E[n]* to refer to such a sequence. Sequence *E[n]* is obtained from manipulation in the *CIELa*b** color space [7,8]. It comes from the fact that the concept of Euclidean distance can be applied in the *CIELa*b** color space. The distance is defined as: $\Delta E = [(\Delta L)^2+(\Delta a)^2+(\Delta b)^2]^{1/2}$ and it is always used in that space to compare two different colors. Here, we defined *E[n]* as:

$$E[n] = \sqrt{(L(n))^2 + (a(n))^2 + (b(n))^2}\ . \qquad (7)$$

With *n = 0* to *N-1*. Equation (7) allows us to reduce the dimension of the sequence from *3* to *1*. For example, the DNA sequence *AATCGGCCTG* with Indicator Sequence Matrix M_{Lab} into *CIELa*b** space [7].

$$M_{Lab}=\begin{bmatrix} 5505 & 5505 & 6486 & 1991 & 0 & 0 & 1991 & 1991 & 6486 & 0 \\ -6667 & -6667 & -3546 & 1021 & 0 & 0 & 1021 & 1021 & -3546 & 0 \\ -2350 & -2350 & 9650 & -7303 & 0 & 0 & -7303 & -7303 & 9650 & 0 \end{bmatrix}$$

becomes

$$E=\begin{bmatrix} 89.64 & 89.60 & 121.56 & 127.14 & 0 & 0 & 127.14 & 127.14 & 121.56 & 0 \end{bmatrix}.$$

Spectrum analysis can be performed using (8) to yield :

$$S_E[k] = |U_E[k]|^2, \qquad k = 0 \text{ to } N/2 \qquad (8)$$

Fig.3 shows the spectrum of the *DNA* sequence *PHIX174* using *E[n]*. Observe the peak at frequency *N/3* and, the fact that the spectrum is again similar to that of [5], *CMYK* color space and other *3-D* color spaces [7]. Thus the period-3 behavior is conserved so are the properties of the *DNA* captured by the techniques discussed in [5] and similar references.

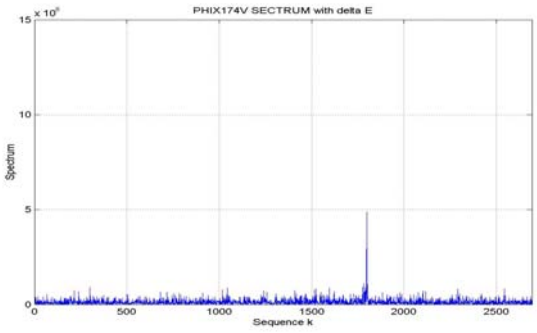

Fig.3. Spectrum analysis using *E[n]*

3. GENE IDENTIFICATION

3.1. Short Time Fourier Transform

To solve the problem of gene identification for DNA with period-3 behavior, references [2,3] use the magnitude square of a variable *W* given by

$$W = \frac{1}{N}\left[aU_A\left[\frac{N}{3}\right] + cU_C\left[\frac{N}{3}\right] + tU_T\left[\frac{N}{3}\right] + gU_G\left[\frac{N}{3}\right]\right]. \qquad (9)$$

In (9), $U_A[N/3]$, $U_C[N/3]$, $U_T[N/3]$, $U_G[N/3]$ are the Discrete Fourier Transform of $u_A[n]$, $u_C[n]$, $u_T[n]$, $u_G[n]$ respectively at frequency *N/3*. The coefficients *a, c, t, g* are complex numbers that are chosen to maximize the discriminatory capability between protein coding regions and random DNA regions, by solving a complex optimization problem (c.f., [2,3]). Here, we solve the same problem by studying the magnitude square of equation of a different variable *W* given by.

$$W = \frac{1}{N}\left[U_E\left[\frac{N}{3}\right]\right] \qquad (10)$$

In (10), $U_E[N/3]$ is the DFT of the *1-D* sequence *E[n]* at frequency *N/3*. Equation (10) can be seen as the *1-D* of equation (9) but in this case, there is no need to solve and optimization problem as with equation (9). In particular, we also observe that the complexity is reduced from 4 to 1. For simulation purposes, we have used the: C.elegans (accession number *AF099922*) *DNA* sequence, with *N = 8,000* nucleotides starting from location 7021. It is known to have a period 3 behavior and 5 Exons. Fig.4 shows the result we obtained. We observe the five genes of C.elegans. Indeed, we get the same results as in [2,3].

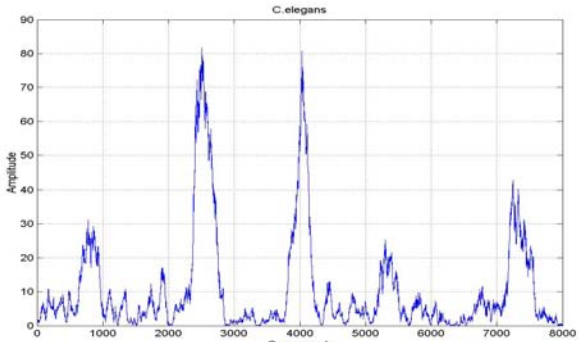

Fig.4. Five exons of C.elegans with STFT and *E[n]*

3.2. Digital Filter Approach

In [4,6], a pass band digital filter with a narrow pass band frequency center at frequency *2π/3* was proposed to solve the problem of gene identification for *DNA* with period-3 behavior. It is shown to be faster than the *STFT*, and to also suppress the *1/f* background noise exhibited by almost every *DNA* sequence. An anti-notch digital filter centered at frequency *2π/3* was therefore designed for this purpose. Details

on how to design the filter can be found in [4,6]. Each binary indicator sequence $u_A[n]$, $u_C[n]$, $u_T[n]$, and $u_G[n]$ is passed through an anti-notch digital filter center at frequency $2\pi/3$. The magnitude square of the sum of the outputs is then used to predict the location of the genes in a *DNA* sequence using the following: equation:

$$Y_{ACTG}[k] = |Y_A[k]|^2 + |Y_C[k]|^2 + |Y_T[k]|^2 + |Y_G[k]|^2 \qquad (11)$$

Instead of using four anti-notch filters we have used one anti-notch filter. That is, the *1-D* sequence *E[n]* is passed through one anti-notch filter centered at frequency $2\pi/3$, and the magnitude square of the output is used to predict the genes in a *DNA* sequence with period-3 behavior using the following equation:

$$Y[k] = |Y_E[k]|^2 \qquad (12)$$

The complexity is therefore again reduced from 4 to 1. Fig.5 shows the result obtained. They are the same as those of [4,6].

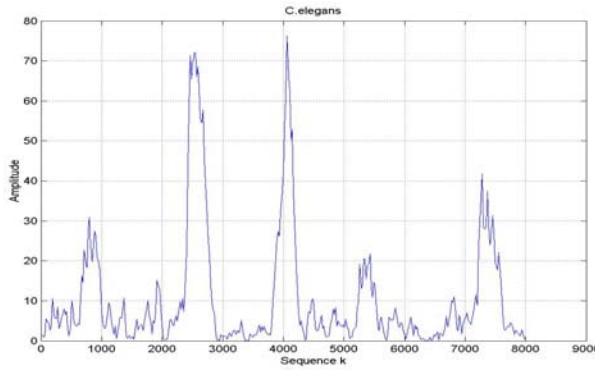

Fig.5. Five exons of C.elegans with Digital filter and *E[n]*

4. CONCLUSION

In this paper, a *1-D DNA* sequences analysis has been proposed. The proposed method treats a *DNA* sequence as a *1-D* sequence even given the fact that four-letter alphabet *{A, C, T, G}* is considered. Therefore, we obtained low complexity that is, the complexity of previous methods found in the literature is reduced from four to one when using the *1-D* analysis to solve the problem of gene identification for example, and there is no need to solve an optimization problem as proposed by some methodologies. Many other properties of the *DNA* have been analyzed using the proposed method, other techniques and properties that are well developed and established in digital color theory. The results obtained so far are promising and, they have a bright future in the field of Genomics.

REFERENCES

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Essential Cell Biology. New York: Garland Publishing, 1998

[2] D. Anastassiou, "Digital Signal Processing of Biomolecular Sequences," Tech. Rep. EE000420-1, Apr.2000. Available: http://www.ce.columbia.edu/cgi-ee-bin/show_archive.pl.

[3] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," Bioinformatics, vol. 16, no.12, pp. 1073-1082, Dec. 2000

[4] P.P. Vaidyanathan, B.J. Yoon, "Digital filter for Gene prediction applications", IEEE Signal processing, 2002

[5] J.A. Berger, S.k. Mitra, J. Astola, "Power Spectrum Analysis for DNA Sequences," IEEE, Signal Processing, 2003

[6] P.P. Vaidyanathan, B.J. Yoon, "The role of signal-processing concepts in genomics and proteomics", The Journal of the Franklin Institute, 200

[7] A.B. Tchagang, "Genomics Signal Processing: Technical Report, University of Minnesota, 2007.

[8] G. Sharma, Digital color imaging Handbook, CRC Press. 2003.