Numerical Representation of DNA Sequences

Hon Keung Kwan and Swarna Bai Arniker
Department of Electrical and Computer Engineering
University of Windsor
401 Sunset Avenue
Windsor, Ontario, Canada N9B 3P4
e-mail: kwan1@uwindsor.ca, arniker@uwindsor.ca

***Abstract*—DNA sequence analysis using digital signal processing requires conversion of a base sequence to a numerical sequence. The choice of the numerical representation of a DNA sequence affects how well its biological properties can be reflected in the numerical domain for the detection and identification of the characteristics of special regions of interest. This paper presents some selected methods of DNA numerical representation for DNA sequence analysis, discusses their relative merits and demerits, and includes some concluding remarks.**

I. INTRODUCTION

Since the completion of the Human Genome Program (HGP) [1], there has been a need to analyze information contained in a growing volume of deoxyribonucleic acid (DNA) sequence database of the human and model organisms. Digital signal processing (DSP) approach has been used increasingly in genomic DNA research to reveal genome structures to identify hidden periodicities and features which cannot be revealed by conventional DNA symbolic and graphical representation techniques [2].

In genomic signal processing (GSP), the mapping of the discrete bases of a DNA sequence to a discrete numerical sequence is required for DSP-based analysis [3]-[5]. A simple and commonly used mapping scheme for this purpose has been the Voss representation [6]. However, many other methods have also been introduced such as the tetrahedron [7], the integer number [8], the real number [9], the complex number [8], the quaternion [10], the electron-ion interaction potentials (EIIP) [11], the atomic number [12], the paired numeric [10], the DNA walk [13], and the Z-curve [14].

Reference [9] describes the effects of binary representation; the integer, real, complex representations; and the DNA walk representation for autoregressive modeling and feature analysis of DNA sequences. In [10] the Voss, the tetrahedron, the real number and its variants, the complex, the quaternion, the EIIP, the paired numeric, and the Z-curve representations are compared using the discrete Fourier transform for period-3 based exon (coding region) prediction. In this paper we shall discuss, classify, and compare, selected DNA numerical representation methods for digital signal processing and analysis.

II. NUMERICAL REPRESENTATION METHODS

DNA contains the genetic instructions of biological processes, stored inside its molecules, for the development of each living organism and virus. DNA consists of double stranded anti-parallel helix built by concatenating nucleotides consisting of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Complementary property exists between DNA double-strands, as an A on one strand always binds with a T on the other strand, and similarly, a C always binds with a G. In order to apply digital signal processing, the bases of a DNA sequence are to be mapped onto their corresponding numerical values. Numerical representation methods of DNA sequences summarized in Table I can be broadly classified into two major groups as described in the following sub-sections.

A. Fixed Mapping

In fixed mapping techniques, the nucleotides of DNA data are transformed into a series of arbitrary numerical sequences. Fixed mapping include the Voss, the tetrahedron, the integer, the real, and the complex representations.

The Voss representation [6] maps the nucleotides A, C, G, and T into four binary indicator sequences as A_n, C_n, G_n, and T_n showing the presence with 1 or absence with 0 of the respective nucleotide. In the tetrahedron method [7], the four sequences [A_n, C_n, G_n, T_n] are mapped to the four vertices of a regular tetrahedron which reduces the number of indicator sequences from four to three but in a manner symmetric to all the four sequences.

The integer representation [8] is a one-dimensional (1-D) mapping of DNA bases which 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). In the real number representation [9], [15], A = -1.5, T = 1.5, C = 0.5, and G = -0.5, which bears complementary property and is efficient in finding the complimentary strand of a DNA sequence. However, the assignment of a real number to each

978-1-4244-3355-1/09/$25.00©2009 IEEE

of the four bases does not necessarily reflect the structure present in a DNA sequence.

The complex representation [3], [8], [16], [17] reflects the complementary nature of A-T and C-G pairs as A = 1+j, C = -1+j, G = -1-j, and T = 1-j. In the quaternion representation [10] of DNA bases, pure quaternions are assigned to each base: A= i+j+k, C= i-j-k, G= -i-j+k, and T= -i+j-k.

B. Physico Chemical Property Based Mapping

In this type of mapping, biophysical and biochemical properties of DNA biomolecules are used for DNA sequence mapping, which is robust and is used to search for biological principles and structures in biomolecules. This mapping includes the EIIP, the atomic number, the paired numeric, the DNA walk, and the Z-curve representations.

EIIP represents the distribution of the free electrons' energies along a DNA sequence. A single EIIP indicator sequence [11], [18] 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. A single atomic number indicator sequence [12] is formed by assigning the atomic number in each nucleotide as: A=70, C=58, G=78 and T=66 in a DNA sequence.

In the paired numeric representation [10], 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.

The DNA-Walk model [13], [19] shows a graph of a DNA sequence in which a step is taken upwards (+1) if the nucleotide is pyrimidine (C or T) or downwards (-1) if it is purine (A or G). The graph continues to move upwards and downwards as the sequence progresses in a cumulative manner, with its base number represented along the x-axis. The DNA walk can be used as a tool to visualize changes in nucleotide composition, base pair patterns, and evolution along a DNA sequence.

The Z-curve [14] is a 3-D curve that provides a unique representation for visualization and analysis of a DNA sequence. The three components of the Z-curve, $\{x_n, y_n, z_n\}$, represent three independent nucleotide distributions that completely describe a DNA sequence. The components x_n, y_n, z_n display respectively the distributions of purine versus pyrimidine (R versus Y), amino versus keto (M versus K), and strong H-bond versus weak H-bond (S versus W) bases along the sequence.

III. MERITS AND DEMERITS

Numerical representation of a DNA sequence when it is being used in conjunction with DSP techniques can identify hidden periodicities, nucleotide distributions, and features that cannot be revealed easily by conventional methods such as DNA symbolic and graphical representations. A summary of the merits and demerits of all the methods presented in this paper is shown in Table II.

TABLE I. DNA NUMERICAL REPRESENTATION (1: VOSS; 2: TETRAHEDRON; 3: INTEGER; 4: REAL; 5: COMPLEX; 6: QUATERNION; 7: EIIP; 8: ATOMIC NUMBER; 9: PAIRED NUMERIC; 10: DNA WALK; 11: Z-CURVE; I: NUMBER OF INDICATOR SEQUENCES)

	Representation	**S(n) = [CGAT]**	**I**
1	$X_n = 1$ for S(n) = X $X_n = 0$ for S(n) ≠ X X_n applies to any of 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
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
3	A = 2, C = 1, G = 3, T = 0	[1, 3, 2, 0]	1
4	A = -1.5, C = 0.5, G = -0.5, T = 1.5	[0.5, -0.5, -1.5, 1.5]	1
5	A = 1+j, C = -1+j, G = -1-j, T = 1-j	[-1+j, -1-j, 1+j, 1-j]	1,4
6	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
7	A= 0.1260, C= 0.1340, G= 0.0806, T= 0.1335	[0.1340, 0.0806, 0.1260, 0.1335]	1,4
8	A = 70, C = 58, G = 78, T = 66	[58, 78, 70, 66]	1,4
9	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
10	C or T = 1, A or G = -1	[1, 0, -1, 0]	1
11	$x_n=(A_n+G_n)-(C_n+T_n)\equiv R_n-Y_n$ $y_n=(A_n+C_n)-(G_n+T_n)\equiv M_n-K_n$ $z_n=(A_n+T_n)-(C_n+G_n)\equiv W_n-S_n$	x_n=[-1, 0, 1, 0] y_n=[1, 0, 1, 0] z_n=[-1,-2,-1, 0]	3

Each of the DNA numerical representations in fixed mapping offers different properties, and maps a DNA sequence into one to four numerical sequences. The Voss binary indicator representation of a DNA sequence does not predefine any mathematical relationship among the bases, but only indicates the frequencies of the bases. Studies [3], [20], [21] indicate that the Voss representation 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, utilized in the program 'Genescan' [20], reveals a peak at frequency 0.33 (period-3) for coding regions and shows no peak for noncoding regions. It operates on a sliding window known as short-time Fourier transform (STFT) to search for a peak with a strength surpasses a threshold (taking as the average strength of the power spectrum within a region) to indicate the presence of a coding region. Thus the Voss mapping is widely utilized in the base distribution and periodicity detection of a sequence. The Voss and the tetrahedron representations are two equivalent representations when being used in power spectrum analysis.

Arbitrarily assigned integer and real number representations may introduce some mathematical property which does not exist in a base sequence. Hence their DSP applications are limited suggesting that these integer and real mappings need to be used carefully for a given application.

The complex representation reflects some of the complementary features of the nucleotides in its mathematical properties. Quaternion based representation can be analyzed only in conjunction with the DQFT to detect certain DNA patterns. It was conjectured [10] that the quaternion approach could improve DNA pattern detection via the discrete quaternionic Fourier transform (DQFT).

Each of the real, complex, quaternion, EIIP, and atomic number representations can be represented by one or four Voss-like indicator sequences. As compared to the Voss representation, the EIIP representation can improve the discrimination capability of gene finding technique and reduce computational overhead by 75% as shown in [11]. There are a number of genes where both the Voss and the EIIP representation fails to detect coding regions. The atomic number representation has been applied to study the nucleotide fluctuations in radiation resistance-repair genes in Deinococcus radiodurans (DR) and E-coli [12]. The atomic number representation is a recent mapping, further exploration is required to reveal its potentials.

The paired numeric representation incorporates a useful DNA structural property and is characterized by reduced complexity. The resultant one-sequence or two-sequence DNA representation offers reduction in DFT processing compared to the Voss, the Tetrahedron, and the Z-curve representations. The paired numeric representation [10] provides an improved accuracy in identifying protein coding regions over Voss, tetrahedron, integer, real, complex, quaternion, EIIP, and Z-curve representations.

The 3-D DNA walk based on complex representation provides useful information such as long range correlation information, sequence periodicities, and changes in nucleotide composition. This technique is suitable only for DNA sequences of a few hundreds of base pairs, and for lengthy sequences the DNA walk tends to become complicated since there is much information for extracting anything useful.

For the Z-curve, the compositional patterns of genomic sequences can be quickly recognized in a perceivable form. There is a strong correlation between the C+G content and gene density. The higher gene density regions for any chromosome are almost situated at CG-isochores with the highest C+G content. The Z-curve was employed [22]-[23] in protein coding measurement and gene identification of DNA. The distribution of the C+G content in the human genome has been studied by using a windowless technique derived from the Z-curve method [23]. The Z-curve's detected distribution of the C+G content along genomic sequences (such as budding yeast and vibro cholerae genomes [22]) exhibits generally better performance than the widely used sliding-window technique. The Z-curve has been used for horizontally transferred genomic islands detection, comparative genomics, studying the distribution of nucleotide composition [24], and identifying replication origins of archaeal genomes with 3-D Z-curve and 2-D Z-curve based on CG and AT disparity [24]. A Z-curve based wavelet denoising technique for DNA sequences has been studied in [25]-[26] for isochore, coding region, and gene detections.

TABLE II. MERITS AND DEMERITS OF DNA NUMERICAL REPRESENTATION (1: VOSS; 2: TETRAHEDRON; 3: INTEGER; 4: REAL; 5: COMPLEX; 6: QUATERNION; 7: EIIP; 8: ATOMIC NUMBER; 9: PAIRED NUMERIC; 10: DNA WALK; 11: Z-CURVE)

	Merits	**Demerits**
1	Efficient spectral detector of base distribution and periodicity features; offering numerical and graphical visualization.	Redundancy; linearly dependent set of representation.
2	Periodicity detection.	Reduced redundancy.
3	Simple integer representation.	(A, G) > (C, T) ; introducing mathematical properties not present in DNA sequence.
4	A-T and C-G are complement.	Introducing mathematical properties not present in DNA sequence.
5	A-T and C-G are complex conjugate; reflecting complementary feature of nucleotides.	Introducing base bias in time domain analysis.
6	Overcoming base bias.	Working with DQFT only.
7	Reflecting DNA physico chemical property; reducing computational overhead; improving gene discrimination capability.	Failing to detect coding region in some genomes.
8	Reflecting DNA physico chemical property.	Requiring further exploration.
9	Reflecting DNA structural property; reduced complexity; reduced DFT processing; improved coding region identification accuracy over other methods.	Requiring further exploration.
10	Providing long range correlation information; sequence periodicities; changes in nucleotide composition; offering numerical and graphical visualization.	Not suitable for lengthy (> 1000 bases) sequences.
11	Clear biological interpretation; independent x_n, y_n, z_n components; reduced computation; superior to sliding window technique; offering numerical and graphical visualization.	

IV. CONCLUDING REMARKS

Which one of the numerical representation techniques is 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 exploits the structural difference of protein coding and noncoding regions to facilitate DSP-based gene and exon predictions. These methods contain less redundant and carry biological interpretations. In particular, studies in [14], [22]-[24] indicate that the Z-curve representation is robust and computationally efficient for DNA sequence analysis in which each of its x_n, y_n, and z_n components is independent and generates a discrete signal that reflects biological properties. Further study is required to reveal more advanced properties of the selected numerical representation methods discussed in this paper.

REFERENCES

[1] R. J. Robbins, B. David, and S. Jay, "Informatics and the Human Genome Project," IEEE Engineering in Medicine and Biology Magazine, vol. 14, pp. 694-701, Nov.-Dec. 1995.

[2] A. Roy, C. Raychaudhury, and A. Nandy, "Novel techniques of graphical representation and analysis of DNA sequences- A review," Journal of Biosciences, vol. 23, pp. 55-71, March 1998.

[3] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, vol. 18, pp. 8-20, July 2001.

[4] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, "Using signal processing techniques for DNA sequence comparison," in Proc. of the 1989 Fifteenth Annual Northeast Bioengineering Conference, March 1989, pp. 173-174.

[5] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "New Approaches to genome sequence analysis based on digital signal processing," in Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), October 2002, pp. 1-4.

[6] Richard F. Voss, "Evolution of Long-range Fractal Correlations and 1/*f* noise in DNA base sequences," Physical Review Leters, vol. 68, pp. 3805-3808, June 1992.

[7] B. D. Silverman and R. Linker, "A measure of DNA periodicity," J. Theor. Biol., vol. 118, pp. 295-300, February 1986.

[8] P. D. Cristea, "Genetic signal representation and analysis," in Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE) conference, vol. 4623, January 2002, pp. 77-84.

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

[10] 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), June 2007, pp. 1-4.

[11] Achuthsankar S. Nair and Sreenadhan S. Pillai, "A coding measure scheme employing electron-ion interaction pseudo potential (EIIP)," Bioinformation, vol. 1, pp. 197-202, October 2006.

[12] 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), vol. 6694, August 2007, pp. 669417-1 to 669417-10.

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

[14] R. Zhang and C. T. Zhang, "Z curves, An Intuitive Tool, for Visualizing and Analyzing the DNA sequences," J. BioMol. Struct. Dyn., vol. 11, pp. 767-782, 1994.

[15] Jing Zhao, Xiu Wen Yang, Jiag Ping Li, and Yuan Yan Tang, "DNA sequence classification based on wavelet packet analysis," in Proc. of the Second International Conf. on Wavelet Analysis and its Applications, Lecture Notes in Computer Science vol. 2251, January 2001. pp. 424-429.

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

[17] 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. vol. 2 , pp. 15-66, 2005.

[18] I. Cosic, "Macromolecular Bioactivity: Is it resonant interaction between macromolecules? Theory and Applications," IEEE Transactions on Biomedical Engg., vol. 41, pp. 1101-1114, December 1994.

[19] C. K. Peng, S.V. Buldyrev, S. Havlin, M. Simmons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," Physical Review E, vol. 49, pp. 1685-1689, February 1994.

[20] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," Computer Applications in the Biosciences (CABIOS), vol. 13, pp. 263-270, June 1997.

[21] J. A. Berger, S. K. Mitra, and J. Astola, "Power spectrum analysis for DNA sequences," in Proc. of Seventh International Symposium on Signal Processing and its Applications, vol. 2, July 2003, pp. 29-32.

[22] C. T. Zhang, J. Wang, and R. Zhang, "A novel method to calculate the G+C Content of Genomic DNA sequences," J. BioMol. Struct. Dyn. vol. 19, pp. 333-341, October 2001.

[23] C. T. Zhang and R. Zhang, "An isochore map of the human genome based on the Z curve method," Gene, vol. 317, pp. 127-135, October 2003.

[24] R. Zhang and C. T. Zhang, "Identification of replication origins in archaeal genomes based on the Z-curve method," Archaea, vol. 1, pp. 335-346, May 2005.

[25] B. Y. M. Kwan, J. Y. Y. Kwan, H. K. Kwan, R. Atwal, and O. T. Shen, "Wavelet analysis of the genome of the model plant *Arabidopsis thaliana*, " in Proc. of TENCON, Hong Kong, China, Nov. 14-17, 2006, pp. 1-4.

[26] H. K. Kwan, R. Atwal, and B. Y. M. Kwan, "Wavelet analysis of DNA sequences," in Proc. of Int. Conf. on Communications, Circuits and Sytems, Xiamen, China, May 25-27, 2008, pp. 917-921.