Exons and Introns Classification in Human and Other Organisms

Benjamin Y. M. Kwan, Jennifer Y. Y. Kwan, and Hon Keung Kwan

***Abstract*—**In the paper, the relative performances on spectral classification of short exon and intron sequences of the *human* and eleven model organisms is studied. In the simulations, all combinations of sixteen one-sequence numerical representations, four threshold values, and four window lengths are considered. Sequences of 150-base length are chosen and for each organism, a total of 16,000 sequences are used for training and testing. Results indicate that an appropriate combination of one-sequence numerical representation, threshold value, and window length is essential for arriving at top spectral classification results. For fixed-length sequences, the precisions on exon and intron classification obtained for different organisms are not the same because of their genomic differences. In general, precision increases as sequence length increases.

***Keywords*—**Exons and introns classification, *Human* genome, Model organism genome, Spectral analysis

I. Introduction

ANALYSIS of DNA sequences requires the 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 of special regions of interest. In the numerical representation of DNA sequences, each nucleotide of a DNA sequence is converted to a numerical value through a mapping function which enables numerical analysis using digital signal processing (DSP) techniques to facilitate identification of hidden periodicities and features, and revealing genome structures [1]. Genome annotation is a process of identifying the locations of the coding regions and genes in a genome and determining their functions. Genome sequencing generates DNA sequences, which in their raw form has no annotation [2]. Some methods focus on sequence similarity or motif matching to known genes in genome annotation. There is a need for other complementary or even more effective approaches to determine if a DNA sequence has a potential to harbour genes. It is known that exons (or coding regions) are rich in nucleotides C and G whereas introns (or noncoding regions) are rich in nucleotides A and T; and that protein coding regions of DNA sequences exhibit a period-3 property which is likely resulted from the three-base-length of codons used to generate amino acids. This period-3 property is relatively less apparent in sequences other than exons and could therefore be used to detect exons, and to distinguish exon regions from intron regions in genome annotation [1]. Consequently, identification of the period-3 regions of a DNA sequence helps predict possible gene locations. In general, classification of short exon and intron sequences is more challenging than that of longer sequences. In this paper, one-sequence numerical representations, thresholding, and windowing are applied to evaluate their performances in classifying short exon and intron sequences of the genomes of the human and other organisms based on their computed discrete Fourier transform (DFT) period-3 values.

This paper is organized as follows: Section II describes sixteen numerical representations used in this paper. In Section III, the DFT-based period-3 value of a numerically represented DNA sequence is derived and four threshold values are specified for classifying exon and intron sequences. In Section IV, the database and parameters used in simulations and the classification results are described. Finally, conclusions are given in Section V.

B. Y. M. Kwan is with the Faculty of Medicine, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, Canada K1H 8M5 (e-mail: bkwan066@uottawa.ca).

J. Y. Y. Kwan is with the School of Medicine, Queen's University, 80 Barrie Street, Kingston, Ontario, Canada K7L 3N6 (e-mail: kwan.j@queensu.ca).

H. K. Kwan is with the Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Windsor, Ontario, Canada N9B 3P4 (e-mail: kwan1@uwindsor.ca).

II. Numerical Representation

There are a variety of numerical representations of DNA sequences. In this paper, the focus is on simple and direct numerical representations which possess the following characteristics: (a) single sequence and compact mapping; (b) fixed magnitude mapping for each nucleotide; and (c) accessibility to DSP analysis. A list of sixteen one-sequence numerical representations obtained from [1] satisfying the above characteristics is shown in Table I.

The Integer Number representation [3] can be obtained by mapping numerals {1, 3, 2, 0} respectively to the four nucleotides as C = 1, G = 3, A = 2, T = 0. The Single Galois Indicator representation maps the CGAT nucleotides to a Galois field of four GF(4) [4] which is formed by assigning the numerical values to the nucleotides C = 1, G = 3, A = 0, T = 2 in a DNA sequence. This representation suggests that C < G and A < T. In the Paired Nucleotide Atomic Number representation [5], the paired nucleotides are assigned with the atomic numbers G, A = 62 and C, T = 42 respectively. In the Atomic Number representation [5], a numerical sequence is formed by assigning the atomic number of each nucleotide as C = 58, G = 78, A = 70, T = 66 in a DNA sequence. The Molecular Mass representation [6] of a DNA sequence is

formed by mapping the four nucleotides to their molecular masses as C = 110, G = 150, A = 134, T = 125, respectively.

The Electron-Ion Interaction Pseudo-potential (EIIP) represents the distribution of the free electrons' energies along a DNA sequence. In the EIIP representation, a single EIIP indicator sequence [7] is formed by substituting the EIIP of the nucleotides as C = 0.1340, G = 0.0806, A = 0.1260, T = 0.1335 in a DNA sequence. In the Paired Numeric representation [4], nucleotides are paired in a complementary manner and values of -1 and +1 are used to denote, respectively, C-G and A-T nucleotide pairs. In the Real Number representation [8], the nucleotide mappings are C = 0.5, G = -0.5, A = -1.5, T = 1.5, in which each of the C-G and A-T pairs bears complementary property. The Complex Number representation [9] reflects the complementary nature of C-G and A-T pairs by mapping nucleotides as C = -1-*j*, G = -1+*j*, A = 1+*j*, T = 1-*j* in which each of the C-G and A-T pairs is symmetrical with respect to the real axis. Seven other numerical representations introduced in [1] are listed as Codes 10-16 in Table I.

III. Spectral Analysis

Given a numerical represented DNA sequence, $X[n]$ for n=1 to N, its finite-length DFT sequence, $X[k]$ for k = 1 to N, is defined by

$$X[k]=\frac{1}{\sqrt{N}}\sum_{n=1}^{N}X[n]W_N^{(k-1)(n-1)} \text{ for } 1\le k\le N,\ W_N=e^{\frac{-j2\pi}{N}} \quad (1)$$

Using the windowing approach with a rectangular window length of L bases and a right-shift 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

$$X_T[k]=\frac{1}{N_w}\sum_{m=1}^{N_w}X_m[k] \quad (2)$$

The spectral content measure can be obtained by taking the normalized power spectrum of (2) as

$$S_n[k]=\frac{N}{L}\left|X_T[k]\right|^2 \quad (3)$$

The finite-length DFT of a numerical represented DNA sequence exhibits a peak at the frequency k = N/3 (corresponding to $2\pi/3$ in the DFT frequency range) called the period-3 property. Therefore, the spectral content measure can be used to detect and identify the period-3 value (P_3) in the spectral domain of a numerical representation as

$$P_3=S_n[N/3+1] \quad (4)$$

The statistics of the period-3 values determined from a training set of exon sequences and intron sequences can be used to classify a given 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, we define [1] a mid threshold value (T_m) and a proportional threshold value (T_p) as

$$T_m=\frac{(meanP_{3i}+meanP_{3e})+(stdP_{3i}-stdP_{3e})}{2} \quad (5)$$

$$T_p=\frac{sdP_{3e}*meanP_{3i}+sdP_{3i}*meanP_{3e}}{sdP_{3e}+sdP_{3i}} \quad (6)$$

Besides (5)-(6), the cross-over point of the cumulative distribution of all the exon period-3 values, $F(P_{3e})$, and the complementary cumulative distribution of all the intron period-3 values, $F_c(P_{3i})$, of a set of exon and intron training sequences [10] can be used to determine a threshold value. We define [1] a cumulative distribution threshold value (T_c) as

$$T_c=\text{Period-3 value at minimum } \left|F(P_{3e})-F_c(P_{3i})\right| \quad (7)$$

In [10], a fixed threshold value T_4 of 4 has been proposed which will also be used in the present study. For each of the above four cases, if a test sequence has a period-3 value P_{3t} greater than or equal to each respective threshold value (T_m, T_p, T_c, T_4), the test sequence is classified as an exon sequence; otherwise it is classified as an intron sequence.

IV. Simulations and Results

The genomes of the *human* and eleven model organisms downloaded from the UCSC Assembly [11]-[14] were used in the simulations. The downloaded genome of each organism consists of different numbers of exon and intron sequences with values summarized in Table II. For each organism, 6000 exon sequences and 6000 intron sequences were used for training and 2000 exon sequences and 2000 intron sequences were used for testing. Therefore, a total of 8000 exon sequences and 8000 intron sequences were used for training and testing. To evaluate the relative performances of each combination of the Codes 1-16, the four threshold values (T_m, T_p, T_c, T_4), and four values of window length (WL) L equal to 9, 15, 24, and 150 bases, each genome of the twelve organisms was trained and tested with identical sequence length (SL) of 150 bases and a right-shift window length of L-3 bases between two adjacent windows. The precision defined in (8) is used to measure the classification performance.

$$precision = \frac{\text{exon classification+intron classification}}{\text{exon number+intron number}} \times 100 \quad (8)$$

In the numerator of (8), the exon (or intron) classification denotes the number of correct exon (or intron) classification. For the twelve organisms, Fig. 1 plots precision (in percentage; Fig. 1, top) and the corresponding code index (from 1 to 16; Fig. 1, bottom). Fig. 2 plots WL index (from 1 to 4; Fig. 2, top), threshold value (Fig. 2, middle), and threshold index (from 1 to 4; Fig. 2, bottom) of top classifications obtained. In each of the two sub-plots of Fig. 1, the four color-bars of each organism represent four precision values (top sub-plot) and four code indexes (from 1 to 16; bottom sub-plot) correspond respectively to the four thresholds T_m (dark blue), T_p (light blue), T_c (yellow), and T_4 (red). Same notations apply to the top sub-plot of WL index (from 1 to 4) and the middle sub-plot (four threshold values) in Fig. 2. The bottom sub-plot of Fig. 2 represents the threshold index (from 1 to 4) of the top classification for each organism. The top classification results shown in Figs. 1-2 can be summarized in Table III which indicate that the Code 13 (the K-Quaternary Code I) achieves top classifications in 10 organisms and ranks second for classification performances in the organisms 4 and 11. A window length of 150 bases appears to be an appropriate choice for short sequences of length 150 bases. The threshold values T_p and T_c exhibit close performances and can often yield top classifications. The precision performances obtained are attractive in view of only relatively short 150-base sequences were used.

V. Conclusion

In this paper, the relative performances of the Codes 1-16 and the four threshold values, and the effect of four window lengths on spectral classification of short exon and intron sequences of twelve organisms have been presented. For short sequences of 150-base length, the simulations have shown that top spectral classification results can often be obtained using a combination of the Code 13 (the K-Quaternary Code I), the threshold value T_p (or T_c), and a window length of 150 bases. In general, classification precision increases as sequence length increases. The work described in this paper offers spectral information to aid identification of potential gene regions which could enhance the effectiveness of some annotation programs such as JIGSAW [15]. Another potential use of the work is in mapping reads (short DNA sequences) generated by next generation sequencing in which enormous short DNA sequences are generated. By being able to classify a read as either being a potential coding region or noncoding region, one can then narrow down possible genomic regions a sequence may belong to and thus aid genomic mapping [16].

References

[1] H. K. Kwan, B. Y. M. Kwan, and J. Y. Y. Kwan, "Novel methodologies for spectral classification of exon and intron sequences," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, 2011 (in press).

[2] R. A. Dalloul, J. A. Long, A. V. Zimin, et al. "Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis", *PLoS Biol*ogy, vol. 8, pii: e1000475, 2010.

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

[4] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal processing in sequence analysis: Advances in eukaryotic gene prediction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, pp. 310-321, June 2008.

[5] T. 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 *Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE)*, vol. 6694, August 2007, pp. 669417-1 to 669417-10.

[6] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S, Havlin, S. M. Ossadnik, C.-K. Peng, and M. Simmons, "Statistical mechanics in biology: How ubiquitous are long-range correlations?" *Physica A*, vol. 205, pp. 214-253, April 1994.

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

[8] 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.

[9] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *Journal of Cellular and Molecular Medicine*, vol. 6, pp. 279-303, April-June 2002.

[10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Bioinformatics* (CABIOS), vol. 13, issue 3, pp. 263-270, 1997.

[11] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, "The UCSC Table Browser data retrieval tool," *Nucleic Acids Research*, vol. 32 (Database issue), pp. D493-496, 1 January 2004.

[12] J. Goecks, A. Nekrutenko, J. Taylor, and The Galaxy Team, "Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, issue 8, article R86, 25 August 2010.

[13] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," *Current Protocols in Molecular Biology*, chapter 19, unit 19.10.1-21, January 2010.

[14] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: A platform for interactive large-scale genome analysis," *Genome Research*, vol. 15, issue 10, pp. 1451-1455, 15 October 2005.

[15] J. E. Allen and S. L. Salzberg, "JIGSAW: Integration of multiple sources of evidence for gene prediction," *Bioinformatics*, vol. 21, no. 18, pp. 3596-603, 2005.

[16] H. Jiang and W. H. Wong, "SeqMap: Mapping massive amount of oligonucleotides to the genome," *Bioinformatics,* vol. 24, no. 20, pp. 2395-2396, 2008.

TABLE I
LIST OF SIXTEEN NUMERICAL REPRESENTATIONS [1]

	Name	C	G	A	T
1	Integer Number	1	3	2	0
2	Single Galois Indicator	1	3	0	2
3	Paired Nucleotide Atomic Number	42	62	62	42
4	Atomic Number	58	78	70	66
5	Molecular Mass	110	150	134	125
6	EIIP	0.1340	0.0806	0.1260	0.1335
7	Paired Numeric	-1	-1	1	1
8	Real Number	0.5	-0.5	-1.5	1.5
9	Complex Number	-1-*j*	-1+*j*	1+*j*	1-*j*
10	K-Twin-Pair Code	-1	-1	*j*	*j*
11	K-Bipolar-Pair Code I	-1	1	*j*	-*j*
12	K-Bipolar-Pair Code II	-1	1	-*j*	*j*
13	K-Quaternary Code I	-1	-*j*	1	*j*
14	K-Quaternary Code II	-1	-*j*	*j*	1
15	K-Quaternary Code III	-*j*	-1	1	*j*
16	K-Quaternary Code IV	-*j*	-1	*j*	1

TABLE II
UCSC GENOMES OF 12 ORGANISMS
(OG: ORGANISM; NUMBER: NUMBER OF SEQUENCES)

OG	Clade	Genome	Type	Number
1	Mammal	*Human*	Exon	195133
			Intron	91529
2	Mammal	*Gorilla*	Exon	91756
			Intron	92315
3	Mammal	*Panda*	Exon	71833
			Intron	134835
4	Vertebrate	*Lizard*	Exon	61145
			Intron	105229
5	Vertebrate	*Tetraodon*	Exon	84691
			Intron	106455
6	Vertebrate	*X. tropicalis*	Exon	23631
			Intron	64458
7	Insect	*A. gambiae*	Exon	37304
			Intron	16599
8	Insect	*D. sechellia*	Exon	239418
			Intron	140870
9	Insect	*D. yakuba*	Exon	229293
			Intron	11669
10	Nematode	*C. brenneri*	Exon	19020
			Intron	23511
11	Nematode	*C. briggsae*	Exon	160061
			Intron	78282
12	Nematode	*P. pacificus*	Exon	32824
			Intron	212043

TABLE III
TOP CLASSIFICATIONS OF 12 ORGANISMS
(OG: ORGANISM; WL: WINDOW LENGTH IN BASES)

OG	Code	WL	Threshold	Precision (%)
1	13	150	T_p	75.8500
2	13	150	T_p	75.1512
3	13	150	T_c	78.5534
4	6	15	T_p	72.7888
4	13	150	T_c	72.6134
5	13	150	T_m	82.8405
6	13	150	T_p	74.0759
7	13	150	T_c	73.0671
8	13	150	T_c	78.0427
9	13	24	T_p	78.4008
10	13	9	T_c	69.7872
11	1	150	T_c	69.7966
11	13	150	T_m	68.8677
12	13	150	T_c	80.8291

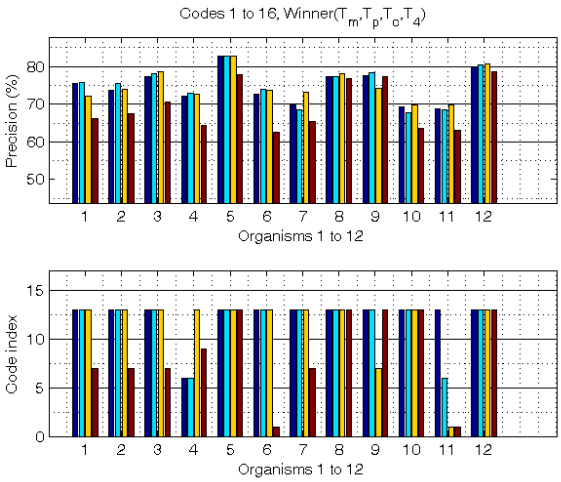

Fig. 1 Precision (top) and code index (bottom) of top classifications of 12 organisms

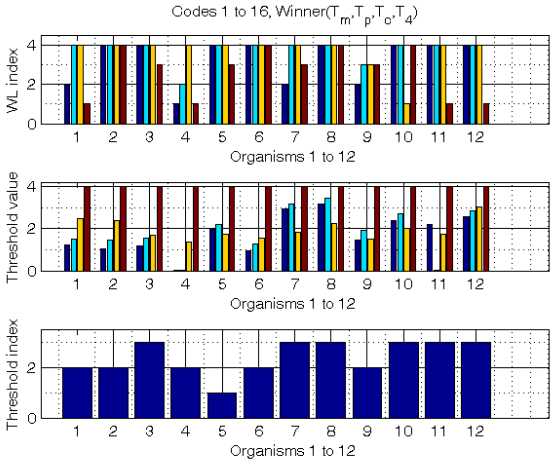

Fig. 2 WL index (top), threshold value (middle), and threshold index (bottom) of top classifications of 12 organisms