EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Novel methodologies for spectral classification of exon and intron sequences

Hon Keung Kwan[1*], Benjamin Y M Kwan[2] and Jennifer Y Y Kwan[3]

Abstract

Digital processing of a nucleotide sequence requires it to be mapped to a numerical sequence in which the choice of nucleotide to numeric mapping affects how well its biological properties can be preserved and reflected from nucleotide domain to numerical domain. Digital spectral analysis of nucleotide sequences unfolds a period-3 power spectral value which is more prominent in an exon sequence as compared to that of an intron sequence. The success of a period-3 based exon and intron classification depends on the choice of a threshold value. The main purposes of this article are to introduce novel codes for 1-sequence numerical representations for spectral analysis and compare them to existing codes to determine appropriate representation, and to introduce novel thresholding methods for more accurate period-3 based exon and intron classification of an unknown sequence. The main findings of this study are summarized as follows: Among sixteen 1-sequence numerical representations, the K-Quaternary Code I offers an attractive performance. A windowed 1-sequence numerical representation (with window length of 9, 15, and 24 bases) offers a possible speed gain over non-windowed 4-sequence Voss representation which increases as sequence length increases. A winner threshold value (chosen from the best among two defined threshold values and one other threshold value) offers a top precision for classifying an unknown sequence of specified fixed lengths. An interpolated winner threshold value applicable to an unknown and arbitrary length sequence can be estimated from the winner threshold values of fixed length sequences with a comparable performance. In general, precision increases as sequence length increases. The study contributes an effective spectral analysis of nucleotide sequences to better reveal embedded properties, and has potential applications in improved genome annotation.

Keywords: DNA sequence, numerical representation, nucleotide to numeric mapping, exon and intron sequences, coding and non-coding sequences, threshold value, thresholding, exon and intron classification, period-3, spectral analysis, discrete Fourier transform, gene detection, genome annotation

1. Introduction

It is known that the coding regions of a nucleotide sequence exhibit a period-3 spectral property because of the presence of codon structure [1], and this property can be used to identify regions of interest. A nucleotide sequence is a discrete sequence consisting of nucleotides C, G, A, and T in which digital spectral analysis can be used to reveal its hidden periodicities, spectral features, and genome structure. Digital spectral analysis is usually carried out by the discrete Fourier transform (DFT) which is a digital signal processing (DSP) technique. Genomic signal processing (GSP) involves the processing and analysis of a digital signal in the form of a numerical sequence mapped from a nucleotide sequence [1,2]. A general description on the biological aspects of this article can be found in [1,2]. To perform digital spectral analysis, each nucleotide of a nucleotide sequence has to be converted to a numerical value through a mapping. Such a mapping is called numerical representation and its choice affects how well the biological properties of a nucleotide sequence can be revealed in numerical domain. A nucleotide sequence can be numerically represented in the form of R-sequence, a survey on mappings with $R \geq 1$ is given in [3,4]. 1-sequence numerical representation (with $R = 1$) is the most compact form of mapping in which one nucleotide

* Correspondence: kwan1@uwindsor.ca
[1]Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada
Full list of author information is available at the end of the article

is mapped to one fixed numerical value to form a single sequence. 1-sequence numerical representations and their relative performances are studied in this article for unknown exon and intron sequence classification with an improved accuracy.

In this article, nine 1-sequence numerical representations (Codes 1-9) [5-16] are identified though a literature search, which can be grouped under positive-integer-value, real-value, and complex-value numerical representations. In addition, seven 1-sequence complex-value numerical representations (Codes 10-16) are derived. In this article, all these sixteen numerical representations and the Voss representation [17] are compared based on the genomes of twelve organisms (including the human).

Genome annotation involves a process of identifying the locations of the coding regions (and associated genes) in a genome. Coding regions exhibit the period-3 property in the spectral domain which is less apparent in sequences other than exon sequences, and can therefore be used to detect exon sequences and to distinguish exon regions from intron regions in genomic sequences. The performance depends on an appropriate choice of a period-3 threshold value to classify between an exon sequence and an intron sequence. In [17,18], the cumulative distributions of coding and non-coding sequences are used to determine a period-3 threshold value. In this article, two threshold value determination methods are defined. Given a set of exon and intron sequences of a fixed length, all the three thresholding values are used to determine the winner threshold value applicable to any nucleotide sequence of the same length. Also, given a number of exon and intron sequence sets of different lengths, the winner threshold values can be interpolated to yield an interpolated winner threshold value applicable to an unknown nucleotide sequence of an arbitrary length. Some commonly used symbols (or abbreviations) are explained in Table 1.

2. Methods and results

2.1. Numerical representation

There are a number of numerical representations for nucleotide sequences. In this article, we shall focus on direct and simple numerical representations which satisfy the following requirements: (a) Single sequence mapping for a nucleotide sequence; (b) Fixed value mapping for each nucleotide; and (c) Accessible to digital signal processing analysis. Sixteen 1-sequence numerical representations (Codes 1-16) which satisfy these three requirements are defined in Table 2.

It is known that exons are rich in nucleotides C and G and introns are rich in nucleotides A and T [8,19]. For ease of viewing, C-G and A-T are paired when defining each of the sixteen numerical representations. The first group of numerical representations consists of five positive-integer-value numerical representations (as Codes 1-5 in Table 2) which shall be described as follows: The Integer Number representation [5,6] is obtained by mapping numerals {1, 3, 2, 0} respectively to the four nucleotides as C = 1, G = 3, A = 2, and T = 0. The Single Galois Indicator representation [7,8] maps the CGAT nucleotides to a Galois field of four, GF(4), which is formed by assigning numerical values to the nucleotides as C = 1, G = 3, A = 0, and T = 2 in a nucleotide sequence. This representation suggests that C < G and A < T. In the Paired Nucleotide Atomic Number representation [9], the paired nucleotides are assigned with atomic numbers as A, G = 62 and C, T = 42 respectively. In the Atomic Number representation [9], a numerical sequence is formed by assigning atomic numbers to each nucleotide as C = 58, G = 78, A = 70, and T = 66 in a nucleotide sequence. The Molecular Mass representation [10] of a nucleotide sequence is formed by mapping the four nucleotides to their molecular masses as C = 110, G = 150, A = 134, and T = 125, respectively.

The second group of numerical representations consists of three real-value numerical representations (as Codes 6-8 in Table 2) which can be described as: The Electron-Ion Interaction Pseudo-potential (EIIP) represents the distribution of the free electrons energies along a nucleotide sequence. In the EIIP representation, a single EIIP indicator sequence [11,12] is formed by substituting the EIIP of the nucleotides as C = 0.1340, G = 0.0806, A = 0.1260, and T = 0.1335 in a nucleotide sequence. In the Paired Numeric representation [6,8], nucleotides (C-G, A-T) are to be paired in a complementary manner and values of -1 and +1 are to be used to denote, respectively, C-G and A-T nucleotide pairs. In the Real Number representation [6,8,13], the nucleotide mappings are C = 0.5, G = -0.5, A = -1.5, and T = 1.5, which bears complementary property. The third group consists of one complex-value numerical representation (as Code 9 in Table 2) called the Complex Number representation [2,5,6,8,14-16], it reflects the complementary nature of C-G and A-T pairs by mapping nucleotides as C = -1-j, G = -1+j, A = 1+j, and T = 1-j.

In addition to the above nine 1-sequence numerical representations, seven 1-sequence numerical representations (listed as Codes 10-16 in Table 2) are derived in which each nucleotide of a sequence is mapped to either a single real-value element (± 1) or a single imaginary-value element (± j). Each of the Codes 10-16, namely, the K-Twin-Pair Code, the K-Bipolar-Pair Codes I and II, and the K-Quaternary Codes I-IV has its equivalent numerical representations. We define a numerical representation $R2$ to be an equivalent numerical

Table 1 List of symbols

Symbol	Description	Symbol	Description
C	Cytosine	G	Guanine
A	Adenine	T	Thymine
DNA	Deoxyribonucleic acid	OG	Organism
DSP	Digital signal processing	GSP	Genomic signal processing
DFT	Discrete Fourier transform	FFT	Fast Fourier transform
SL	Sequence length (bases)	WL	Window length (bases)
WRS	Window right-shift (bases)	NW	Number of windows
CDP3	Cumulative distribution period-3	CCDP3	Complementary CDP3
PSV	Power spectral value	P_3	Period-3 power spectral value
T_m	Mid threshold value	T_p	Proportional threshold value
T_c	Cumulative distribution threshold value	T_4	Fixed threshold value (= 4.0)
T_w	Winner threshold value	T_i	Interpolated winner threshold value
$meanP_{3e}$	Mean of the period-3 values obtained from specified exon sequences		
sdP_{3e}	Standard deviation of the period-3 values obtained from specified exon sequences		
$meanP_{3i}$	Mean of the period-3 values obtained from specified intron sequences		
sdP_{3i}	Standard deviation of the period-3 values obtained from specified intron sequences		

Note: $T_m(P_3)$, $T_p(P_3)$, $T_c(P_3)$, $T_w(P_3)$, $T_i(P_3)$ are threshold values computed from P_3.

representation of $R1$ if $R2$ gives rise to the same power spectrum as that of $R1$. An equivalent numerical representation can be obtained by multiplying the numerical represented value of each of the four bases [C, G, A, T] of a nucleotide sequence by the same constant which includes any of -1, and ± j. In particular, a complementary numerical representation obtained by inverting the signs of all the four bases [C, G, A, T] in a numerical representation is an equivalent numerical representation. Other forms of equivalent numerical representation exist, for example, the K-Twin-Pair Code [C, G, A, T] = [-1, -1, j, j] has an equivalent form [C, G, A, T] = [-j, -j, 1, 1]. The K-Quaternary Code III (Code 15) [C, G, A, T] = [-j, -1, 1, j] is identical to the pentanary code in [20] (with [20] specifies that $x[n] = 0$ for an unknown nucleotide at position n) which was derived for the three-dimensional DNA walk for graphical representation of nucleotide sequences. The K-Bipolar-Pair Code II (Code 12) [C, G, A, T] = [-1, 1, -j, j] has an equivalent numerical representation of [C,G,A,T] = [-j, j, 1, -1] adopted in [21]. The K-Quaternary Code IV (Code 16) [C, G, A, T] = [-j, -1, j, 1] has a complementary numerical representation of [C, G, A, T] = [j, 1, -j, -1] mentioned in [22]. In [23], complex numerical representations with simultaneous non-zero real part and non-zero imaginary part were adopted which are different from those of the Codes 10-16 in which either a real value or an imaginary value is used for each nucleotide to numeric mapping. It should be noted that the numerical representations in [21-23] were formulated for sequence alignment but not for period-3 power spectral analysis.

Table 2 List of sixteen numerical representation codes

	Name	Code			
		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

The Voss representation (i.e., the Code 17) is a commonly used numerical representation for period-3 spectral classification of exon and intron sequences [1,2,17,24,25]. The Voss representation is a 4-sequence representation (R = 4) in which each of the four nucleotides of a nucleotide sequence is represented by a separate numerical sequence (as C-sequence, G-sequence, A-sequence, and T-sequence) such that the n^{th} position of the C-sequence is coded by 1 if the n^{th} nucleotide of the sequence is C, otherwise it is coded by 0. In a similar manner, the coding procedure just described applies to the remaining three numerical sequences. A threshold value can be determined from the cumulative

distributions of the signal-to-noise ratio of the peak at f = 1/3 for each set of exon and intron sequences [17]. Using the Voss representation, two known threshold values (T_c, T_4) are adopted for comparison in this article. The T_c thresholding is determined from the exon and intron cumulative distribution functions [17] and the T_4 thresholding is set to a fixed value four [17]. The Codes 1-16 are to be compared to the Voss representation (Code 17) in Section 2.4.

2.2. Spectral analysis

The purpose of the spectral analysis of a numerically represented nucleotide sequence is to compute its period-3 spectral component located at a frequency equal to 2π/3 in the DFT spectrum. Given a numerical sequence, $x[l]$ for l = 1 to L, its finite-length DFT sequence, $X[k]$ for k = 1 to L, and its inverse DFT, $x[l]$ for l = 1 to L, are defined by

$$X[k] = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} x[l] W_L^{(k-1)(l-1)} \text{ for } 1 \le k \le L \tag{1}$$

$$x[l] = \frac{1}{\sqrt{L}} \sum_{k=1}^{L} X[k] W_L^{-(k-1)(l-1)} \text{ for } 1 \le l \le L \tag{2}$$

$$W_L = e^{\frac{-j2\pi}{L}} \tag{3}$$

In this article, a rectangular windowing of length N bases is used and each consecutive window is right-shifted by three bases with an overlap window length of N-6 bases between two adjacent windows. The rectangular windowing is adopted to avoid distortion on the cumulative DFT spectrum. We define the mean cumulative DFT spectrum, $X_T[k]$, of all the windowed sequences, $X_m[k]$ for m = 1 to M, as

$$X_T[k] = \frac{1}{M} \sum_{m=1}^{M} X_m[k] \text{ for } 1 \le k \le N \tag{4}$$

The power spectrum $S[k]$ is obtained from Equation 4 as

$$S[k] = \left| X_T[k] \right|^2 \text{ for } 1 \le k \le N \tag{5}$$

From Equation 5, we define the normalized power spectrum $S_n[k]$ of a numerically represented nucleotide sequence of length L as

$$S_n[k] = \frac{L}{N} S[k] \text{ for } 1 \le k \le N \tag{6}$$

The period-3 power spectral value (or period-3 value) P_3 can be obtained from the normalized power spectrum of a numerically represented sequence at k = N/3+1 as

$$P_3 = S_n[\frac{N}{3} + 1] \tag{7}$$

In all computer simulations, for a window length of WL equals to a sequence length of SL, one complete DFT analysis is adopted. For WL less than SL, each consecutive window is right-shifted by three bases with an overlap window length of "WL-6" bases between two adjacent windows. The period-3 values P_3 computed by Equation 7 for 4583 exon sequences and that for 4583 intron sequences (both of sequence length 1250 bases, window length 24 bases, and window right-shift 3 bases) obtained by the K-Quaternary Code I are plotted in Figures 1 and 2. From these figures, it can be observed that the period-3 values of exon sequences are higher than those of the intron sequences. This is an important property which will be used to classify exon sequences from intron sequences.

The $\log_{10}$ power spectrum $S_n[k]$ for $1 \le k \le N$ of an exon sequence and that of an intron sequence (both of sequence length 1250 bases, window length 24 bases, and window right-shift 3 bases) obtained by the K-Quaternary Code I are plotted in Figures 3 and 4. As shown in Figure 3 (or Figure 4), the exon (or intron) period-3 value at 2π/3 defined by Equation 7 has in fact been able to reveal a prominent power spectral value; also, the exon period-3 value is higher than the intron period-3 value. The DC power spectral value is usually the highest. Besides period-3 and DC power spectral value, power spectral values of other periodicities are clearly shown in Figures 3 and 4.

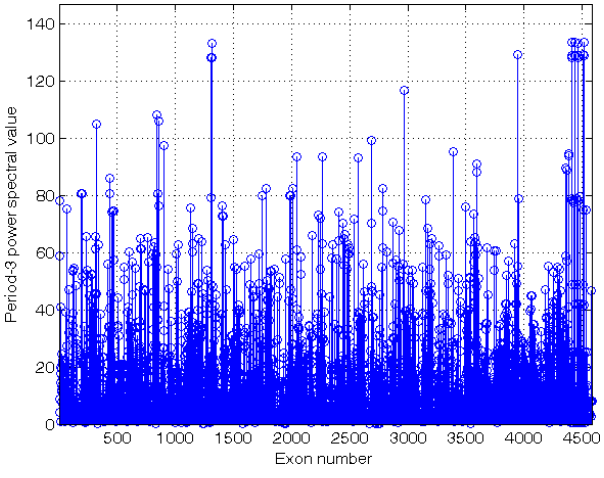

Figure 1 Period-3 power spectral values of 4583 exon sequences using K-Quaternary Code I. SL = 1250 bases, WL = 24 bases, and WRS = 3 bases.

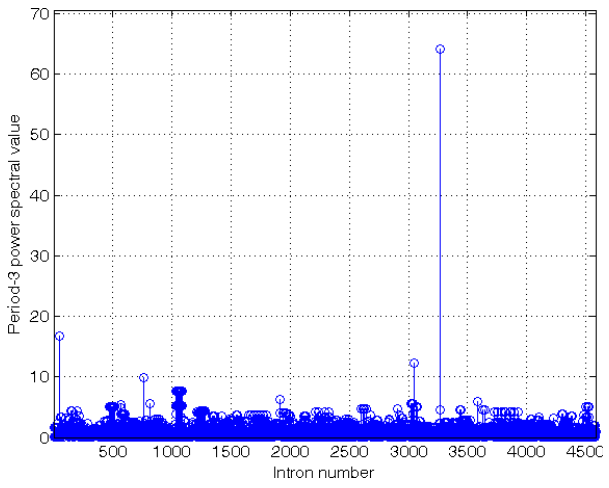

Figure 2 Period-3 power spectral values of 4583 intron sequences using K-Quaternary Code I. SL = 1250 bases, WL = 24 bases, and WRS = 3 bases.

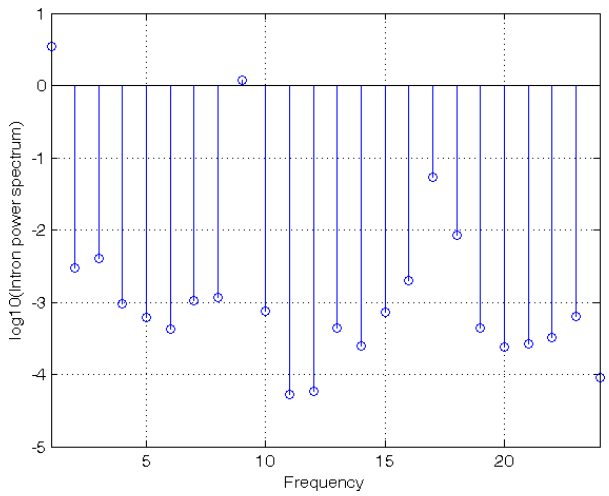

Figure 4 Log_{10}(Intron power spectrum) $S[k]$ for k = 1 to 24 of intron sequence # 4583 using K-Quaternary Code I. SL = 1250 bases, WL = 24 bases, and WRS = 3 bases.

2.3. Winner threshold value

In this section, the statistics of the period-3 values computed from a training set of exon sequences and intron sequences are used to define the threshold values of two novel thresholding methods to classify an untrained sequence to be either an exon sequence or an intron sequence. As explained in Table 1, $meanP_{3e}$ 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 two threshold values called the mid threshold value, T_m, and the proportional threshold value, T_p, as

$$T_m = \frac{(meanP_{3i} + meanP_{3e}) + (stdP_{3i} - stdP_{3e})}{2} \quad (8)$$

$$T_p = \frac{sdP_{3e} \times meanP_{3i} + sdP_{3i} \times meanP_{3e}}{sdP_{3e} + sdP_{3i}} \quad (9)$$

In both definitions, the exon cluster is centred at $meanP_{3e}$ and the intron cluster is centred at $meanP_{3i}$. In Equation 8, the mid threshold value is determined by the mid-point between the exon cluster and the intron cluster, whereas in Equation 9, the proportional threshold value is determined in proportion to the standard deviations of the two clusters.

Besides the above two methods for determining a threshold value, 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 can be used to determine a threshold value. We define such a cumulative distribution threshold value, T_c, as

$$T_c = \text{Period - 3 value at minimum } \left|F(P_{3e}) - F_c(P_{3i})\right| \quad (10)$$

Figure 5 shows a plot of $F(P_{3e})$ and $F_c(P_{3i})$ versus corresponding exon/intron period-3 power spectral value. Each of the three threshold values T_m, T_p, and T_c can be used to classify an unknown nucleotide sequence. For illustration, let T_t be any of T_m, T_p, and T_c, if an unknown nucleotide sequence has a computed period-3 power spectral value greater than or equal T_t, the

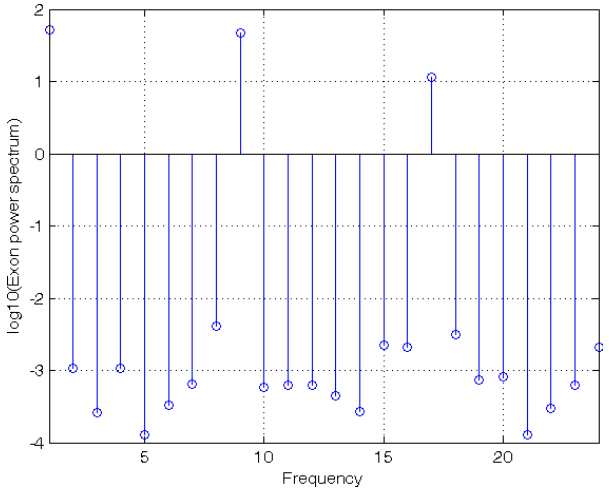

Figure 3 Log_{10}(Exon power spectrum) $S[k]$ for k = 1 to 24 of exon sequence # 4583 using K-Quaternary Code I. SL = 1250 bases, WL = 24 bases, and WRS = 3 bases.

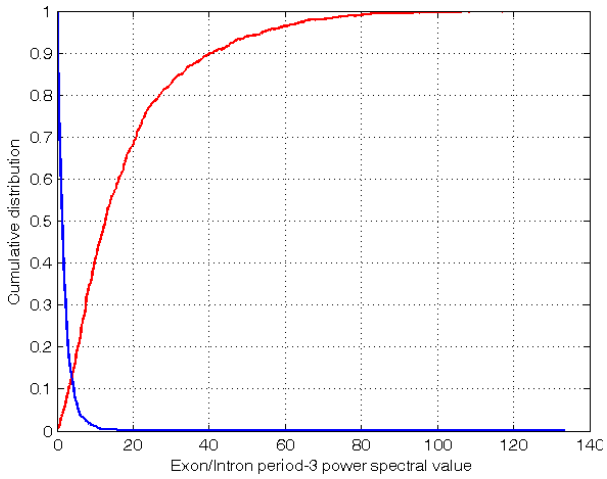

Figure 5 Cumulative distribution threshold value (T_c) = 3.881141 at cross-over point of exon CDP3 PSV (red) and intron CCDP3 PSV (blue) using K-Quaternary Code I. Exon/intron sequence number = 3437, SL = 1250 bases, WL = 24 bases, and WRS = 3 bases.

unknown nucleotide sequence is classified as an exon sequence; otherwise it is classified as an intron sequence.

The performance is measured objectively in terms of the exon classification (%) which is the percentage of correct classification of exon sequences, the intron classification (%) which is the percentage of correct classification of intron sequences, and the precision (%). These three terms are defined in Equations 11-13 as

$$\text{exon classification} = \frac{\text{number of correct exon classification}}{\text{exon number}} \times 100\% \quad (11)$$

$$\text{intron classification} = \frac{\text{number of correct intron classification}}{\text{intron number}} \times 100\% \quad (12)$$

$$\text{precision} = \frac{\text{number of correct exon classification} + \text{number of correct intron classification}}{\text{exon number} + \text{intron number}} \times 100\% \quad (13)$$

We define a winner threshold value, T_w, as the threshold value chosen among T_m, T_p, and T_c that yields the top classification (or the highest precision).

2.4. Classification and speed performances of codes 1-17

In this article, all the exon and intron sequences of the human and eleven model organisms were downloaded from the UCSC Genomes [26-29] as inputs to be processed by Matlab programs.

2.4.1. Classification performance of short sequences

The short sequence group I listed in Table 3 downloaded from the UCSC Human genome [26-29] consists of the thirteen short sequence sets ranging from 50 bases to 650 bases (at intervals of 50 bases) as [50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650]. For each of the thirteen short sequence sets, the training sequence numbers are 1 to 10005 and the testing sequence numbers are 10006 to 13340 with a train/test ratio of three for both exon and intron sequences within a set. Each of the thirteen sequence sets is trained and tested using a combination of sixteen numerical representations, four window lengths, and four threshold values (T_m, T_p, T_c, and T_4). Figures 6(a-b) and 7 (a-c) plots the top classifications obtained using each of the thirteen untrained sub-sets of 3335 exon sequences and 3335 intron sequences of the short sequence group I. Figure 6 (a, top) displays a total of 13 × 4 precision values obtained using the thirteen sequence sets in thirteen combined columns with each combined column consists of four sub-columns of precision values. Each of the four sub-column precision values shown in Figure 6 (a, top) corresponds to a top classification with its code index, window length index, threshold index, and threshold value displayed, respectively, in Figures 6 (b, bottom) and Figure 7 (a-c). In each of the five sub-plots in Figures 6 (a-b) and 7 (a-c), the first sub-column corresponds to the top classification obtained by the top

Table 3 UCSC Human genome consisting of 2 short sequence groups.

Short sequence group I			Short sequence group II		
SL	Type	Number	SL	Type	Number
50	Exon	542910	90	Exon	4705
	Intron	653640		Intron	1495
100	Exon	379835	120	Exon	4582
	Intron	621266		Intron	723
150	Exon	195133	180	Exon	2355
	Intron	588070		Intron	427
200	Exon	90211	210	Exon	1188
	Intron	567031		Intron	436
250	Exon	51626	270	Exon	612
	Intron	548406		Intron	271
300	Exon	35685	330	Exon	147
	Intron	532489		Intron	322
350	Exon	28497	390	Exon	122
	Intron	519074		Intron	242
400	Exon	24007	420	Exon	126
	Intron	506839		Intron	240
450	Exon	20957	480	Exon	65
	Intron	495007		Intron	263
500	Exon	18492	510	Exon	54
	Intron	483412		Intron	272
550	Exon	16605	570	Exon	75
	Intron	472689		Intron	214
600	Exon	14822	630	Exon	25
	Intron	462572		Intron	189
650	Exon	13340			
	Intron	452529			

Clade: Mammal. Genome: Human. Assembly: Feb. 2009 (GRCh37/hg19). Group: Genes and Gene Prediction Tracks. Track: UCSC Genes. Table: knownGene.

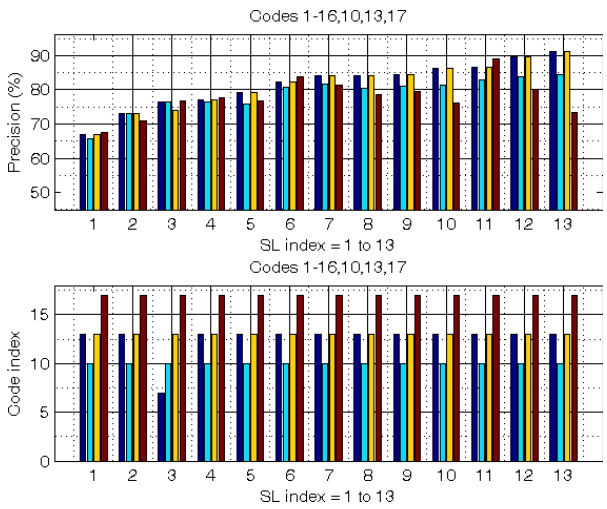

Figure 6 (a-b) Precision (a, top) and code index (b, bottom) of top classifications of 13 short sequence sets obtained by best of Codes 1-16 (dark blue), Code 10 (light blue), Code 13 (yellow), and Code 17 (red).

performer among the Codes 1 to 16, the second sub-column corresponds to the top classification obtained by the Code 10, the third sub-column corresponds to the top classification obtained by the Code 13, and the fourth sub-column corresponds to the top classification obtained by the Code 17. The Codes 10 and 13 are found to be among the top performers and therefore they are displayed in two separate sub-columns. As the Codes 1-16 are to be compared to the Code 17 (described in Section 2.1), therefore the Code 17 is also displayed as one sub-column.

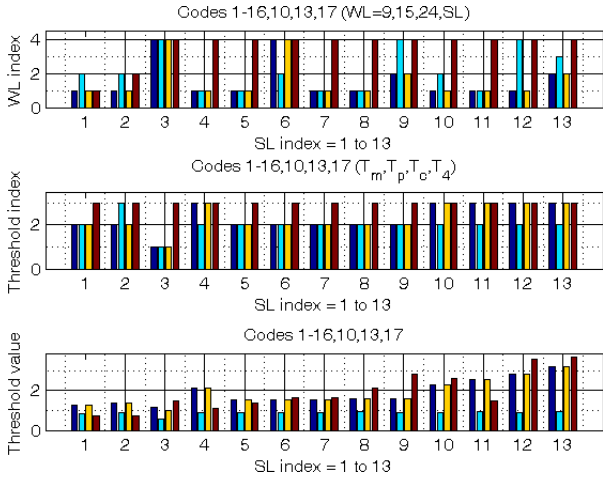

Figure 7 (a-c) WL index (a, top), threshold index (b, middle), and threshold values (c, bottom) of top classifications of 13 short sequence sets obtained by best of Codes 1-16 (dark blue), Code 10 (light blue), Code 13 (yellow), and Code 17 (red).

Table 4 Code (Figure 6b, bottom), WL (bases) (Figure 7a, top), threshold method (Figure 7b, middle), and precision (%) (Figure 6a, top) of top classifications of 13 short sequence sets

SL	Code	WL	Threshold	Precision
50	17	9	T_c	67.5412
50	13	9	T_p	-
100	13	9	T_p	73.2234
150	17	150	T_c	76.8066
150	10	150	T_m	-
200	17	200	T_c	77.8411
200	13	9	T_c	-
250	13	9	T_p	79.1004
300	17	300	T_c	83.8231
300	13	300	T_p	-
350	13	9	T_p	84.1829
400	13	9	T_p	84.2579
450	13	15	T_p	84.3628
500	13	9	T_c	86.2519
550	17	550	T_c	89.1604
550	13	9	T_c	-
600	13	9	T_c	89.7151
650	13	15	T_c	91.2744

The classification performance shown in Figures 6(a-b) and 7 (a-c) and summarized in Table 4 indicate that for the thirteen short sequence sets: (a) The top performer for 8 of these thirteen sequence sets is the Code 13, followed by the Code 17 as the top performer for the remaining 5 sequence sets (with the Code 13 ranks 2nd in 4 cases and the Code 10 ranks 2nd in 1 case, each of these 5 cases is also shown in Table 4 on the row following that of the top performer). (b) The window length of 9 bases yields 7 top classifications; the window length of 15 bases yields 2 top classifications; and for the window length equals to the sequence length, 4 top classifications are obtained. (c) The T_c thresholding yields 8 top classifications; followed by the T_p thresholding which yields 5 top classifications. In general, top classifications can often be achieved by the Code 13, using WL = 9 bases and T_p thresholding for SL = 50 to 450 bases; and using WL = 9 bases and T_c thresholding for SL = 500 to 650 bases.

2.4.2. Classification performance of long sequences

The long sequence group I listed in Table 5 downloaded from the UCSC Human genome [26-29] consists of the thirteen long sequence sets ranging from 650 bases to 1250 bases (at intervals of 50 bases) as [650, 700, 750, 800, 850, 900, 950, 1000, 1050, 1100, 1150, 1200, 1250]. For each of the thirteen long sequence sets, the training sequence numbers are 1 to 3437 and the testing sequence numbers are 3438 to 4583 with a train/test ratio of three for both exon and intron sequences within

Table 5 UCSC Human genome consisting of 2 long sequence groups

Long sequence group I			Long sequence group II		
SL	Type	Number	SL	Type	Number
650	Exon	13340	690	Exon	35
	Intron	452529		Intron	178
700	Exon	12103	720	Exon	25
	Intron	443645		Intron	205
750	Exon	11075	780	Exon	37
	Intron	434145		Intron	191
800	Exon	10132	810	Exon	26
	Intron	425427		Intron	125
850	Exon	9312	870	Exon	5
	Intron	417108		Intron	119
900	Exon	8587	930	Exon	42
	Intron	409022		Intron	177
950	Exon	7608	990	Exon	24
	Intron	401490		Intron	110
1000	Exon	6846	1020	Exon	14
	Intron	393725		Intron	217
1050	Exon	6340	1080	Exon	31
	Intron	385722		Intron	143
1100	Exon	5726	1110	Exon	12
	Intron	378374		Intron	160
1150	Exon	5276	1170	Exon	12
	Intron	371372		Intron	174
1200	Exon	4953	1230	Exon	14
	Intron	364598		Intron	106
1250	Exon	4583			
	Intron	357831			

Clade: Mammal. Genome: Human. Assembly: Feb. 2009 (GRCh37/hg19). Group: Genes and Gene Prediction Tracks. Track: UCSC Genes. Table: knownGene.

a set. Each of the thirteen sequence sets is trained and tested using a combination of sixteen numerical representations, four window lengths, and four threshold values (T_m, T_p, T_c, and T_4). The top classification results of Codes 1-16 are compared to those obtained by the Code 17 [17] (as described in Section 2.1). Figures 8(a-b) and 9 (a-c) plots the top classifications obtained using each of the thirteen untrained sub-sets consisting of 1146 exon sequences and 1146 intron sequences of the long sequence group I. The notations and explanations of Figures 6(a-b) and 7 (a-c) regarding the short sequence set I in Table 3 apply to Figures 8(a-b) and 9 (a-c) which display the corresponding results of the long sequence set I in Table 5.

The classification performance shown in Figures 8(a-b) and 9 (a-c) and summarized in Table 6 indicates that for the thirteen long sequence sets: (a) The top performer for the thirteen sequence sets is the Code 13, except for SL = 1100 bases in which the Code 1 (with WL = 24 bases and T_4 thresholding) is marginally higher than that of the Code 13 (with WL = 9 bases

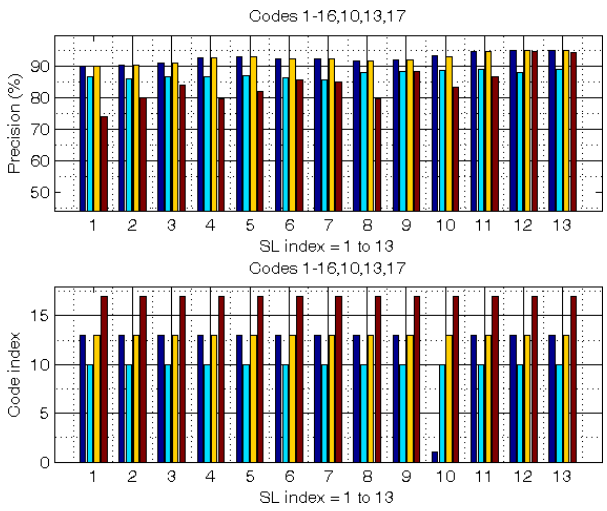

Figure 8 (a-b) Precision (a, top) and code index (b, bottom) of top classifications of 13 long sequence sets obtained by best of Codes 1-16 (dark blue), Code 10 (light blue), Code 13 (yellow), and Code 17 (red).

and T_p thresholding as shown in Table 6 on the row following that of the top performer). (b) The window length of 9 bases yields 5 top classifications; the window length of 15 bases yields 4 top classifications; and for window length equals to 24 bases, 3 top classifications are obtained. (c) The T_p thresholding yields 6 top classifications; followed by the T_4 thresholding which yields 4 top classifications; and the T_c thresholding which yields 3 top classifications. In general, top classifications are achievable often using a combination of the Code 13, WL = 9 bases, and T_p thresholding.

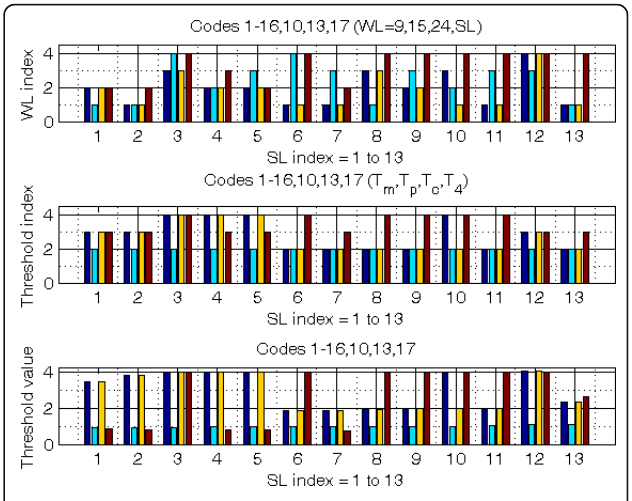

Figure 9 (a-c) WL index (a, top), threshold index (b, middle), and threshold value (c, bottom) of top classifications of 13 long sequence sets obtained by best of Codes 1-16 (dark blue), Code 10 (light blue), Code 13 (yellow), and Code 17 (red).

Table 6 Code (Figure 8b, bottom), WL (bases) (Figure 9a, top), threshold method (Figure 9b, middle), and precision (%) (Figure 8a, top) of top classifications of 13 long sequence sets

SL	Code	WL	Threshold	Precision
650	13	15	T_c	90.1396
700	13	9	T_c	90.4014
750	13	24	T_4	91.2304
800	13	15	T_4	92.8447
850	13	15	T_4	93.2810
900	13	9	T_p	92.4084
950	13	9	T_p	92.4520
1000	13	24	T_p	91.7103
1050	13	15	T_p	92.1902
1100	1	24	T_4	93.5428
1100	13	9	T_p	-
1150	13	9	T_p	94.6771
1200	13	1200	T_c	94.9825
1250	13	9	T_p	94.9825

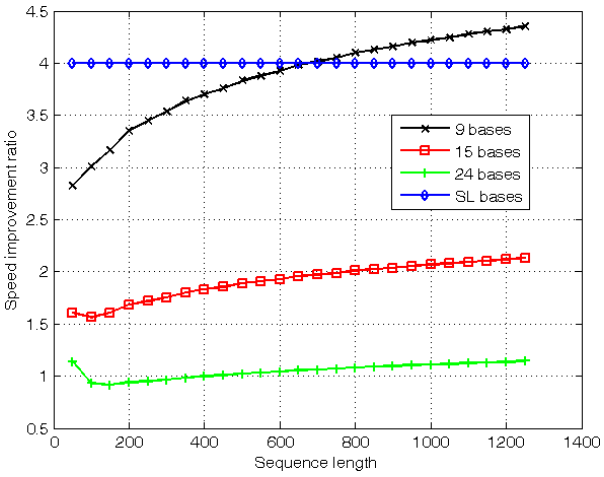

Figure 10 Speed improvement ratio of 1-sequence representation over 4-sequence representation for window length = 9, 15, 24, SL bases in FFT spectral analysis.

2.4.3. Speed performance

In this article, the DFT-based spectral analysis of a numerical sequence is computed using the FFT. In general, the computational complexity of a 4-sequence numerical representation (such as the Voss representation [17]) is always four times that of a 1-sequence numerical representation if either windowing or non-windowing is used in the comparison. The FFT computational complexity for a numerical sequence of length L is $L \times \log_2(L)$ and for a windowed sequence of WL = N is $N \times \log_2(N)$. As each consecutive window is right-shifted by three bases, the number of windows NW = 1 + fix((*L*-*N*)/3) where fix(X) rounds the elements of X to the nearest integer. For a non-windowed 4-sequence numerical representation of SL = L, the computational complexity is $4L \times \log_2(L)$ whereas for a windowed 1-sequence numerical representation of SL = L, the computational complexity is NW $\times$ $N \times \log_2(N)$. The ratio of $4L \times \log_2(L)$ over NW $\times$ $N \times \log_2(N)$ gives a speed improvement ratio. A plot of the speed improvement ratio against SL for WL = [9,15,24] bases is shown in Figure 10. As seen from Figure 10, it can be observed that a windowed 1-sequence representation with WL = [9,15,24] bases offers a possible speed advantage over a non-windowed 4-sequence representation and the speed improvements increases as SL increases. Also shown in Figure 10, for WL = SL, $N = L$, the speed improvement ratio is four.

2.5. Interpolated winner threshold value

With the use of the winner threshold value described in Section 2.3, a top precision for an unknown nucleotide sequence of length equals to any of the thirteen long or short sequence sets can be obtained. However, if the length of an unknown nucleotide sequence is neither any of the thirteen short sequence sets nor any of the thirteen long sequence set, an interpolated winner threshold value can be determined as described in this section. For illustration, let us consider the K-Quaternary Code I but the interpolated winner threshold methodology also applies to other numerical representations. For a set of exon and intron training sequences, the winner threshold value, $T_w(P_3)$, is first chosen among the three threshold values, $T_m(P_3)$, $T_p(P_3)$, and $T_c(P_3)$ that yields a top precision during testing. For each window length, the winner threshold value, $T_w(P_3)$, versus sequence length of the set of thirteen short sequences with SL = 50 to 650 bases is plotted in Figure 11 and similarly in Figure 12 for the set of thirteen long sequences with SL = 650 to 1250 bases. From Figures 11 and 12, given an unknown nucleotide sequence of arbitrary length (within the length of trained sequences), an interpolated winner threshold value, $T_i(P_3)$, at any of the four specified window lengths can be obtained. Using an interpolated winner threshold value, $T_i(P_3)$, if the unknown nucleotide sequence has a period-3 power spectral value greater than or equal $T_i(P_3)$, the unknown nucleotide sequence is classified as an exon sequence; otherwise it is classified as an intron sequence.

2.5.1. Short sequences

For each WL, an interpolated threshold value for an arbitrary sequence length can be estimated using cubic spline interpolation from the corresponding discrete winner threshold values shown in Figure 11. For each of the four WL, the interpolated threshold values and the corresponding precisions obtained on testing each of the twelve untrained sequence sets (under the short

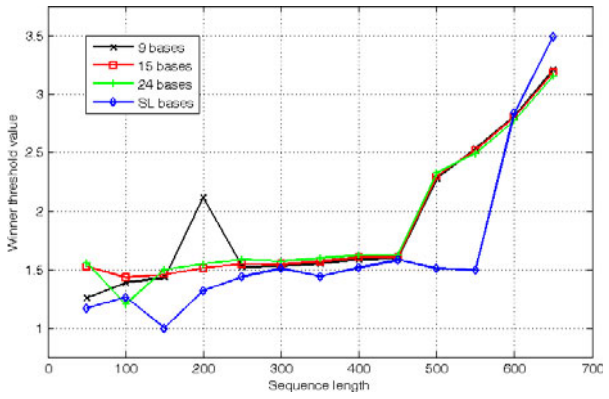

Figure 11 Winner threshold value $T_w(P_3)$ versus SL of 13 short sequence sets with WL = 9, 15, 24, SL bases using K-Quaternary Code I.

sequence group II of Table 3) of sequence lengths [90, 120, 180, 210, 270, 330, 390, 420, 480, 510, 570, 630] bases, are displayed in Figures 13 and 14 using thin red bars. For comparison, the winner threshold values and the corresponding precisions of the thirteen short sequence sets for each of the four WL are displayed in Figures 13 and 14 using thick blue bars. From Figures 13 and 14, it can be observed that the interpolated winner threshold values and their precisions perform well and follow the general trends.

2.5.2. Long sequences

In a similar manner to the short sequence sets, for each WL, an interpolated winner threshold value for an arbitrary sequence length can be determined using cubic spline interpolation from the corresponding discrete winner threshold values of the thirteen long sequence sets shown in Figure 12. For each of the four WL,

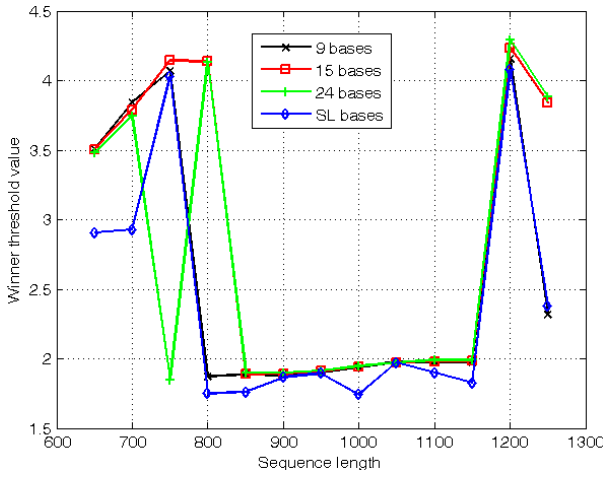

Figure 12 Winner threshold value $T_w(P_3)$ versus SL of 13 long sequence sets with WL = 9, 15, 24, SL bases using K-Quaternary Code I.

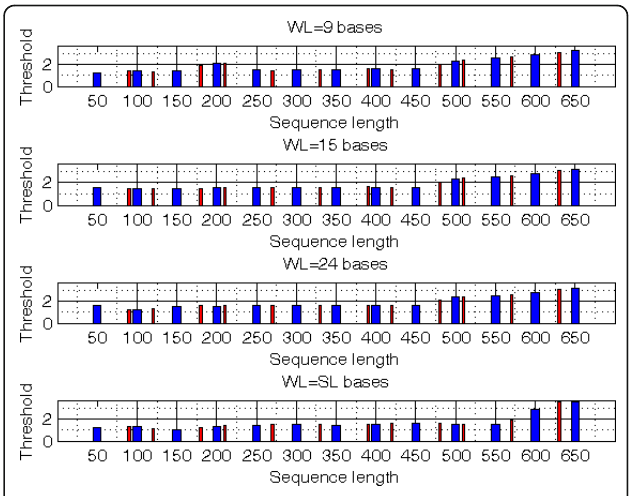

Figure 13 Interpolated winner threshold values (thin red bar) of 12 untrained short sequence sets obtained from winner threshold values (thick blue bar) of 13 short sequence sets using K-Quaternary Code I.

twelve untrained sequence sets (under the long sequence group II of Table 5) of sequence lengths [690, 720, 780, 810, 870, 930, 990, 1020, 1080, 1110, 1170, 1230] bases are tested, the interpolated winner threshold values and the corresponding precisions obtained are displayed in Figures 15 and 16 using thin red bars. For comparison, the winner threshold values and the corresponding precisions for each of the four WL of the thirteen long sequence sets are displayed in Figures 15 and 16 using thick blue bars. From Figures 15 and 16, it can be observed that the interpolated winner threshold

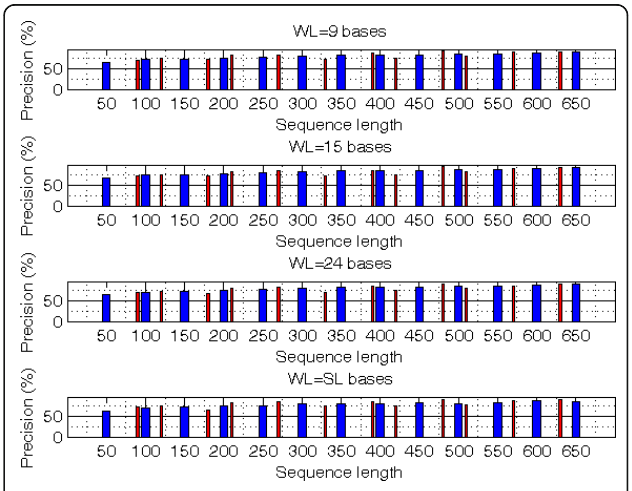

Figure 14 Precisions (thin red bar) of 12 untrained short sequence sets (obtained using interpolated winner threshold values) versus precisions (thick blue bar) of 13 short sequence sets (obtained using winner threshold values), both using K-Quaternary Code I.

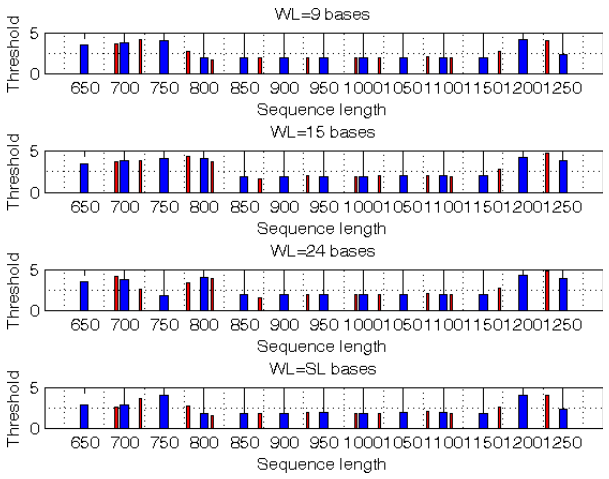

Figure 15 Interpolated winner threshold values (thin red bar) of 12 untrained long sequence sets obtained from winner threshold values (thick blue bar) of 13 long sequence sets using K-Quaternary Code I.

values and their precisions work well and follow the general trends.

2.6. Classification performances of 12 organisms

To demonstrate the efficiency of the numerical representations and thresholding methods when applied to other model organisms, the UCSC genomes [26-29] of twelve organisms (including the human) listed in Table 7 are tested. Each of the genomes of the twelve organisms is trained and tested with identical SL of 300 bases, four window lengths, and four threshold values, the results are compared based on the best of the Codes 1-16 (dark blue), the Code 10 (light blue), the Code 13 (yellow), and the Code 17 (red) [17] (as described in Section 2.1) as shown in Figures 17-18. For the twelve organisms, Figures 17(a-b) plot precision (a, top) and the corresponding code index (b, bottom); and Figure 18(a-c) plot window length index (a, top), threshold index (b, middle), and threshold value (c, bottom) of the top classifications obtained. The results summarized in Table 8 indicate that top classifications are obtained by the Code 13 in seven organisms; followed by the Code 17 with four top classifications (with the Code 13 ranks 2nd in 2 cases and the Code 10 ranks 2nd in 1 case, each of these 3 cases is shown in Table 8 on the row following that of the top performer); and the Code 6 with one top classification. The results also indicate that the Code 13 with (a) WL = 24 bases and T_p thresholding is the top choice for organisms 1, 3, 4; (b) WL = 24 bases and T_4 thresholding is the top choice for organism 7; (c) WL = 15 bases and T_c thresholding is the top choice for organisms 11 and 12; and (d) WL = 300 bases and T_m thresholding is the top choice for organism 2.

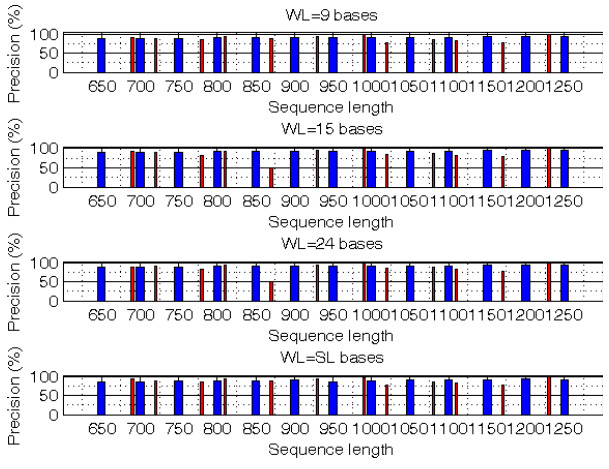

Figure 16 Precisions (thin red bar) of 12 untrained long sequence sets (obtained using interpolated winner threshold values) versus precisions (thick blue bar) of 13 long sequence sets (obtained using winner threshold values), both using K-Quaternary Code I.

3. Discussion

In this article, the ability of the K-Quaternary Code I (the Code 13) through the use of the discrete Fourier transform to capture the periodicities of an exon sequence or an intron sequence across the entire spectrum at a resolution defined by the window length is shown in Figures 3 and 4. Such a spectral tracking ability is shared among all of the sixteen numerical representations. As seen from Figures 3 and 4, there are three prominent peaks located at frequency equal to 0, 2π/3, and 4π/3. The peak at 2π/3 corresponds to the period-3 power spectral value, and the peak at 0 corresponds to the power spectral value at DC which usually exhibits the highest value within a spectrum. The power spectral values at other frequencies are lower and different but all serve to reflect their actual power spectral properties across the spectrum. For the Codes 9-16, their numerical represented sequences $x[n]$ are complex; therefore, DFT spectrum shows non-symmetrical peaks at frequency equal to 2π/3 and 4π/3. However, for the Codes 1-8 in which their numerical represented sequences $x[n]$ are real and therefore symmetrical peaks at frequency equal to 2π/3 and 4π/3 are obtained from DFT analysis.

The interpolated winner threshold values, $T_i(P_3)$, and their corresponding precisions shown in Figures 13 and 14 for short sequences, and in Figures 15 and 16 for long sequences of the human genome indicate that $T_i(P_3)$ obtained using cubic spline interpolation from either Figure 11 or Figure 12 can yield a similar precision as compared to that of the winner threshold value, $T_w(P_3)$, of its nearby SL. It should be noted that each of

Table 7 UCSC genome of 12 organisms

OG	Clade	Genome	Assembly	Track	Table	Type	Number
1	Mammal	Human	Feb. 2009 (GRCh37/hg19)	UCSC Genes	knownGene	Exon Intron	35685 532489
2	Mammal	Mouse	July 2007 (NCBI37/mm9)	UCSC Genes	knownGene	Exon Intron	27452 43041
3	Mammal	Pig	Nov. 2009 (SGSC Sscrofa9.2/susScr2)	Ensembl Genes	ensGene	Exon Intron	11499 113946
4	Vertebrate	Chicken	May 2006 (WUGSC 2.1/galGal3)	Ensembl Genes	ensGene	Exon Intron	12045 57770
5	Vertebrate	Zebrafish	Jul. 2010 (Zv9/danRer7)	RefSeq Genes	refGene	Exon Intron	7838 78776
6	Deuterostome	*C. intestinalis*	Mar. 2005 (JGI 2.1/ci2)	Ensembl Genes	ensGene	Exon Intron	7360 86885
7	Deuterostome	*S. purpuratus*	Sep. 2006 (Baylor 2.1/strPur2)	Other RefSeq	xenoRefGene	Exon Intron	4888 193451
8	Insect	*D. melanogaster*	Apr. 2006 (BDGP R5/dm3)	RefSeq Genes	refGene	Exon Intron	37925 35403
9	Insect	*A. mellifera*	Jan. 2005 (Baylor 2.0/apiMel2)	Ensembl Genes	ensGene	Exon Intron	21827 47605
10	Nematode	*C. elegans*	May 2008 (WS190/ce6)	RefSeq Genes	refGene	Exon Intron	25360 37359
11	Nematode	*C. japonica*	Mar. 2008 (WUGSC 3.0.2/caeJap1)	Other RefSeq	xenoRefGene	Exon Intron	7978 30958
12	Other	Sea hare	Sept. 2008 (Broad 2.0/aplCal1)	Other RefSeq	xenoRefGene	Exon Intron	6914 431792

Group: Genes and Gene Prediction Tracks. SL = 300 bases.

$T_m(P_3)$, $T_p(P_3)$, and $T_c(P_3)$ required to determine $T_w(P_3)$ has to be computed directly from the training portion of each short/long sequence set whereas $T_i(P_3)$ requires minimal computation but can achieve a comparable precision.

Given an unknown human nucleotide sequence of an arbitrary length L, if L is equal to the length of any of the thirteen short sequence sets or any of the thirteen long sequence sets, the choice of a suitable set of code, WL, and threshold can be obtained as a table-look-up from either Table 4 or 6. If L is not equal to the length of any of these thirteen short or long sequence sets, the closest SL, its code, and WL can also be obtained from either Table 4 or 6. For the latter case, once the code

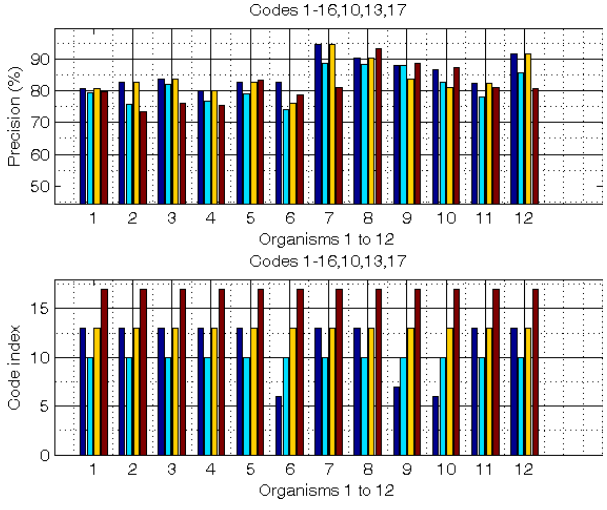

Figure 17 (a-b) Precision (a, top) and code index (b, bottom) of top classifications of 12 organisms obtained by best of Codes 1-16 (dark blue), Code 10 (light blue), Code 13 (yellow), and Code 17 (red).

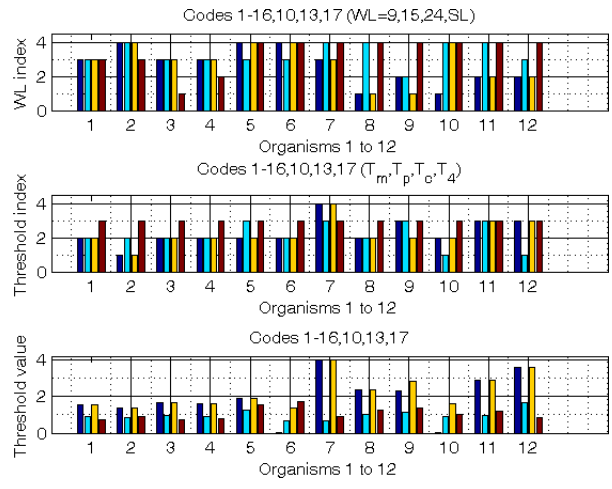

Figure 18 (a-c) WL index (a, top), threshold index (b, middle), and threshold value (c, bottom) of top classifications of 12 organisms obtained by best of Codes 1-16 (dark blue), Code 10 (light blue), Code 13 (yellow), and Code 17 (red).

Table 8 Code index (Figure 17b, bottom), WL (bases) (Figure 18a, top), threshold method (Figure 18b, middle), and precision (%) (Figure 17a, top) of top classifications of 12 organisms (SL = 300 bases)

OG	Code	WL	Threshold	Precision
1	13	24	T_p	80.8281
2	13	300	T_m	82.9433
3	13	24	T_p	83.6184
4	13	24	T_p	80.1530
5	17	300	T_c	83.3933
5	13	300	T_p	-
6	6	300	T_p	82.6283
7	13	24	T_4	94.9145
8	17	300	T_c	93.2943
8	13	9	T_p	-
9	17	300	T_c	88.7039
9	10	15	T_c	-
10	17	300	T_c	87.3087
11	13	15	T_c	82.3132
12	13	15	T_c	91.8992

and WL are determined, its threshold value can then be determined using the interpolated winner threshold described in Section 2.5. Besides the human genome, the methodologies described in this article can be applied to the genome of other model organisms as verified by the results shown in Figures 17(a-b) and 18 (a-c) and Tables 7 and 8.

4. Conclusions

In this article, two methods for determining threshold values have been defined, and together with the cumulative distribution threshold value, determine the winner threshold value for classifying an unknown nucleotide sequence of a fixed length. An interpolated winner threshold value has also been introduced to classify an unknown nucleotide sequence of an arbitrary length with a comparable performance to that obtained by the winner threshold value of its nearby SL (in classifying an unknown nucleotide sequence of a fixed length). In general, precision increases as sequence length increases, and classification performance depends on a suitable choice of numerical representation and window length. Sixteen 1-sequence numerical representations have been presented and compared to classify untrained exon and intron sequences in the spectral domain, in which the K-Quaternary Code I yields attractive performance. When comparing each of the sixteen windowed 1-sequence numerical representations using WL = [9,15,24] bases to a non-windowed 4-sequence numerical representation (such as the Voss representation), the speed improvement ratio increases as SL increases which favours long nucleotide sequence analysis. The results obtained indicate the methodologies introduced in this article for exon and intron sequence classification are applicable to the genomes of the human and other model organisms. Overall, the study has developed novel methodologies in numerical representation for improved nucleotide to numeric mapping, spectral analysis for effective period-3 spectral value computation, and thresholding for more accurate classification of unknown exon and intron sequences of fixed and arbitrary length.

Author details
[1]Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada [2]Faculty of Medicine, University of Ottawa, 451 Smyth Road, Ottawa, ON K1H 8M5, Canada [3]School of Medicine, Queen's University, 80 Barrie Street, Kingston, ON K7L 3N6, Canada

Competing interests
The authors declare that they have no competing interests.

Received: 6 October 2011 Revised: 6 October 2011
Accepted: 28 February 2012 Published: 28 February 2012

References

1. PP Vaidyanathan, Genomics and proteomics: A signal processor's tour. IEEE Circ. Syst. Mag. 4th Q 6–29 (2004)
2. D Anastassiou, Genomic signal processing. IEEE Signal Process. Mag. **18**, 8-20 (2001)
3. HK Kwan, SB Arniker, Numerical representation of DNA sequences, in Proceedings of *IEEE International Conference on Electro/Information Technology (EIT)*, Windsor, Ontario, Canada,.307–310 (7-9 June 2009)
4. SB Arniker, HK Kwan, Graphical representation of DNA sequences, in Proceedings of *IEEE International Conference on Electro/Information Technology (EIT)*, Windsor, Ontario, Canada, .311–314 (7-9 June 2009)
5. PD Cristea, in *BIOS'02: Genetic Signal Representation and Analysis*, SPIE International Conference on Biomedical Optics Symposium, Molecular Analysis and Informatics, San Jose, CA, USA, 21-24 January 2002. Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, vol. 4623 (SPIE, 2002). 77–84
6. M Akhtar, J Epps, E Ambikairajah, On DNA numerical representations for period-3 based exon prediction. 4 pages in *Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS),* Tuusula, Finland, (10-12 June 2007)
7. GL Rosen, *Signal Processing for Biologically-inspired Gradient Source Localization and DNA Sequence Analysis*, (PhD dissertation, Georgia Institute of Technology, Atlanta, August 2006)
8. M Akhtar, J Epps, E Ambikairajah, Signal processing in sequence analysis: Advances in eukaryotic gene prediction. IEEE J Sel Top Signal Process. **2**, 310–321 (2008)
9. T Holden, R Subramaniam, R Sullivan, E Cheng, C Sneider, G Tremberger Jr, A Flamholz, DH Leiberman, TD Cheung, ATCG nucleotide fluctuation of *Deinococcus radiodurans* radiation genes. in *Instruments, Methods, and Missions for Astrobiology X. Proceedings of Society of Photo-Optical 2 Instrumentation Engineers (SPIE) Conference, vol. 6694 (SPIE, 12 September 2007)*, ed. by B Hoover, GV Levin, AY Rozanov, and PCW Davies669417-1–669417-10
10. HE Stanley, SV Buldyrev, AL Goldberger, ZD Goldberger, S Havlin, SM Ossadnik, C-K Peng, M Simmons, Statistical mechanics in biology: how ubiquitous are long-range correlations?. Physica A. **205**, 214–253 (1994). doi:10.1016/0378-4371(94)90502-9
11. AS Nair, SS Pillai, A coding measure scheme employing electron-ion interaction pseudo potential (EIIP). Bioinformation. **1**, 197–202 (2006)
12. I Cosic, Macromolecular Bioactivity: is it resonant interaction between macromolecules? Theory and applications. IEEE Trans Biomed Eng. **41**, 1101–1114 (1994). doi:10.1109/10.335859

13. N Chakravarthy, A Spanias, LD Lasemidis, K Tsakalis, Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal of Genomic Signal Processing. **1**, 13–28 (2004)
14. PD Cristea, Conversion of nucleotides sequences into genomic signals. J Cell Mol Med. **6**, 279–303 (2002). doi:10.1111/j.1582-4934.2002.tb00196.x
15. PD Cristea, Representation and analysis of DNA sequences. in *Genomic Signal Processing and Statistics, EURASIP Book Series in Signal Processing and Communications, volume 2 (Hindawi Publishing Corporation, 2005)*, ed. by ER Dougherty, I Shmulevich, J Chen, ZJ Wang15–65
16. AK Brodzik, O Peters, Symbol-balanced Quaternionic periodicity transform for latent pattern detection in DNA sequences. in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol 5, Philadelphia, USA, 373–376 (March 2005)
17. S Tiwari, S Ramachandran, A Bhattacharya, S Bhattacharya, R Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences. Bioinformatics (CABIOS). **13**(3), 263–270 (1997). doi:10.1093/bioinformatics/13.3.263
18. S Datta, A Asif, A fast DFT based gene prediction algorithm for identification of protein coding regions. in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol 5, Philadelphia, USA, 653–656 (March 2005)
19. BYM Kwan, JYY Kwan, HK Kwan, R Atwal, OT Shen, Wavelet analysis of the genome of the model plant Arabidopsis thaliana. in *Proceedings of TENCON*, Hong Kong, China, 1-4 (14-17 November 2006)
20. JA Berger, SK Mitra, M Carli, A Neri, New approaches to genome sequence analysis based on digital signal processing. in *Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, North Carolina*, 1-4 (October 2002)
21. EA Cheever, DB Searls, W Karunaratne, GC Overton, Using signal processing techniques for DNA sequence comparison, in *Proceedings of the 1989 Fifteenth Annual Northeast Bioengineering Conference*, Boston, MA.173–174 (27-28 March 1989)
22. S Rajasekaran, H Nick, PM Pardalos, S Sahni, G Shaw, Efficient algorithms for local alignment search. J Comb Optim. **5**, 117–124 (2001). doi:10.1023/A:1009893719470
23. GD Avenio, M Grigioni, G Orefici, R Creti, SWIFT (sequence-wide investigation with Fourier transform): a software tool for identifying proteins of a given class from the unannotated genome sequence. Bioinformatics. **21**(13), 2943–2949 (2005). doi:10.1093/bioinformatics/bti468
24. CC Yin, SS-T Yau, Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol. **247**, 687–694 (2007). doi:10.1016/j.jtbi.2007.03.038
25. J Tuqan, A Rushdi, A DSP approach for finding the codon bias in DNA sequences. IEEE J Sel Topics Signal Process. **2**(3), 343–356 (2008)
26. 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: D493–496 (2004). doi:10.1093/nar/gkh103
27. 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. **11**(8) (2010). Article R86. doi:10.1186/gb-2010-11-8-r86
28. 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**(Unit 19.10), 1–21 (2010)
29. 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: a platform for interactive large-scale genome analysis. Genome Res. **15**(10), 1451–1455 (2005). doi:10.1101/gr.4086505

doi:10.1186/1687-6180-2012-50
Cite this article as: Kwan *et al.*: Novel methodologies for spectral classification of exon and intron sequences. *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:50.