JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 10, 2008

Pp. 1237–1256
DOI: 10.1089/cmb.2008.0041

Coding Region Prediction Based on a Universal DNA Sequence Representation Method

XIANYANG JIANG,[1] DOMINIQUE LAVENIER,[2] and STEPHEN S.-T. YAU[3]

ABSTRACT

Graphical representation of DNA sequences provides a simple and intuitive way of viewing, anchoring, and comparing various gene structures, so a simple and non-degenerate method is attractive to both biologists and computational biologists. In this study, a universal graphical representation method for DNA sequences based on S.S.-T. Yau's method is presented. The method adopts a trigonometric function to represent the four nucleotides A, G, C, and T. Some interesting characteristics of the universal representation are introduced. We exploit frequency analysis with our representation method on DNA sequences, demonstrating possible applications in coding region prediction, and sequence analysis. Based on the statistically experimental results from this frequency analysis, a simple coding region predictor and an optimized one are presented. An experiment on the broadly accepted ROSETTA data set demonstrates that the performance of the optimized predictor is comparable to that of other popular methods.

Key words: bioinformatics, frequency analysis, mining methods and algorithms, representations, signal processing.

1. INTRODUCTION

FOR DECADES, one of the key challenges for biologists is to understand the structure and function of DNA sequences. More recently, computer scientists have tried to provide powerful and flexible computational tools for biologists to make more progress in this area. In particular, graphical representation of DNA sequences provides a simple way of viewing, anchoring, and comparing various gene structures. It is an attractive and promising research tool for bioinformatics.

The first important and simple method in this direction is the three-dimensional curve used about 20 years ago to represent a DNA sequence (Hamori and Ruskin, 1983; Hamori, 1985). Unfortunately, sophisticated computer graphic tools are needed to produce the H curve (Hamori, 1994). In the 1980s, a two-dimensional graphical representation was proposed that was simpler than the H curve (Gates, 1985, 1986). Gates' graphical representation, however, has high degeneracy (i.e., repetitive closed loops or circuits will appear

[1]Institute of Microelectronics and Information Technology, Wuhan University, Wuhan, China.
[2]IRISA-INRIA, Campus de Beaulieu, Rennes cedex, France.
[3]Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, Illlinois.

in the DNA representation graph) for some sequences. In 1990, Jeffrey (1990) presented a chaos game representation (CGR) of gene structures. In his method, however, no strict mathematical description was provided. On the other hand, such mathematical work was done by Tiňo (1999), but the work remains far from the real graphical representation of a DNA sequence. In 1992, Berthelsen et al. used a fractal method to represent DNA sequences (Berthelsen et al., 1992). Their main contribution is the estimation of the global fractal dimension of a DNA sequence, but the estimation is affected seriously by the length and the embedding dimension of the sequence. In 1993, Wu et al. used an iterated function system to unify the H-curve, the chaos game representation, and W-curve (Wu et al., 1993). The newly presented W-curve, however, cannot provide detailed representation when the sequence is too long. This shortcoming is due to the compactness of the W-curve. In 1994, Zhang and Zhang used the Z-curve to represent and analyze DNA sequences (Zhang and Zhang, 1994). It seems that the Z-curve representation was the most successful method for the graphical representation of DNA sequences.

Recently, digital signal processing methods and other mathematical tools provide another flexible way to analyze this kind of representation (Anastassiou, 2000). Cheever et al. mapped DNA symbols into the plane of complex numbers, i.e., "A" to 1, "T" to -1, "G" to i (where $i = \sqrt{-1}$), and "C" to $-i$. Based on this mapping, they tried to find similarities between two sequences by correlating their corresponding complex sequences (Cheever et al., 1989). Wang and Johnson (2002) directly mapped "A," "C," "G," "T" into 1, 2, 3, 4, respectively, and performed scalogram and spectrogram analysis on various sequences, but this mapping faces a fundamental problem of adding a property in which one symbol is larger than another. By doing so, this mapping may not reflect the original information from DNA. Su et al. (2003) provided a mapping method based on a "pattern" filtering, the resulting gap sequences abstractly represent the original sequence in the form of gaps. This method has side effects at both the beginning and the end of the output signal.

To improve the performance of a graphical representation method, Yau et al. (2003) presented a method which used a two quadrant Cartesian coordinate system for denoting DNA sequences. The authors have proved that their two-dimensional graphical representation method is the best method for graphical representation. Obviously, Yau's idea works well as long as the nucleotides "A," "C," "G," "T" are represented by four linearly independent vectors in the right half plane.

In this study, we rewrite Yau's idea into a universal representation method as follows:

$$\begin{aligned} (\cos(-(\pi/2-\theta)), \sin(-(\pi/2-\theta))) &\to A, \\ (\cos\theta, \sin\theta) &\to C, \\ (\cos(-\theta), \sin(-\theta)) &\to G, \\ (\cos(\pi/2-\theta)), \sin(\pi/2-\theta)) &\to T. \end{aligned} \tag{1}$$

that is,

$$\begin{aligned} (\sin\theta, -\cos\theta) &\to A, \\ (\cos\theta, \sin\theta) &\to C, \\ (\cos\theta, -\sin\theta) &\to G, \\ (\sin\theta, \cos\theta) &\to T. \end{aligned} \tag{2}$$

where $0 < \theta < \pi/2$ and $\theta \neq \pi/4$.

In principle, a good representation method should not cause degeneracy; otherwise, not only is it worse than a traditional one in this point, but also loses the ability to retain the biological information. By a method similar to one found in Yau's paper, one can easily prove that there is no circuit or degeneracy in the universal graphical representation.

There is a one-to-one correspondence between the universal graphical representation and the original DNA sequence, and sequence alignment can be done by simply identifying similar segments of the graph. On the other hand, the original DNA sequence can be recovered from its graph mathematically without loss of any biological information.

The rest of the article is organized as follows. The universal representation method's characteristics are given in Section 2. In Section 3, coding region prediction based on our representation method is detailed. Section 4 concludes the article.

2. CHARACTERISTICS OF OUR REPRESENTATION GRAPH

Though our representation method seems to be a minor modification on that by Yau et al. (2003), it has some interesting characteristics not possessed by other similar representation methods.

2.1. Unification with other representation methods

From our idea we can easily draw out that the unit vectors presented by Gates in the Cartesian coordinate plane can be rewritten as follows:

$$\begin{aligned} (\cos(-\pi/2), \sin(-\pi/2)) &\to A, \\ (\cos(-\pi), \sin(-\pi)) &\to C, \\ (\cos 0, \sin 0) &\to G, \\ (\cos(\pi/2), \sin(\pi/2)) &\to T. \end{aligned} \tag{3}$$

Based on Equation (3), Gates' representation satisfies a trigonometric function to a certain extent, though such kind of representation generates degeneracy as pointed out by Yau et al. (2003).

In this point of view, with a variable θ, our representation method can unify some different representation methods, which is new for a representation method.

2.2. Amplitude characteristic

It is easy to see that the amplitude of each point in our representation graph is the sum of the amplitudes of the preceding nucleotides. Therefore, looking at the differentials from a piece of the representation line, we get:

$$\Delta y_0 = (\Delta n_t - \Delta n_a)\cos\theta + (\Delta n_c - \Delta n_g)\sin\theta \tag{4}$$

where Δn_a, Δn_c, Δn_g, Δn_t, and Δy_0 represent the changes of the four nucleotides A, C, G, T, and the change of amplitude respectively between two points in the DNA representation graph.

From Equation (4), if the change of the four nucleotides are almost the same, i.e., $\Delta n_t = \Delta n_a = \Delta n_c = \Delta n_g$, then $\Delta y_0 \simeq 0$. In other words, if the distribution of the four nucleotides is the same for one segment of a sequence, then the amplitudes for the two ends of the segment in the representation will be the same.

If we apply the symmetry of the four nucleotides, we also get other similar representations of the same segment:

$$\begin{aligned} \Delta y_1 &= (\Delta n_a - \Delta n_t)\cos\theta + (\Delta n_c - \Delta n_g)\sin\theta \\ \Delta y_2 &= (\Delta n_g - \Delta n_a)\cos\theta + (\Delta n_c - \Delta n_t)\sin\theta \\ \Delta y_3 &= (\Delta n_a - \Delta n_g)\cos\theta + (\Delta n_c - \Delta n_t)\sin\theta \end{aligned} \tag{5}$$

If $\Delta y_0 = \Delta y_1 = \Delta y_2 = \Delta y_3 = 0$, then we get $\Delta n_t = \Delta n_a = \Delta n_c = \Delta n_g$. In other words, if the amplitudes for the two ends of the segments in the representative graphs are the same, then the distributions of the four nucleotides in the segments are almost the same.

This characteristic is useful for predicting the probability of the appearance of each nucleotide as well as the locations of exons and introns. Similar measures for discrimination between exons and introns are well developed by Kotlar and Lavner (2003) and C. Mathé et al. (2002).

2.3. Frequency characteristic

For the frequency analysis of a sequence of complex numbers, we shall use a *Discrete Fourier Transform* (DFT).

Let $S = \{a_1, a_2, \ldots, a_N\}$ be a given DNA sequence. For each A, T, C, and G, we shall associate it with a numerical vector by our graphical representation method. In general, let $y = \{y_1, y_2, \ldots, y_N\}$, where y_k is the numerical representation for a_k by our method and $1 \leq k \leq N$. The DFT of sequence y is another sequence $Y[k]$ of the same length as defined by:

$$Y[k] = \sum_{n=1}^{N} y_n e^{-j2\pi(k-1)(n-1)/N}, \qquad k = 1, 2, \ldots, N. \tag{6}$$

The sequence $Y[k]$ provides a measure of the frequency content at "frequency" k, which corresponds to an underlying "period" of $N/(k-1)$ samples (Anastassiou, 2000).

Here we first do the following to transform the original sequence into four sequences:

$$\begin{aligned}
S_a[k] &= \begin{cases} \sin\theta - i\cos\theta & \text{for } a_k = A \\ 0 & \text{for } a_k \neq A \end{cases} \\
S_c[k] &= \begin{cases} \cos\theta + i\sin\theta & \text{for } a_k = C \\ 0 & \text{for } a_k \neq C \end{cases} \\
S_g[k] &= \begin{cases} \cos\theta - i\sin\theta & \text{for } a_k = G \\ 0 & \text{for } a_k \neq G \end{cases} \\
S_t[k] &= \begin{cases} \sin\theta + i\cos\theta & \text{for } a_k = T \\ 0 & \text{for } a_k \neq T \end{cases}
\end{aligned} \tag{7}$$

Sequence $S_a[k]$ has a relation with sequence $u_A(k)$ adopted by Yin and Yau (2005), which is expressed by

$$S_a[k] = (\sin\theta - i\cos\theta)u_A(k). \tag{8}$$

Other three sequences $S_g[k]$, $S_c[k]$, and $S_t[k]$ have the same relation with their corresponding sequences $u_G(k)$, $u_C(k)$, and $u_T(k)$ defined by Yin and Yau (2005), respectively.

Then the DFT is applied to each of these four sequences $S_a[k]$, $S_g[k]$, $S_c[k]$, and $S_t[k]$ to get four spectral representations $S_A(k)$, $S_G(k)$, $S_C(k)$, and $S_T(k)$. The power spectrum of DNA sequence S is defined as

$$S(k) = |S_A(k)|^2 + |S_C(k)|^2 + |S_G(k)|^2 + |S_T(k)|^2. \tag{9}$$

Because the DFT is linear, we have

$$|S_A(k)|^2 = |\sin\theta - i\cos\theta|^2 PS_A(k) = PS_A(k), \tag{10}$$

where $PS_A(k)$ is the Fourier power spectrum of sequence $u_A(k)$. Similar relationships exist between $S_C(k)$, $S_G(k)$, and $S_T(k)$ and the respective sequences $PS_C(k)$, $PS_G(k)$, and $PS_T(k)$. Therefore, our new spectrum $S(k)$ is equal quantitatively to the spectrum $PS(k)$ by Yin and Yau (2005) and power spectrum $S(k)$ can be used to demonstrate the distinctive feature of protein coding regions in DNA. Namely, the power spectrum has an absolute peak at frequency $k = N/3$ for a coding region, but such a phenomenon does not occur in a non-coding region. As pointed out by Yin and Yau (2005), this is due to the fact that the DNA code consists of triplets(codons) and that not all nucleotides are used equally in codons/triplets positions.

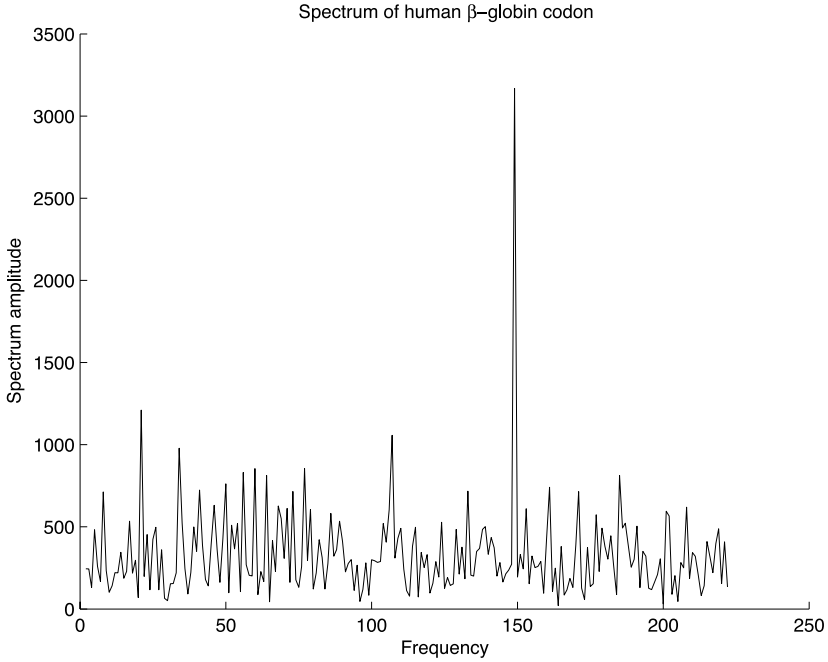

FIG. 1. Spectrum of human β-globin gene coding region when $\theta = \pi/3$.

Our analysis method is slightly different from those presented by Anastassiou (2000) because our method is based on our two-dimensional representation for coding regions.

We show the spectrum of the DNA coding region of the human β-globin gene (from AF527577 or gi:22094826) and the mouse β-globin gene (from J00413 or gi:193793) in Figures 1 and 2, respectively. The spectrum of mouse β-globin major gene based on the above analysis is shown in Figure 3. From these figures we can see that the Fourier spectrum of a coding DNA typically has a peak at frequency $k = N/3$,

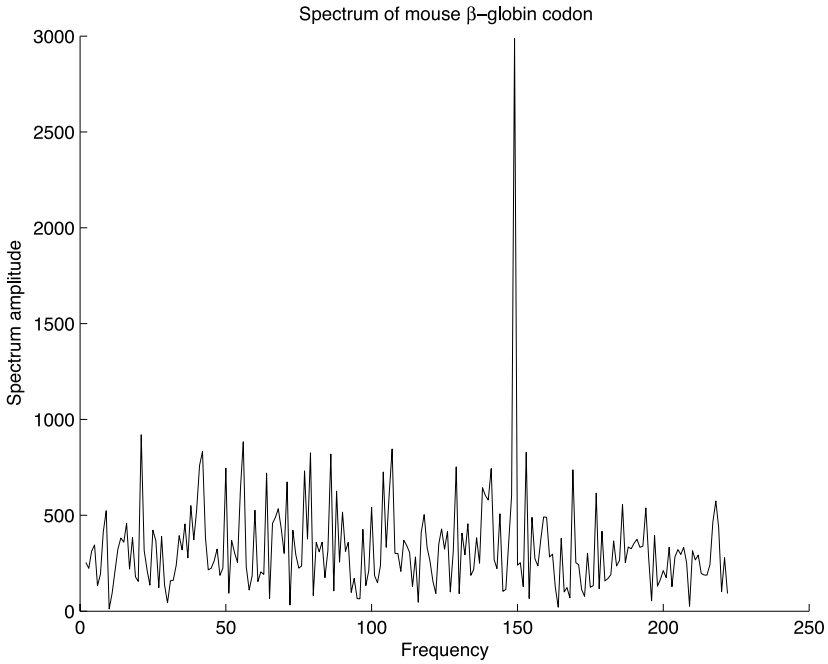

FIG. 2. Spectrum of mouse β-globin gene coding region when $\theta = \pi/3$.

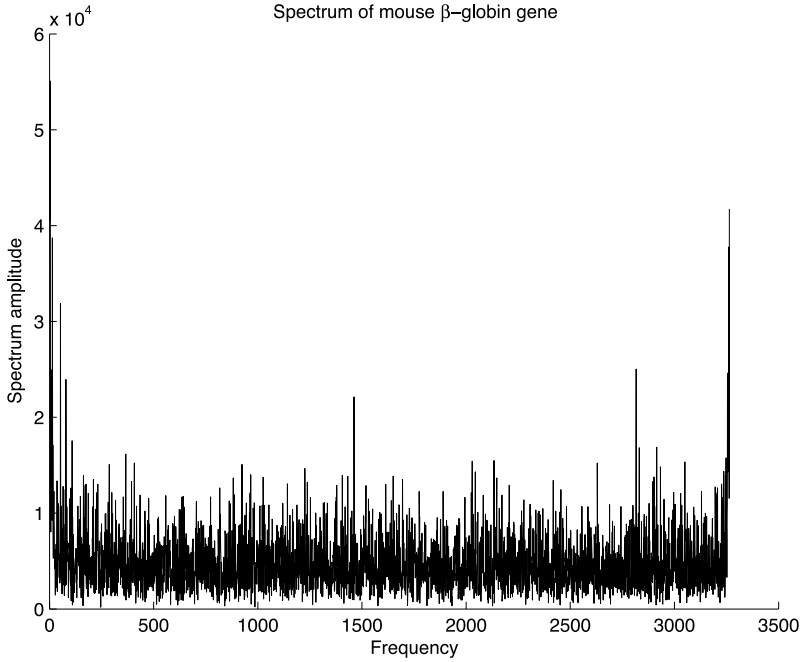

FIG. 3. Spectrum of mouse β-globin major gene when $\theta = \pi/3$, without significant peak due to including non-coding regions.

but for a non-coding DNA sequence the power spectrum generally does not have any significant peaks. In Figures 1 and 2, both coding regions consist of 444 bp, and the peaks are correctly shown at $N/3 = 148$. In Figure 3, it is clearly demonstrated that there is no significant peak for the power spectrum; this is due to the fact that the mouse β-globin major gene includes some non-coding regions.

3. CODING REGION PREDICTION BASED ON OUR REPRESENTATION

3.1. Peak significance evaluation for the spectrum

Quantitatively, we use a two-step method with z-score to evaluate the "significance" of the peak in the spectrum of a coding region indicated by $S(k)$.

Step 1: This step is applied to each sequence to pick out those with a peak of a certain significance at frequency $N/3$.

For a certain θ, let m_1 and d_1 be the mean and the standard deviation, respectively, of the series $S(k)$. We calculate the z-score for each nucleotide position of the sequence by

$$Z_1(k) = (S(k) - m_1)/d_1, k = 1, 2, \ldots, N. \tag{11}$$

For series $Z_1(k)$, when the largest value is larger than 3.0 and located at frequency $N/3$, we call the peak corresponding to the largest z-score in the spectrum "significant."

Step 2: This step is used to verify the peak's significance obtained in Step 1.

In this step, each coding sequence with a significant peak picked out by Step 1 is shuffled p times. For each shuffled sequence we get a series $S(k)_i$, where $i = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, N$. From them we get a series $S(N/3)_i$, where $i = 1, 2, \ldots, p$.

Let m_2 and d_2 be the mean and the standard deviation, respectively, of series $S(N/3)_i$, where $i = 0, 1, \ldots, p$. $S(N/3)_0$ is from the real coding sequence. Then we get z-scores for these shuffled sequences and the real sequence

$$Z_2(i) = (S(N/3)_i - m_2)/d_2, i = 0, 1, \ldots, p. \tag{12}$$

If all $Z_2(i)$ for $i = 1, 2, \ldots, p$ are less than $Z_2(0)$ (the z-score for the real coding sequence at frequency $N/3$), then we verify that the sequence has a "real" significant peak at frequency $N/3$.

In order to demonstrate the possibility of applying our method for detecting coding regions, we have done some statistical experiments on human gene coding regions by the two-step method. In our experiments, all the test data were downloaded from NCBI gene database (*www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene*). The shortest coding region is 17 bp (from AC073127 or gi:14589774), and the longest coding region is 6858 bp (from NM_006015 or gi:21264564 and AF231056 or gi:11320941). All the coding regions are randomly selected by a fair chance from genes in the database. The annotation information of their source genes is used to make sure the coding regions are from real genes. Therefore, by the two-step method, we obtained the coding regions with significant peak as indicated by z-score Z_1 and the sequences with significant peak as verified by z-score Z_2 for a certain θ.

It is interesting that when θ changes from $\pi/30$ to $7\pi/15$ with a step $\pi/30$, the sets obtained in Step 1 and Step 2 remain the same respectively, in other words, the statistical results for Step 1 and Step 2 keep consistent for different values θ. So we only list the experimental results without showing the θ we used.

The experimental results on human gene coding regions are shown in Table 1, where the first column is the length range for real coding regions in the test set; N_{total} is the total number of real coding regions within the length range in the first column in the table. This number can be obtained from the annotation information of DNA sequences downloaded from the genebank; N_{Z_1} is the number of coding regions with significant peak checked by Step 1 from the total N_{total} coding regions; N_{Z_2} is the number of coding regions verified by Step 2 from those checked by Step 1.

From these results, we can see that, for long coding regions with length greater than 300 bp, there is a high probability ($> 70\%$) that their spectra will contain a significant peak at frequency $k = N/3$, but for short coding regions with length less than 300 bp, the probability is low. After detailed investigation, we have found an interesting phenomenon. The spectrum of a short coding region sometimes contains one or more significant peaks. Although the frequency of these peaks is not fixed, most of them are at frequency $k = N/3$. When there are one or more peaks in the spectrum of a short coding region the peak is not as distinctive as that shown in the spectrum of a long coding region.

Most of peaks gotten from Step 1 can be verified by Step 2. In other words, when the length of each sequence is less than 300 bp, the significance of the peak in Step 1 for some sequences cannot be verified by the results of Step 2; when the length of the sequence is larger than 300 bp, all the significance of the peak in Step 1 can be verified by Step 2 and so the peak does not appear by chance.

TABLE 1. STATISTICALLY EXPERIMENTAL RESULTS ON HUMAN GENE CODING REGIONS

Length	N_{total}	N_{Z_1}	$\frac{N_{Z_1}}{N_{total}}$ (%)	N_{Z_2}	$\frac{N_{Z_2}}{N_{Z_1}}$ (%)
0–100	16	3	18.8	2	66.7
101–200	28	10	35.7	5	50.0
201–300	18	11	61.1	7	63.6
301–400	32	23	71.9	23	100
401–500	70	68	97.1	68	100
501–600	17	14	82.4	14	100
601–700	11	10	90.9	10	100
701–800	16	13	81.3	13	100
801–900	9	9	100	9	100
901–1000	6	5	83.3	5	100
1001–2000	66	61	92.4	61	100
2001–3000	45	45	100	45	100
3001–4000	5	5	100	5	100
4001–5000	4	4	100	4	100
5001–6000	6	6	100	6	100
6001–7000	3	3	100	3	100

We have also done some statistical experiments on coding regions of *arabidopsis thaliana*. The test data are also from the NCBI gene database and chosen by the same method as for human coding regions to make sure the coding regions are segments from real genes. The shortest coding region is 60 bp (from AC004747 or gi:20197263), and the longest coding region is 8055 bp (from AC023673 or gi:7543635). We adopted the same quantitative measure on human gene coding regions, the experimental results are shown in Table 2.

From Table 2, we can get results similar to those from Table 1; i.e., for long coding regions with length greater than 500 bp, significant peak at frequency $k = N/3$ in their spectra has a high probability (>70%), but for short coding regions with length less than 500 bp, the probability of this occurring is low. The characteristics are also verified by the results from Step 2.

From the above statistical experiments, we can see that for coding regions long enough (e.g., >1000 bp), the power spectra each have a greater than 90% chance with a significant peak at frequency $N/3$. With such a high probability, our representation method is thus useful for detecting/predicting coding regions in long DNA sequences.

In order to investigate deeply whether the power spectrum can be used to find gene coding regions, we have taken two kinds of experiments on human gene coding regions with lengths above 6000 bp. The first kind is called DNA walk, which starts from the beginning or end of the sequence. Also, each walk increases by 500 bp. Thus, the first walk region is from 1 to 500 bp, and the second walk region is from 1 to 1000 bp, etc. Then the same frequency analysis is applied to these walk regions. The mechanism of the DNA walk is shown in Figure 4. The other kind is a random mode; i.e., the subregion for frequency

TABLE 2. STATISTICALLY EXPERIMENTAL RESULTS ON *Arabidopsis thaliana* GENE CODING REGIONS

Length	N_{total}	N_{Z_1}	$\frac{N_{Z_1}}{N_{total}}$ (%)	N_{Z_2}	$\frac{N_{Z_2}}{N_{Z_1}}$ (%)
0–100	2	0	0.0	0	—
101–200	11	2	18.2	1	50.0
201–300	62	34	54.8	18	52.9
301–400	30	11	36.7	11	100
401–500	40	16	40.0	16	100
501–600	21	17	81.0	17	100
601–700	37	30	81.1	30	100
701–800	30	24	80.0	24	100
801–900	40	30	75.0	30	100
901–1000	42	34	81.0	34	100
1001–1100	48	46	95.8	46	100
1101–1200	26	24	92.3	24	100
1201–1300	33	30	90.9	30	100
1301–1400	27	26	96.3	26	100
1401–1500	43	41	95.3	41	100
1501–1600	25	25	100	25	100
1601–1700	24	23	95.8	23	100
1701–1800	18	18	100	18	100
1801–1900	11	11	100	11	100
1901–2000	17	17	100	17	100
2001–3000	58	58	100	58	100
3001–4000	15	15	100	15	100
4001–5000	7	7	100	7	100
5001–6000	2	2	100	2	100
6001–7000	1	1	100	1	100
7001–8000	0	0	—	0	—
8001–9000	1	1	100	1	100

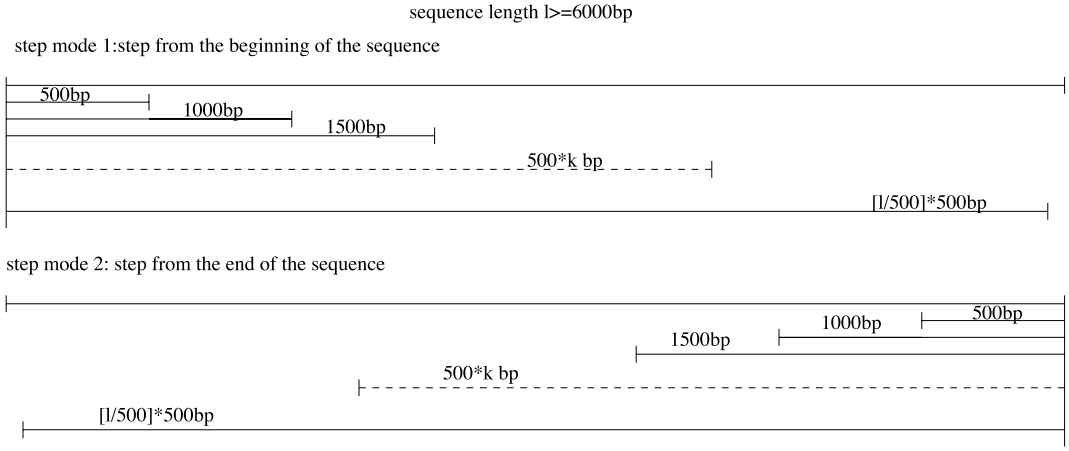

FIG. 4. DNA walk mode of frequency analysis for sub-coding regions.

analysis is chosen from the sequence with a fixed length of $(1000 \pm \delta)$ bp randomly, where $0 \leq \delta < 3$. The mechanism of the random mode is shown in Figure 5.

We show the statistical results in Figures 6 and 7.

It is interesting to note that both experiments obtain the same results; i.e., each spectrum of the sub-coding regions chosen based on the above two rules has its own significant peak at frequency $k = N/3$; on the contrary, for non-coding regions and with the same experiments, there is not such significant peak. Thus we conclude that using our representation method, the characteristic of a peak appears at frequency $k = N/3$ is easily detected for coding regions (or their pieces) but is absent for non-coding regions.

3.2. A simple coding region predictor

Based on the experimental results in the above subsection, we provide an efficient predictor for gene coding regions. We define $W(j, L)$ as the predictor for a coding region at nucleotide position j and with a window size L:

$$W(j, L) = |a|A_L|^2 + c|C_L|^2 + g|G_L|^2 + t|T_L|^2|, \tag{13}$$

where parameters a, c, g, and t get the same values in our representation; i.e.,

$$\begin{aligned} a &= \sin\theta - i\cos\theta, \\ c &= \cos\theta + i\sin\theta, \\ g &= \cos\theta - i\sin\theta, \\ t &= \sin\theta + i\cos\theta. \end{aligned} \tag{14}$$

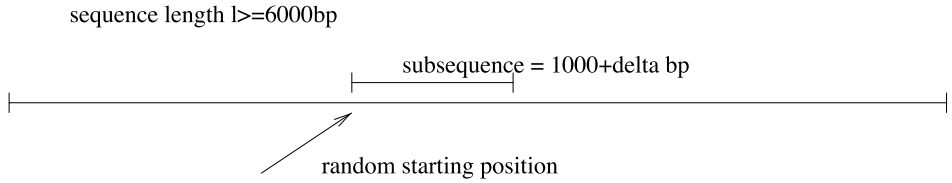

FIG. 5. Random mode of frequency analysis for sub-coding regions.

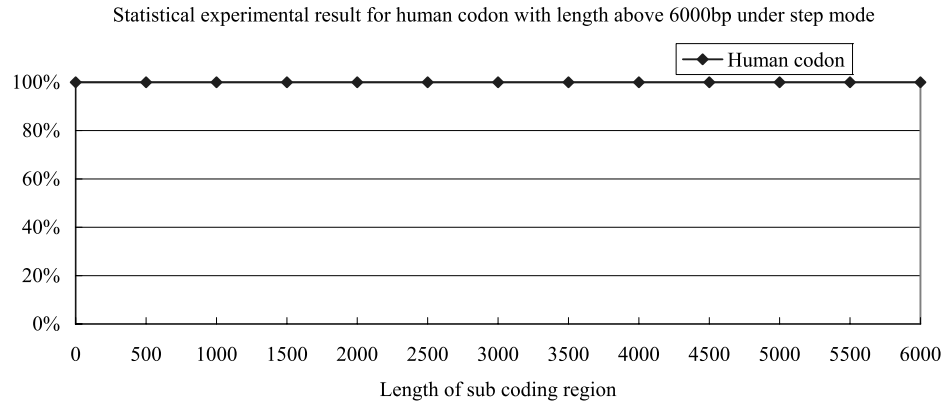

FIG. 6. Statistically experimental result on human codon with length above 6000 bp under DNA walk mode.

and

$$
\begin{aligned}
A_L &= \frac{1}{L} S_A(L/3)_j, \\
C_L &= \frac{1}{L} S_C(L/3)_j, \\
G_L &= \frac{1}{L} S_G(L/3)_j, \\
T_L &= \frac{1}{L} S_T(L/3)_j.
\end{aligned}
\tag{15}
$$

where $S_A(L/3)_j$, $S_C(L/3)_j$, $S_G(L/3)_j$, and $S_T(L/3)_j$ are DFT coefficients at frequency $k = L/3$ for a subsequence with the window width L starting at nucleotide j. When the window slides by one or more bases (we choose one base) each step along the DNA sequence, we get all $W(j, L)$.

Furthermore, we use a window with the same size L sliding along $W(j, L)$ curve and calculate the mean of $W(j, L)$ in the window; then we get a smoothed version $W_s(j, L)$ from $W(j, L)$. Based on the smoothed curve, the mean $W_{sm}(L)$ of all $W_s(j, L)$ is used as a threshold. When $W_s(j, L)$ is above the threshold, an exon is picked out; otherwise, an intron is picked out. The boundaries of each exon are decided by the positions j where $|W_s(j, L) - W_{sm}(L)|$ is minimum.

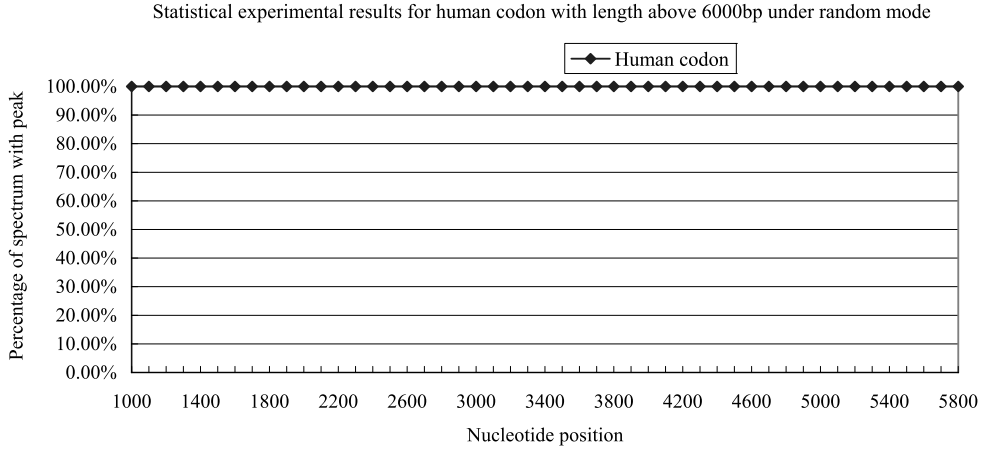

FIG. 7. Statistically experimental result for human codon with length above 6000 bp under random mode.

The window width L can be larger than or equal to the minimum length of coding regions of the target gene to suppress the background $1/f$ noise spectra (Li and Holste, 2005). Initial studies of the $1/f$ noise in DNA sequences were motivated by a model of the spatial $1/f$ noise of the symbolic sequence evolution. Subsequently, empirical $1/f$ noise spectra were indeed observed in non-protein coding DNA sequences, and their generality in DNA sequences was further illustrated by Voss (1992). Here we leave out the details about it because its description is beyond the main scope for this paper. In our study, first let L be an arbitrary value between the minimum length and the maximum length of the coding regions of the target gene to simplify the calculation. For example, we let L be 153 this step in the following experiments; then when we get the predicted exons with length $L_1, L_2, \ldots, L_e$, we changed L to be L_m and did the prediction to get the final prediction results, where e is the number of the predicted exons gotten by the first prediction step and $L_m = \frac{\sum_{num=1}^{e} L_{num}}{e}$.

We calculated $W(j, L)$ and its smoothed version $W_s(j, L)$ of gene F56F11.4 in the C-elegans chromosome III with widow width $L = L_m = 327$ and $\theta = \pi/30$; the results are shown in Figure 8. The gene has five exons (relative position 929–1135, 2528–2857, 4114–4377, 5465–5644, 7255–7605, respectively); we can see that the boundaries of these five exons are picked out accurately at positions j where $|W_s(j, L) - W_{sm}(L)|$ is minimum. The diagram of $W(j, L)$ and its smoothed version $W_s(j, L)$ for the human germ line β-globin gene (from V00499 or gi:29440) with window width $L = L_m = 147$ and $\theta = \pi/3$ are shown in Figure 9. The gene has three exons (relative position 104–245, 376–598, 1449–1709, respectively) whose boundaries are also correctly obtained by our method.

Now, we use statistics to evaluate the efficiency of the predictor $W(j, L)$. In order to do this, we calculate the percentage of the picked number N_p of coding regions. N_p is the number of coding regions picked out by the mean $W_{sm}(L)$ based on $W_s(j, L)$ from the N_{total} coding regions within a length range shown in the left columns in Tables 3 and 4.

The statistically experimental results for predictor $W(j, L)$ on *arabidopsis thaliana* and human gene coding regions are shown in Tables 3 and 4, respectively. It is interesting that the statistical results for predictor $W(j, L)$ are consistent with those gotten from power spectra by Step 1 shown in Tables 1 and 2,

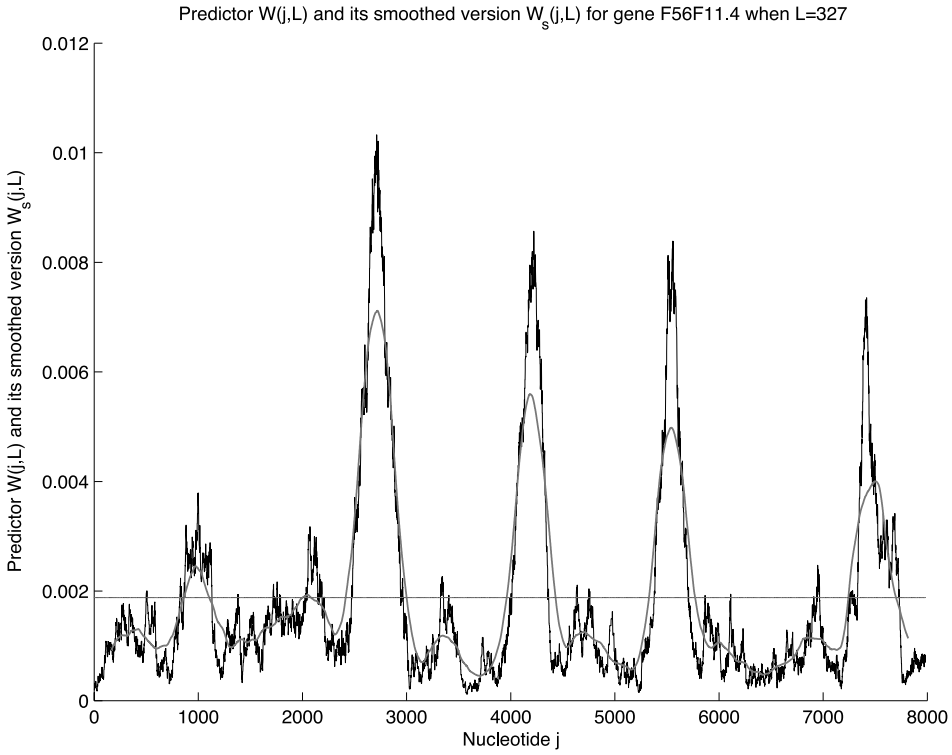

FIG. 8. Predictor $W(j, L)$ and its smoothed version $W_s(j, L)$ for gene F56F11.4. The horizontal line is the mean $W_{sm}(L)$.

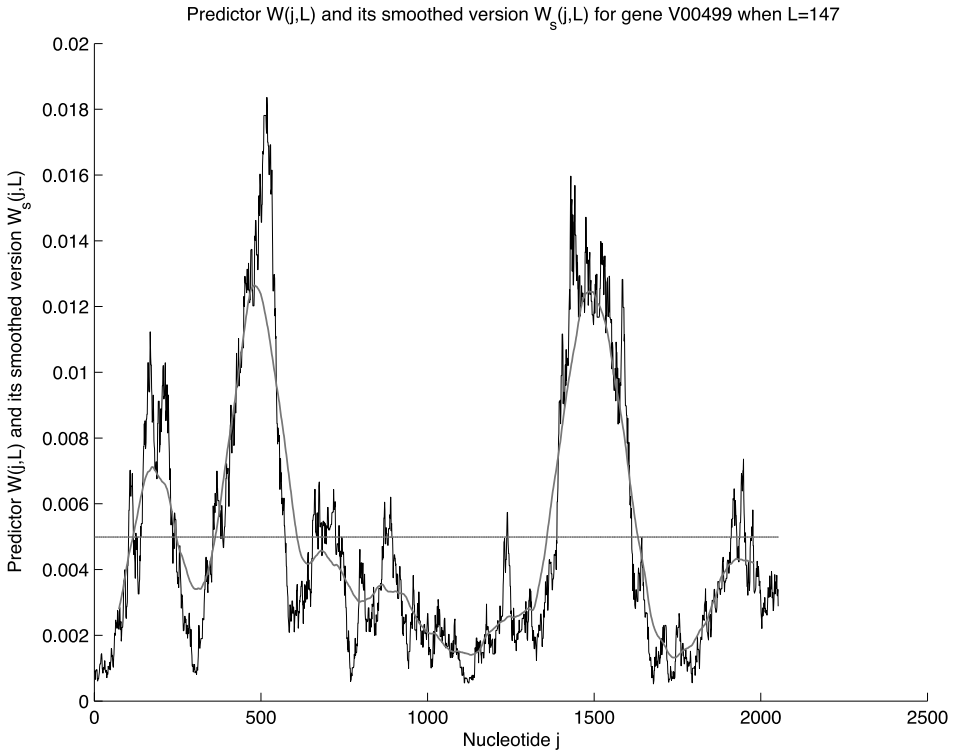

FIG. 9. Predictor $W(j, L)$ and its smoothed version $W_s(j, L)$ for gene V00499. The horizontal line is the mean $W_{sm}(L)$.

this demonstrates the consistency between $W(j, L)$ and the power spectrum S of a DNA sequence with our representation.

Anastassiou (2000) provided one method for the identification of gene coding regions. In that method, however, one needs to use complicated optimization techniques to yield the parameters a, c, g, and t. For our method, we just need to take the values of a, c, g, and t from our representation, thus our method is simpler from this point of view. In addition, if we use some of the filter techniques mentioned by Vaidyanathan and Yoon (2002), the result can be improved.

3.3. An optimized coding region predictor

Based on predictor $W(j, L)$, we can define an optimized coding region predictor as follows.

Definition 1. *Using the above defined DNA walk mode with walk length equal to 1, for a sequence with length N, we calculate all $W(j, L)$ for its sub-pieces with window size L at nucleotide position j from 1 to $N - L$, the predictor $P(L)$ is the mean of all ratios between the value of $W(j + [\frac{L}{3}] + 1, L)$ (where j is the position offset from the sequence starting point) and window size L for the same sequence; i.e.,*

$$P(L) = \frac{\sum_{j=1}^{N-L} \frac{W(j+[\frac{L}{3}]+1,L)}{L}}{N - L}, \tag{16}$$

where $[x]$ is the maximum integer less than or equal to x.

The reason to use value $W(j + [\frac{L}{3}] + 1, L)$ in our optimized predictor is based on the characteristic which is demonstrated by the above investigation on power spectrum of DNA sequences.

We also use similar statistics to evaluate the efficiency of the optimized predictor. Exons and introns are decided based on the following rule: at first, we use predictor $W(j, L)$ to get exon-pieces and intron-pieces and their corresponding boundaries. Then for each of them, we calculate their individual $P(L)$ values.

TABLE 3. STATISTICALLY EXPERIMENTAL RESULTS FOR PREDICTOR $W(j, L)$ ON *Arabidopsis thaliana* GENE CODING REGIONS

Length	N_{total}	N_p	$\frac{N_p}{N_{total}}$ (%)
0–100	2	0	0.0
101–200	11	2	18.2
201–300	62	34	54.8
301–400	30	11	36.7
401–500	40	16	40.0
501–600	21	17	81.0
601–700	37	30	81.1
701–800	30	24	80.0
801–900	40	30	75.0
901–1000	42	34	81.0
1001–1100	48	46	95.8
1101–1200	26	24	92.3
1201–1300	33	30	90.9
1301–1400	27	26	96.3
1401–1500	43	41	95.3
1501–1600	25	25	100
1601–1700	24	23	95.8
1701–1800	18	18	100
1801–1900	11	11	100
1901–2000	17	17	100
2001–3000	58	58	100
3001–4000	15	15	100
4001–5000	7	7	100
5001–6000	2	2	100
6001–7000	1	1	100
7001–8000	0	0	—
8001–9000	1	1	100

TABLE 4. STATISTICALLY EXPERIMENTAL RESULTS FOR PREDICTOR $W(j, L)$ ON HUMAN GENE CODING REGIONS

Length	N_{total}	N_p	$\frac{N_p}{N_{total}}$ (%)
0–100	16	3	18.8
101–200	28	10	35.7
201–300	18	11	61.1
301–400	32	23	71.9
401–500	70	68	97.1
501–600	17	14	82.4
601–700	11	10	90.9
701–800	16	13	81.3
801–900	9	9	100
901–1000	6	5	83.3
1001–2000	66	61	92.4
2001–3000	45	45	100
3001–4000	5	5	100
4001–5000	4	4	100
5001–6000	6	6	100
6001–7000	3	3	100

TABLE 5. $P(L)$ VALUE FOR HUMAN GERM LINE β-GLOBIN GENE

Pieces gotten from $W(j, L)$	$P(L)$ *value*	*Exon*
1–103	0.65868	No
104–245	1.18960	Yes
246–375	0.52682	No
376–598	1.64480	Yes
599–1448	0.67244	No
1449–1709	1.69860	Yes
1710–2052	0.44925	No

When the value of $P(L)$ is greater than or equal to 1, it is recognized as an exon at last; otherwise, an intron is picked out. The boundaries of introns and exons remain unchanged. By this rule, we found that each exon predicted by $W(j, L)$ can be predicted by $P(L)$; at the same time, some intron-pieces predicted by $W(j, L)$ (actually these pieces are real coding regions based on the annotation information) are predicted as exons by $P(L)$. This improvement is due to the fact that predictor $P(L)$ takes advantage of both the characteristics of significant peaks in the power spectra of coding sequences and a smoothing technology.

As an example, we show the $P(L)$ results for human germ line β-globin gene (from V00499 or gi:29440) with window width $L = 147$ and $\theta = \pi/3$ in Table 5 based on the boundaries of its seven pieces (they are either introns or exons) first decided by $W(j, L)$. We can see that the gene's three exons (relative position 104–245, 376–598, 1449–1709, respectively) are correctly picked out.

As to the improvement by using $P(L)$, there are two cases:

1. When an exon is predicted as an intron by $W(j, L)$;
2. When an intron is predicted as an exon by $W(j, L)$.

An example for the first case is as follows. Applying predictor $W(j, L)$ on human hemoglobin A beta chain A01592.1 with window width $L = 153$ and $\theta = \pi/3$, we can get the result shown in Figure 10. The gene is an exon as a whole; however, only a part of the gene is predicted and its relative position is from 178 to 321, which is different from reality. We show the $P(L)$ prediction result for A01592.1 with window width $L = 153$ and $\theta = \pi/3$ in Table 6 based on the boundaries of the three pieces first decided by $W(j, L)$. The result shows the two "introns" predicted by $W(j, L)$ are now predicted as exons by $P(L)$; the exon predicted by $W(j, L)$ coincides with the prediction result by $P(L)$. Because the three exons predicted by $P(L)$ border each other they are combined into a whole exon which corresponds to reality. In this example, we can see prediction is improved by predictor $P(L)$.

An example for the other case is as follows. Applying predictor $W(j, L)$ on human gene for delta-globin (V00505.1 or gi:30510) with window width $L = L_m = 135$ and $\theta = \pi/3$, we can get the result shown in Figure 11. The gene has three exons (relative position 123–265, 394–615, 1505–1763, respectively), which are correctly predicted by predictor $W(j, L)$. Besides the three exons, another "exon" with relative position from 993 to 1010 is also predicted by predictor $W(j, L)$, which is different from reality. We show the $P(L)$ prediction result for V00505.1 with window width $L = 135$ and $\theta = \pi/3$ in Table 7 based on the boundaries of the nine pieces first decided by $W(j, L)$. The result shows that the "exon" with relative position from 993 to 1010 predicted by predictor $W(j, L)$ is now predicted as an intron correctly by $P(L)$; the left prediction by $P(L)$ coincides with the prediction result by $W(j, L)$. In this case, prediction is also improved by predictor $P(L)$.

TABLE 6. $P(L)$ VALUE FOR A01592.1

Pieces gotten from $W(j, L)$	$P(L)$ *value*	*Exon*
1–177	1.5554	Yes
178–321	1.1581	Yes
322–438	1.1339	Yes

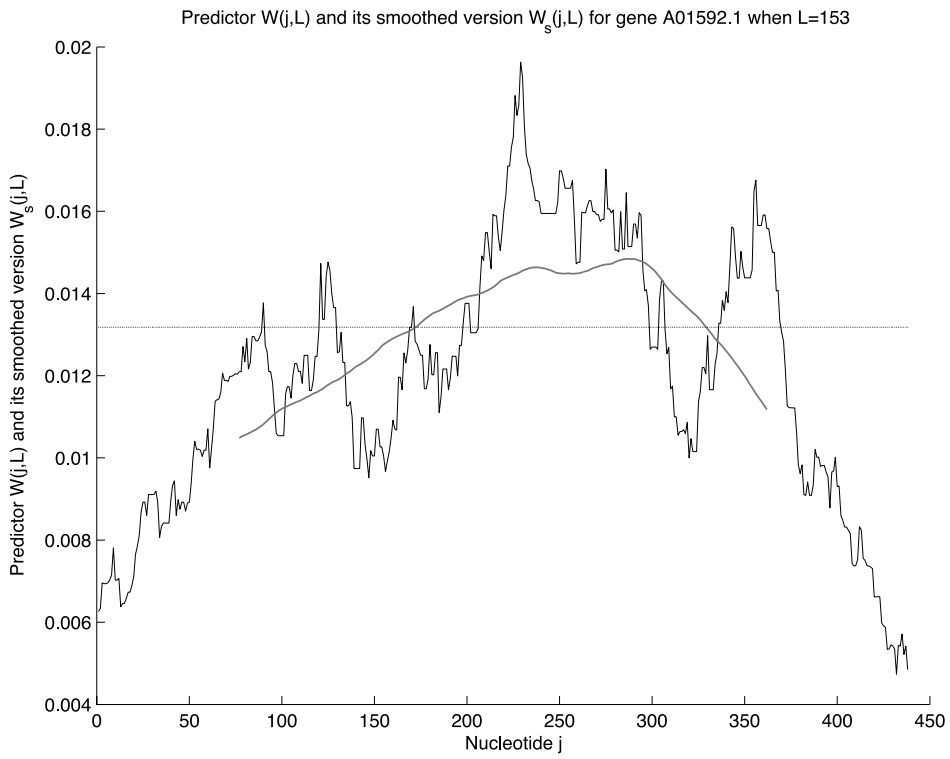

FIG. 10. Predictor $W(j, L)$ and its smoothed version $W_s(j, L)$ for gene A01592.1. The horizontal line is the mean $W_{sm}(L)$.

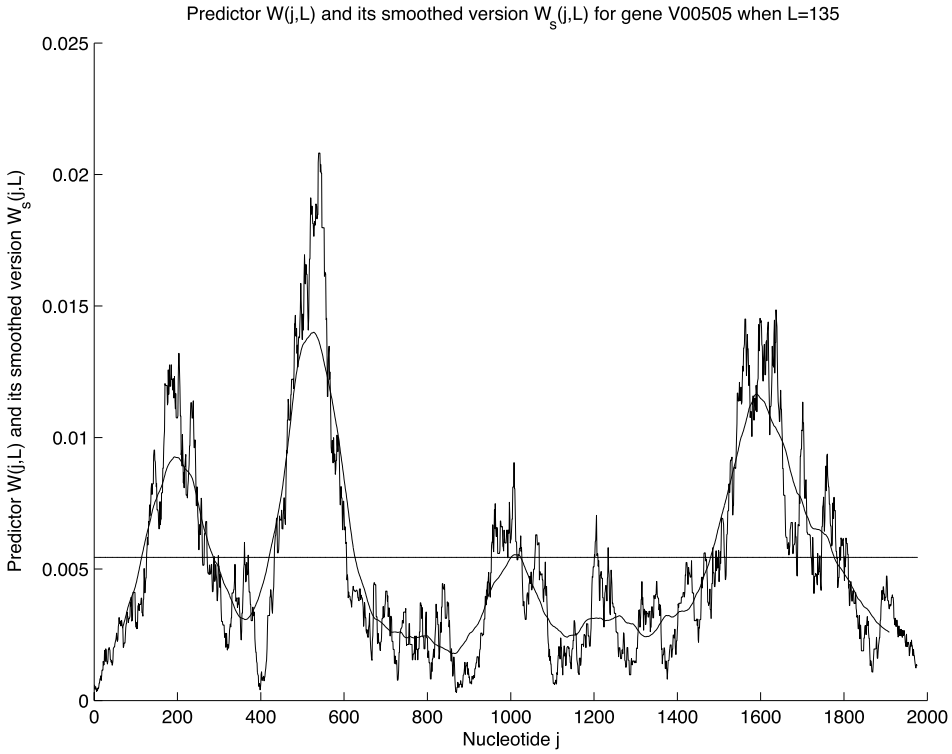

FIG. 11. Predictor $W(j, L)$ and its smoothed version $W_s(j, L)$ for gene V00505.1. The horizontal line is the mean $W_{sm}(L)$.

TABLE 7. $P(L)$ VALUE FOR V00505.1

Pieces gotten from $W(j, L)$	$P(L)$ *value*	*Exon*
1–122	0.4631	No
123–265	1.0858	Yes
266–393	0.6749	No
394–615	1.2926	Yes
616–992	0.3996	No
993–1010	0.6203	No
1011–1504	0.5480	No
1505–1763	1.5170	Yes
1764–1976	0.5859	No

The statistically experimental results for *arabidopsis thaliana* gene coding regions are shown in Table 8, where N_s is the number of coding regions with $P(L)$ value being greater than or equal to 1 from the total N_{total} coding regions within the length range shown in the left column in the table; i.e., N_s is the number of coding regions picked out by $P(L)$. The results show that for most coding regions, $P(L)$ is greater than or equal to 1; only for a small portion of them is $P(L)$ less than 1.

Another result is that for coding regions of certain lengths, the percentage of coding regions with $P(L)$ value greater than or equal to 1 is much higher than the percentage of coding regions predicted by $W(j, L)$

TABLE 8. STATISTICALLY EXPERIMENTAL RESULTS FOR PREDICTOR $P(L)$ ON *Arabidopsis thaliana* GENE CODING REGIONS

Length	N_{total}	N_s	$\frac{N_s}{N_{total}}$ *(%)*
0–100	2	1	50.0
101–200	11	5	45.5
201–300	62	60	96.8
301–400	30	21	70.0
401–500	40	28	70.0
501–600	21	18	85.7
601–700	37	33	89.2
701–800	30	29	96.7
801–900	40	38	95.0
901–1000	42	40	95.2
1001–1100	48	47	97.9
1101–1200	26	26	100
1201–1300	33	30	90.9
1301–1400	27	27	100
1401–1500	43	43	100
1501–1600	25	25	100
1601–1700	24	24	100
1701–1800	18	18	100
1801–1900	11	11	100
1901–2000	17	17	100
2001–3000	58	58	100
3001–4000	15	15	100
4001–5000	7	7	100
5001–6000	2	2	100
6001–7000	1	1	100
7001–8000	0	0	—
8001–9000	1	1	100

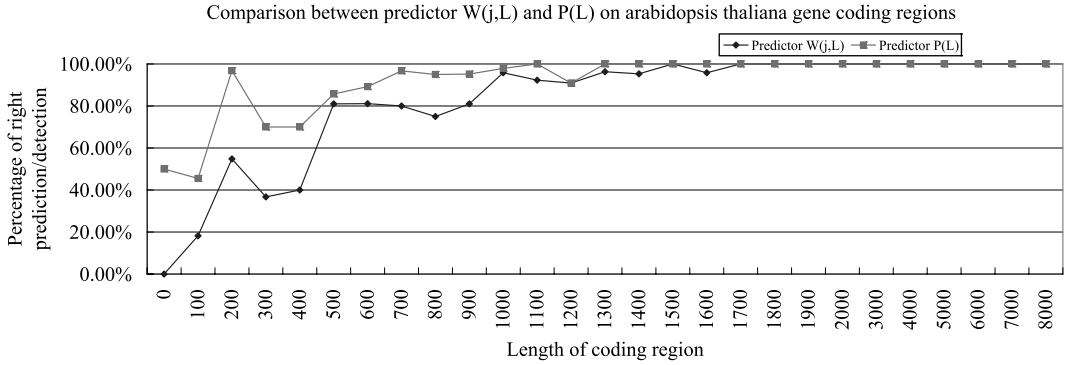

FIG. 12. Comparison between predictor $W(j, L)$ and $P(L)$ on *Arabidopsis thaliana* gene coding regions.

diagrams with the smooth technology, this can be easily seen by comparison between Tables 3 and 8, which is shown in Figure 12. At the same time, we can see that even for short coding regions, $P(L)$ has a high probability ($\geq$70%) of a value being greater than or equal to 1.

The statistically experimental results for human gene coding regions are shown in Table 9. From Table 9, we can obtain the same results as those from Table 8; i.e., for most coding regions, $P(L)$ is greater than or equal to 1; only for a small portion of them is $P(L)$ less than 1, and for coding regions of certain lengths, the percentage of coding regions for which $P(L)$ value is greater than or equal to 1 is much higher than the percentage of coding regions predicted by $W(j, L)$ diagrams with the smooth technology. This can be seen by comparing Table 4 with Table 9, which is shown in Figure 13. In the same way, even for short coding regions, the probability that $P(L)$ value is greater than or equal to 1 exceeds 80%. Thus, this is useful for the detection of coding regions.

Based on our optimization and the statistically experimental results, the predictor $P(L)$ can be used to detect coding region efficiently, and $P(L)$ is a more effective predictor than $W(j, L)$.

In both predictors, θ is a parameter. Here we choose an arbitrary value for it, which is based on two reasons. The first reason is that we find the prediction keeps stable when θ changes to any possible value. The other reason is that we have proved that our new spectrum $S(k)$ is equal quantitatively to the spectrum

TABLE 9. STATISTICALLY EXPERIMENTAL RESULTS FOR PREDICTOR $P(L)$ ON HUMAN GENE CODING REGIONS

Length	N_{total}	N_s	$\frac{N_s}{N_{total}}$ (%)
0–100	16	13	81.3
101–200	28	24	85.7
201–300	18	15	83.3
301–400	32	29	90.6
401–500	70	68	97.1
501–600	17	15	88.2
601–700	11	10	90.9
701–800	16	14	87.5
801–900	9	9	100
901–1000	6	6	100
1001–2000	66	66	100
2001–3000	45	45	100
3001–4000	5	5	100
4001–5000	4	4	100
5001–6000	6	6	100
6001–7000	3	3	100

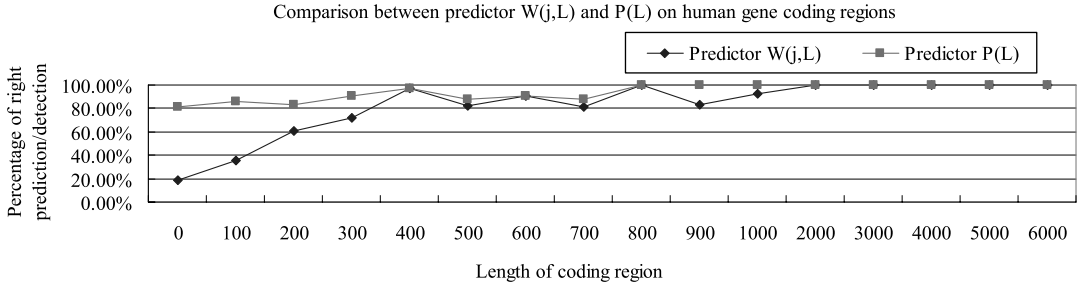

FIG. 13. Comparison between predictor $W(j, L)$ and $P(L)$ on human gene coding regions.

$PS(k)$ by Yin and Yau (2005). Naturally, evaluation of the detailed effects introduced by parameter θ will be an interesting work in the future, which is not the main topic of this article.

3.4. Performance comparison of predictor $P(L)$ and other popular methods

In order to compare the performance of predictor $P(L)$ and other popular methods, we have tested the predictor $P(L)$ on the broadly adopted ROSETTA data set of 117 homologous gene pairs (Batzoglou et al., 2000) and compared the results with those of ROSETTA, SGP-1, SGP2, SLAM, TWINSCAN, GENSCAN, and EXONSCAN. Similarly, we measure the prediction accuracy at the exon level by sensitivity S_n and specificity S_p. The two quantities are defined as follows:

$$\begin{aligned} S_n &= TP/(TP + FN), \\ S_p &= TP/(TP + FP). \end{aligned} \tag{17}$$

where *TP* (True Positives) is the number of nucleotides in coding exons predicted as coding ones, *FP* (False Positives) is the number of nucleotides in non-coding exons which are predicted as coding ones, *FN* (False Negatives) is the number of nucleotides in coding exons that are predicted as non-coding ones.

The experimental results are shown in Table 10. In the table, the results for GENSCAN, TWINSCAN, ROSETTA, SGP-1, and SLAM were retrieved from Alexandersson et al. (2003). SGP2 results were obtained from Parra et al. (2003). EXONSCAN results were gotten from Hsieh et al. (2005).

In our experiment, an exon is assumed to be correct predicted only when both of its boundaries are predicted exactly. The average of S_n and S_p summarizes the overall exon sensitivity and specificity. *ME* (Missing Exons) is the proportion of annotated exons not overlapped by any predicted exon, and *WE* (Wrong Exons) is the proportion of predicted exons not overlapped by any annotated exons.

With regard to S_n, S_p, and $(S_n + S_p)/2$, experimental results demonstrate that predictor $P(L)$ is comparable with or slightly better than all the other eight popular methods in performance. *ME* and *WE* for predictor $P(L)$ are bigger than most of those from other popular methods, so we surmise that this is due to the calculation side effects of our method. At the same time, the optimized predictor has an advantage that it does not need to be trained by a data set, thus it is simpler than other popular methods.

TABLE 10. PERFORMANCE COMPARISON ON ROSETTA SET

Program	S_n	S_p	$(S_n + S_p)/2$	*ME*	*WE*
GENSCAN	0.82	0.77	0.79	0.06	0.11
TWINSCAN	0.85	0.77	0.80	0.03	0.12
SGP2 (single)	0.84	0.85	0.84	0.05	0.03
SGP2 (multiple)	0.71	0.79	0.75	0.12	0.03
ROSETTA	0.83	0.83	0.83	0.05	0.05
SGP-1	0.70	0.76	0.73	0.12	0.04
SLAM	0.78	0.76	0.77	0.04	0.06
EXONSCAN	0.87	0.89	0.88	0.04	0.03
$P(L)$	0.83	0.86	0.84	0.08	0.07

One might argue why we do not use EGASP data as a benchmark (EGASP stands for ENCODE GASP, which was inspired by the Genome Annotation Assessment Project [Guigó and Reese, 2005]) to test our predictor. This is attributable firstly to the fact that we are exploring a good representation method and basing on it to characterize coding regions; on the contrary, ENCODE project has different objectives (Encode). Secondly, because our adopted test data are broadly accepted and used by many experts in their rich methods recently (Alexandersson et al., 2003; Parra et al., 2003; Hsieh et al., 2005), and we can get the experimental results on them easily for performance comparison. At this moment, we consider that the data set used is sufficient to support our claims. However, we agree that ideally a larger data set will be needed to provide evidence for stronger claims.

4. CONCLUSION

Graphical representation of DNA sequence provides a simple and intuitive way of viewing, anchoring, and comparing various gene structures, so a simple and non-degenerate graphical representation is attractive to both biologists and computational biologists.

The two-dimensional graphical representation of DNA sequences presented in this study, which is simple and does not cause degeneracy, is a universal version of the method presented by Yau et al. (2003). We introduce some interesting characteristics of the representation and their possible use.

We have done some statistical experiments on the frequency analysis of human and *arabidopsis thaliana* gene coding regions that show there is a high possibility of a significant peak appearing at frequency $k = N/3$ in their spectra based on our graphical representation. These results demonstrate possible applications of our representation in coding region detection or prediction. We have also performed two kinds of frequency analysis experiments, i.e., the DNA walk and random mode, on coding region pieces and non-coding region pieces. The results show that a significant peak appears at the frequency $k = N/3$ in their spectra for coding regions but such phenomena is absent for pieces of non-coding regions.

Based on the frequency analysis, we provide an efficient predictor $W(j, L)$ for gene coding regions; our method is simpler than those of Anastassiou (2000). Unlike the method of Anastassiou which uses optimization techniques to yield parameters for his predictor, the parameters of our predictor are naturally given by our graphical representation. Results from statistical experiments demonstrate the consistency between $W(j, L)$ and the power spectra of DNA sequences with our representation.

An optimized predictor $P(L)$ based on our universal representation is provided for the detection of coding regions. The statistically experimental results have shown that it is more effective for the detection than those using predictor $W(j, L)$.

Experimentation on a broadly accepted ROSETTA data set demonstrates that the optimized predictor $P(L)$ is also comparable with other popular methods such as GENSCAN, TWINSCAN, ROSETTA, SGP-1, SLAM, and EXONSCAN in performance. Unlike most other methods, our method does not require the train data.

In addition, in our method we can select the most preferable unit vectors to represent the four nucleotides. Therefore, it is an efficient and flexible approach to analyze DNA sequences for both computational scientists and molecular biologists.

ACKNOWLEDGMENTS

This work is supported by NSFC grant 60403025 (to Prof. Xianyang Jiang) and PRA SI04-04 (to Prof. Xianyang Jiang and Prof. Lavenier). The authors also thank the associated editor and the anonymous reviewers for their valuable suggestions.

DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Alexandersson, M., Cawley, S., and Pachter, L. 2003. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* 13, 496–502.

Anastassiou, D. 2000. Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 16, 1073–1081.

Batzoglou, S., Pachter, L., Mesirov, J., et al. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. *Proc. 4th Annu. Int. Conf. Comput. Mol. Biol.* 46–53.

Berthelsen, C.L., Glazier, J.A., and Skolnick, M.H. 1992. Global fractal dimension of human DNA sequences treated as pseudorandom walks, *Phys. Rev. A* 45, 8902–8913.

C. Mathé, T.S., Sagot, M. F., Rouzé, P., et al. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103–4117.

Cheever, E.A., Searls, D.B., Karunaratne, W., et al. 1989. Using signal processing techniques for DNA sequence comparison. *Proc. 15th Annu. Northeast Bioeng. Conf.* 173–174.

Encode. Available at: *www.genome.gov/10005107*. Accessed August 1, 2008.

Gates, M.A. 1985. Simpler DNA sequence representations. *Nature* 316, 219.

Gates, M.A. 1986. A simple way to look at DNA. *J. Theor. Biol.* 119, 319–328.

Guigó, R., and Reese, M.G. 2005. EGASP: collaboration through competition to find human genes. *Nat. Methods* 2, 575–577.

Hamori, E. 1985. Novel DNA sequence representations. *Nature* 314, 585–586.

Hamori, E. 1994. *Visualization of Biological Information Encoded in DNA*. Wiley, New York.

Hamori, E., and Ruskin, J. 1983. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* 258, 1318–1327.

Hsieh, S.J., Chung, Y.S., Lin, C.-Y, et al. 2005. EXONSCAN: exon prediction with signal detection and coding region alignment in homologous sequences. *Proc. 2005 ACM Symp. Appl. Comput.* 202–203.

Jeffrey, H.J. 1990. Chaos game representation of gene structure. *Nucleic Acids Res.* 18, 2163–2170.

Kotlar, D., and Lavner, Y. 2003. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Res.* 13, 1930–1937.

Li, W., and Holste, D. 2005. Universal $1/f$ noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. *Phys. Rev. E* 71, p. 041910.

Parra, G., Agarwal, P., Abril, J.F., et al. 2003. Comparative gene prediction in human and mouse. *Genome Res.* 13, 108–117.

Su, S.-C., Yeh, C.H., and Kuo, C.-C. 2003. Structural analysis of genomic sequences with matched filtering. *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 3, 2893–2896.

Tiňo, P. 1999. Spatial representation of symbolic sequences through iterative function systems. *IEEE Trans. Syst. Man Cybernet. Part A* 29, 386–393.

Vaidyanathan, P.P., and Yoon, B.-J. 2002. Digital filters for gene prediction applications. *Proc. 36th Asilomar Conf. Signals Syst. Comput.* 1, 306–310.

Voss, R.F. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68, 3805–3808.

Wang, W., and Johnson, D.H. 2002. Computing linear transforms of symbolic signals. *IEEE Trans. Signal Process.* 50, 628–634.

Wu, D., Robergé, J., Cork, D.J., et al. 1993. Computer visualization of long genomic sequences. *Proc. 4th Conf. Visualization '93* 308–315.

Yau, S.S.T., Wang, J., Niknejad, A., et al. 2003. DNA sequence representation without degeneracy. *Nucleic Acids Res.* 31, 3078–3080.

Yin, C., and Yau, S.S.T. 2005. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. *J. Comput. Biol.* 12, 1153–1165.

Zhang, R., and Zhang, C.T. 1994. Z-curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dynamics* 11, 767–782.

Address reprint requests to:
Dr. Stephen S.-T. Yau
Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607-7045

E-mail: yau@uic.edu