JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 9, 2005

Pp. 1153–1165

A Fourier Characteristic of Coding Sequences: Origins and a Non-Fourier Approximation

CHANGCHUAN YIN and STEPHEN S.-T. YAU

ABSTRACT

The 3-base periodicity, identified as a pronounced peak at the frequency $N/3$ (N is the length of the DNA sequence) of the Fourier power spectrum of protein coding regions, is used as a marker in gene-finding algorithms to distinguish protein coding regions (exons) and noncoding regions (introns) of genomes. In this paper, we reveal the explanation of this phenomenon which results from a nonuniform distribution of nucleotides in the three coding positions. There is a linear correlation between the nucleotide distributions in the three codon positions and the power spectrum at the frequency $N/3$. Furthermore, this study indicates the relationship between the length of a DNA sequence and the variance of nucleotide distributions and the average Fourier power spectrum, which is the noise signal in gene-finding methods. The results presented in this paper provide an efficient way to compute the Fourier power spectrum at $N/3$ and the noise signal in gene-finding methods by calculating the nucleotide distributions in the three codon positions.

Key words: Fourier spectral analysis, 3-base periodicity, genetic codon.

1. INTRODUCTION

THE INITIAL STEP IN GENOMIC ANNOTATION is to identify protein coding regions of the genomes, which is a challenging problem especially in the study of eukaryote genomes. In an eukaryote genome annotation, protein coding regions (exons) are usually not continuous, and they are interrupted by noncoding regions (introns). It is difficult to distinguish protein coding regions from noncoding regions by the sequences since there is no obvious sequence feature between exons and introns.

To tackle this problem, a variety of computational techniques have been developed such as neural networks, Markov models, and Fourier power spectrum analysis (Tiwari *et al.*, 1997; Anastassiou, 2000). During the last decade, more and more efforts have been placed on Fourier analysis in gene-finding algorithms which are based on the well-recognized fact that an exon DNA sequence with length N has a 3-base periodicity property, which is identified as a pronounced peak at the frequency $N/3$ of the Fourier power spectrum, but the 3-base periodicity does not exist in most of intron sequences (Fickett and Tung, 1992). Knowledge of the origin of the 3-base periodicity in protein coding regions will provide useful information for designing more effective gene-finding algorithms and understanding the other periodicities

Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607-7045.

within the genomes. Tiwari *et al.* (1997) showed that the 3-base periodicity is not caused by the gene codons bias in protein coding regions. However, the exact origin of the 3-base periodicity is still an open question.

Stochastic simulation is a powerful approach to study many events and has been used to investigate many biological systems. This paper presents the stochastic simulation of DNA sequences to reveal the origins of the 3-base periodicity in protein coding regions of DNA sequences and the noise signal in the Fourier power spectrum.

2. SYSTEMS AND METHODS

2.1. Numerical representations of DNA sequences

DNA molecules are composed of four linearly linked nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G). A DNA sequence can be represented as a permutation of four characters A, T, C, G at different lengths. The character strings of DNA molecules are mapped into one or more numerical sequences (Voss, 1992; Tiwari *et al.*, 1997; Yau *et al.*, 2003). One method found in the literature is to use binary indicator sequences. Consider a DNA sequence denoted as $x(0), x(1), \ldots, x(N-1)$ that can be decomposed into four binary indicator sequences, $u_A(n)$, $u_T(n)$, $u_C(n)$, and $u_G(n)$ which indicate the presence or absence of four nucleotides, A, T, C, and G at the nth position, respectively (Tiwari *et al.*, 1997). For example, the indicator sequence $u_A(n) = 0001010111\ldots$ indicates that the nucleotide A presents in positions 4, 6, 8, 9, and 10 of the DNA sequence.

2.2. Discrete Fourier transform (DFT) power spectrum analysis

The Discrete Fourier Transform (DFT) converts a signal in time domain to a set of new values in the frequency domain. The DFT of a signal of length N, $f(n), n = 0, 1, \ldots, N-1$, at frequency k is defined as follows

$$F(k) = \sum_{n=0}^{N-1} f(n) e^{-i\frac{2\pi}{N}kn} \tag{2.1}$$

where $i = \sqrt{-1}$. The DFT power spectrum of a signal at frequency k is defined as

$$PS(k) = |F(k)|^2, k = 0, 1, 2, \ldots, N-1 \tag{2.2}$$

where $F[k]$ is the kth DFT coefficient. The DFT power spectrum analysis is often used to find the frequency components of a signal buried in a noisy time domain signal. Consider a signal s containing 50 Hz and 120 Hz frequencies and being corrupted with some zero-mean random noise:

$$s = sin(2 * \pi * 50 * t) + sin(2 * \pi * 120 * t) + random. \tag{2.3}$$

It is difficult to identify the two frequency components, 50 Hz and 120 Hz, in the original time domain signal (Fig. 1(a)), but these two frequency components can be detected after the signal is converted to the frequency domain by the DFT (Fig. 1(b)).

For DFT power spectrum analysis of a DNA sequence, the total Fourier power spectral content of a DNA sequence is the sum of power spectrum of its four binary indicator sequences (Tiwar *et al.*, 1997):

$$PS(k) = PS_A(k) + PS_T(k) + PS_C(k) + PS_G(k) \tag{2.4}$$

where $PS_A(k)$, $PS_T(k)$, $PS_C(k)$, and $PS_G(k)$ are the Fourier power spectrum of the four indicator sequences $u_A(n)$, $u_T(n)$, $u_C(n)$, and $u_G(n)$, respectively. Due to the symmetry property of the DFT spectrum of real number signals, all figures of DFT spectra in this paper show only the first half of the original figures. The average power spectrum over all the frequencies is considered as noise background in gene-finding methods.

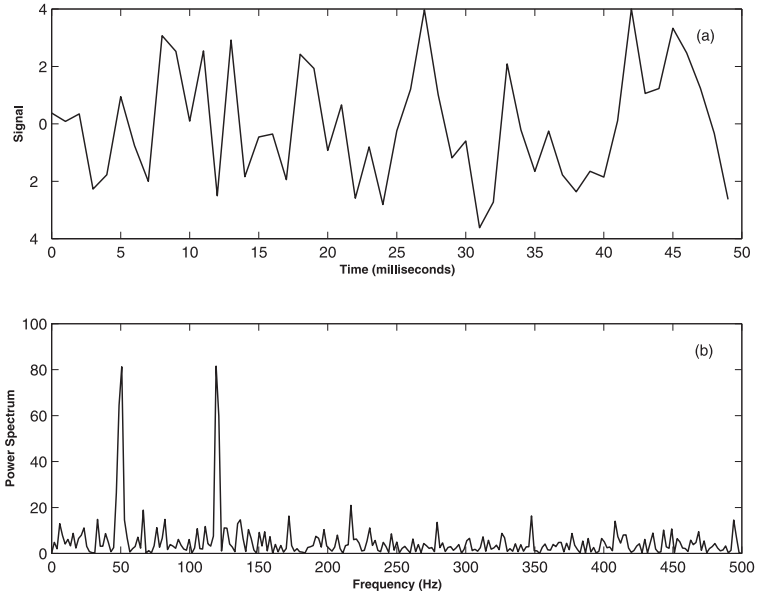

FIG. 1. DFT power spectra analysis of a test signal. (**a**) Plotting the test signal versus time t. (**b**) Plotting DFT power spectrum versus frequency.

Let E denote the average power spectrum over all frequencies, then

$$E = \frac{1}{N}\sum_{f=0}^{N-1} PS(f). \tag{2.5}$$

2.3. Stochastic simulations

Stochastic simulation programs, implemented using the MATLAB software package, generate artificial DNA sequences to demonstrate the 3-base periodicity according to specific purposes described in the text. If there is no explicit specification, the artificial DNA sequences used in the experiment are those with different lengths and with any possible occurrence probabilities on the first, second, and third codon positions. The simulation is also performed on the protein coding region of the *GroEL* gene from the *Escherichia coli* genome (GeneBank access number NC000913). Similar tests are performed on other natural DNA sequences.

2.4. Database

Exon and intron sequences used for the statistics study are from the Xpro database (Gopalan *et al.*, 2004), which contains annotated eukaryotic exon and intron sequences from GeneBank release 139. The dataset was downloaded from the Xpro web site as flat files (Xpro version v.1.2, 2004, *www.origin.bic.nus.edu.sg/xpro*) and parsed to exon or intron sequences based on the header information of each sequence entry.

3. RESULTS AND DISCUSSION

3.1. The 3-base periodicity is not determined by the genetic codon bias usages in DNA sequences

Before the stochastic simulation study, the 3-base periodicity in the *GroEL* gene is verified as a pronounced peak in the DFT power spectrum at frequency $N/3$ (where $N = 1{,}647$ is the length of the gene, Fig. 2(a)), but it does not appear in an intron or a uniformly random DNA sequence (as will be shown later, in Figs. 8(a) and 9(b)).

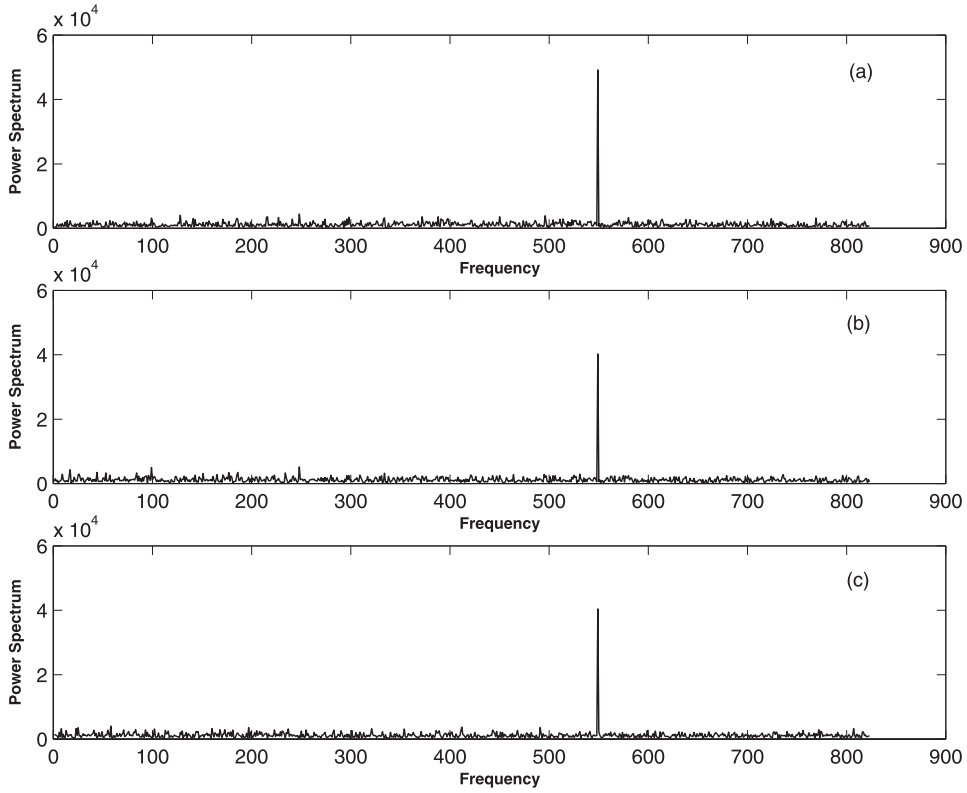

FIG. 2. DFT power spectra of artificial gene sequences. (**a**) Original *GroEL* gene. (**b**) Gene sequence with same amino acid sequence of *GroEL*, but with uniform codon usage for an amino acid. (**c**) DNA sequence from a perfect permutation of genetic codons of *GroEL* gene.

Tiwari *et al.* (1997) demonstrated that 3-base periodicity was not affected by the genetic codon bias. To make the complete simulation, we create an artificial gene sequence that encodes the same protein sequence as *GroEL* in which each codon for an amino acid has an equal probability of occurrence and then perform DFT analysis on this artificial gene sequence. The result, shown in Fig. 2(b), indicates that the 3-base periodicity is not affected by the codon bias, validating the conclusion of Tiwari *et al.* (1997).

3.2. The 3-base periodicity is affected by the amino acid compositions, but not by the ordering of the amino acids encoded by the DNA sequence

To investigate whether the special ordering of the amino acids encoded by the gene contributes to the 3-base periodicity, we generate a new *GroE* gene by a perfect permutation. By *perfect permutation*, we mean the amino acid sequence of a gene is randomly shuffled, followed by reverse translation of the amino acid sequence to a gene sequence, which is done by randomly picking one of its codons. Figure 2(c) shows the 3-base periodicity is still present in this sequence of shuffled codons, implying that amino acid composition, not the ordering of the amino acids in proteins, determines the 3-base periodicity. This result indicates that the codon frequencies, or the nucleotide distribution within the codons, plays an important role in the determination of the 3-base periodicity in protein coding regions.

3.3. The 3-base periodicity of a DNA sequence is determined by the unbalanced nucleotide distributions of the three codon positions

To investigate the relationship between the nucleotide distributions and the 3-base periodicity of a DNA sequence, we calculate the nucleotide frequencies of the first codon positions: $1, 4, 7, \ldots, 3m + 1$, or

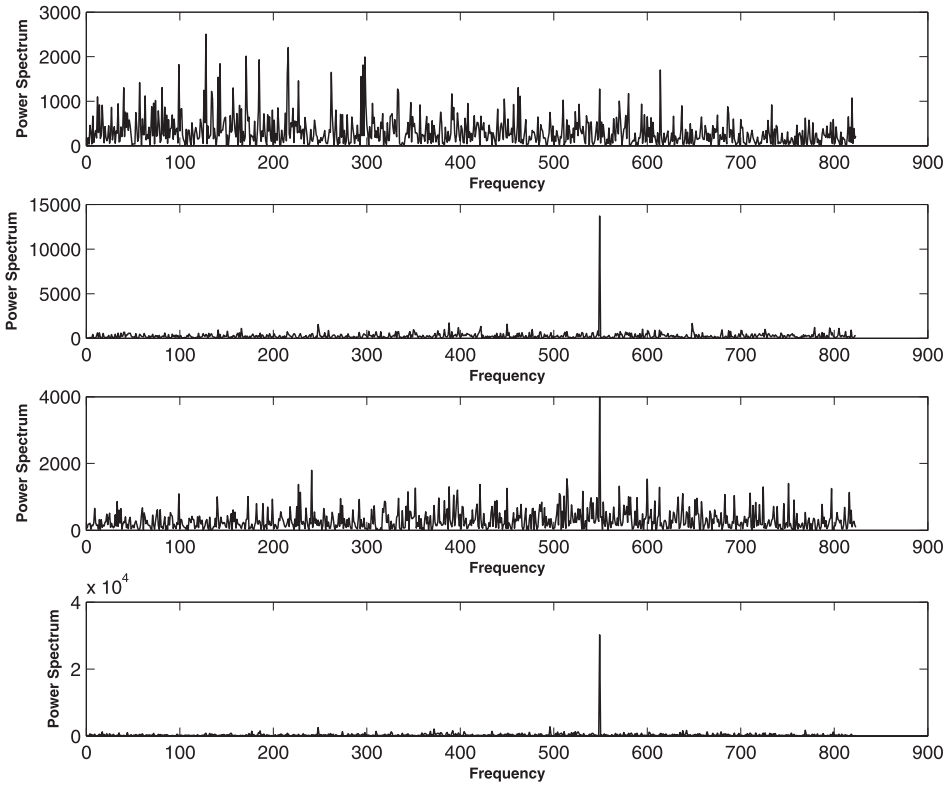

FIG. 3. DFT power spectra of four indicator sequences of *GroEL* gene. (**a**) DFT power spectrum of indicator sequence $u_A(n)$. (**b**) DFT power spectrum of indicator $u_T(n)$. (**c**) DFT power spectrum of indicator $u_C(n)$. (**d**) DFT power spectrum of indicator sequence $u_G(n)$.

the second codon positions: $2, 5, 8, \ldots, 3m+2$, or the third codon positions: $3, 6, 9, \ldots, 3m+3$ ($m = 0, 1, 2, \ldots, N/3$) of the *GroEL* gene and conduct the correlation study between the nucleotide distribution and the 3-base periodicity. Let F_{Xi} denote the occurrence frequency of a nucleotide $\mathbf{x}$ ($\mathbf{x} \in (A, T, C, G)$) in the ith ($i = 1, 2, 3$) codon positions of a DNA sequence; then the nucleotide frequencies in *GroEL* gene are

$$\begin{pmatrix} F_{A1} & F_{A2} & F_{A3} \\ F_{T1} & F_{T2} & F_{T3} \\ F_{C1} & F_{C2} & F_{C3} \\ F_{G1} & F_{G2} & F_{G3} \end{pmatrix} = \begin{pmatrix} 148 & 165 & 124 \\ 36 & 161 & 143 \\ 92 & 138 & 164 \\ 273 & 85 & 118 \end{pmatrix}.$$

This shows that nucleotide A is distributed in the three codon positions approximately equally in this gene, but nucleotides, T, C, and G have nonuniform distribution in the three codon positions. Figure 3 is the power spectra of the four indicator sequences of the *GroEL* gene. The indicator sequence for nucleotide A does not have a peak at frequency $N/3$, while all the indicator sequences for T, C, and G do have a pronounced peak. These results imply that the unbalanced nucleotide distributions on the three codon positions contribute to the 3-base periodicity in the indicator sequences. To validate this assumption, we generate an artificial DNA sequence with the following probabilities for nucleotides on the three codon positions. Let P_{Xi} denote the probability of nucleotide X on the ith positions:

$$\begin{pmatrix} P_{A1} & P_{A2} & P_{A3} \\ P_{T1} & P_{T2} & P_{T3} \\ P_{C1} & P_{C2} & P_{C3} \\ P_{G1} & P_{G2} & P_{G3} \end{pmatrix} = \begin{pmatrix} 0.10 & 0.25 & 0.25 \\ 0.25 & 0.40 & 0.10 \\ 0.25 & 0.25 & 0.25 \\ 0.40 & 0.10 & 0.40 \end{pmatrix}.$$

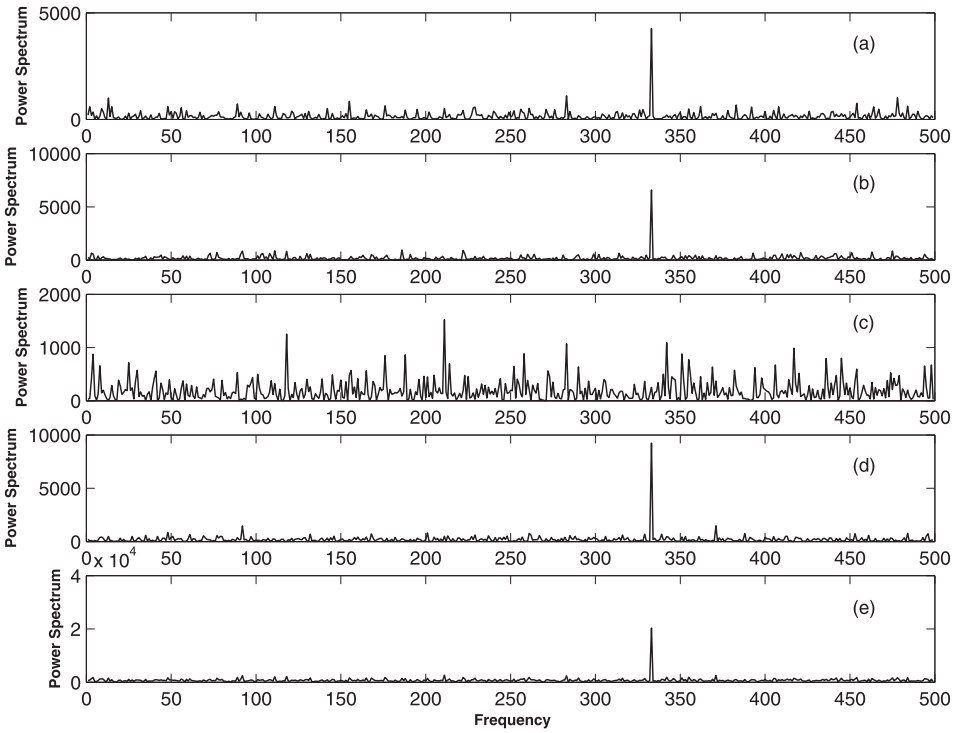

FIG. 4. DFT power spectra analysis of an artificial DNA sequence with the probabilities of nucleotides of the three codon position as described in text. (**a**) DFT power spectrum of indicator sequence $u_A(n)$. (**b**) DFT power spectrum of indicator $u_T(n)$. (**c**) DFT power spectrum of indicator $u_C(n)$. (**d**) DFT power spectrum of indicator sequence $u_G(n)$. (**e**) DFT power spectrum of the DNA sequence, which is the sum of the power spectra of the four indicator sequences.

The DFT power spectrum of this artificial DNA sequence in Fig. 4 shows that there are clear peaks at frequency $N/3$ for indicator sequences $u_A(n)$, $u_T(n)$, and $u_G(n)$, all of which have unbalanced nucleotide distributions (Fig. 4(a), (b), (d)). But the 3-base periodicity does not exist in the indicator sequence $u_C(n)$ (Fig. 4(c)). The reason for the absence of the 3-periodicity in this indicator sequence is that the distribution of nucleotide C on the three codon positions is uniform, which does not contribute to the 3-base periodicity. Since the DFT power spectrum of the DNA sequence is the sum of the power spectra of four indicator sequences, there is a pronounced peak at frequency $N/3$. A conclusion can be obtained: when a nucleotide has an unbalanced distribution in the three codon positions, there will be a 3-base periodicity in the corresponding indicator sequence. The more detailed relationship between nucleotide distributions and the 3-base periodicity is investigated in following sections.

3.4. The power spectrum at frequency $N/3$, $PS(N/3)$ of a DNA sequence depends on the variance of nucleotide distributions in the three codon positions

Let F_{x1}, F_{x2}, F_{x3} be the occurrence frequencies of the nucleotide $\mathbf{x} \in (A, T, C, G)$ in the first codon position, the second codon position, and the third codon position, respectively. The variance of the nucleotide distribution in the three codon positions, which is a measurement of the unbalanced distribution of this nucleotide distribution in the three codon positions, is believed to contribute the 3-base periodicity in DNA sequences. The nucleotide x distribution in the three codon positions can be represented as the variance of frequencies of the nucleotide x in the three codon positions:

$$\sigma_x = \sum_{i=1,2,3} \left(F_{xi} - \frac{1}{3} \sum_{j=1,2,3} F_{xj} \right)^2 . \tag{3.1}$$

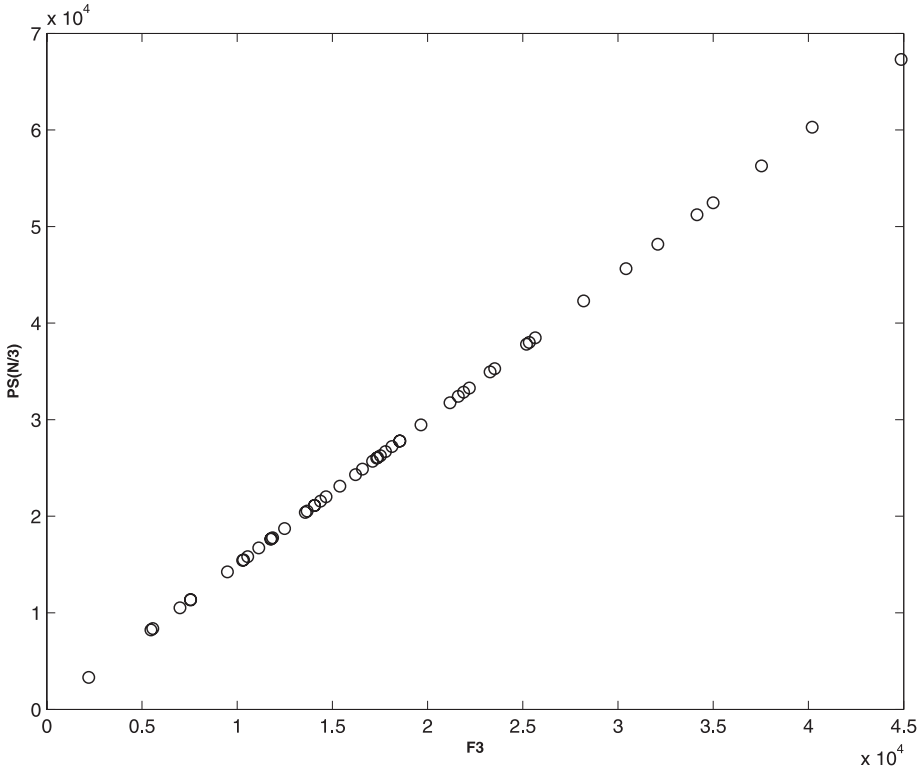

FIG. 5. The correlation between nucleotide distribution of three codon positions and the power spectrum at $N/3$. The DNA sequences used in the experiment are those with different lengths and with any possible occurrence probabilities on the first, second, and third codon positions.

The total measurement of unbalance of the nucleotide distributions in the three codon positions, F_3, can be represented as the sum of the variances of each nucleotide on the three codon positions, i.e.,

$$F_3 = \sum_{x=A,T,C,G} \sigma_x \tag{3.2}$$

and equally,

$$F_3 = \sum_{x=A,T,C,G} \sum_{i=1,2,3} \left(F_{xi} - \frac{1}{3} \sum_{j=1,2,3} F_{xj} \right)^2 . \tag{3.3}$$

Equation (3.3) can be simplified as

$$F_3 = \frac{2}{3} \sum_{x=A,T,C,G} (F_{x1}^2 + F_{x2}^2 + F_{x3}^2 - (F_{x1} * F_{x2} + F_{x1} * F_{x3} + F_{x2} * F_{x3})). \tag{3.4}$$

To validate the correlation between the frequency variance of the three codon positions and the value of the power spectrum at frequency $N/3$ of a DNA sequence, we plot the power spectrum coefficients at frequency $N/3$ versus the corresponding frequencies F_3 of the test DNA sequences. The result, shown in Fig. 5, indicates an accurate linear correlation between $PS(N/3)$ and F_3. The slope of the line is $\frac{3}{2}$. Thus, from the simulation data, we can get a simple and fast calculation method for the power spectrum at $N/3$:

$$PS(N/3) = \frac{3}{2} F_3 \tag{3.5}$$

and equally,

$$PS(N/3) = \sum_{x=A,T,C,G} (F_{x1}^2 + F_{x2}^2 + F_{x3}^2 - (F_{x1} * F_{x2} + F_{x1} * F_{x3} + F_{x2} * F_{x3})) \tag{3.6}$$

where F_3 is described in formula (3.2).

3.5. *The average power spectrum of a DNA sequence depends only on the length of the DNA sequence and the variance of nucleotide distributions of all the codon positions*

The ratio of the Fourier power spectrum at the frequency $N/3$ to the average power spectrum over all the frequencies (formula (2.5)), denoted by R, is considered as the signal-to-noise ratio in identification of protein coding regions. Regions of a DNA sequence having a value of R greater or equal to a certain threshold value can be identified as protein-coding regions.

We conduct computational experiments to investigate whether there exists a relationship between the average power spectrum and the nucleotide distribution on each codon position. Since F_{xi} is the frequency of nucleotide x at ith codon position, then obviously, $F_{Ai} + F_{Ti} + F_{Ci} + F_{Gi} = \frac{N}{3}$, $i = 1, 2, 3$. Let D_i denote the distribution of the four nucleotides at the ith codon position as follows:

$$D_i = \sum_{x=A,T,C,G} \left(F_{xi} - \frac{N}{3}\right)^2. \tag{3.7}$$

The nucleotide distribution of all the three codon positions can then be represented as F_c,

$$F_c = \sum_{i=1,2,3} \sum_{x=A,T,C,G} \left(F_{xi} - \frac{N}{3}\right)^2, \tag{3.8}$$

which is equal to the following equation:

$$F_c = \frac{3}{4}N^2 + \sigma_F \tag{3.9}$$

where σ_F is the variance of nucleotide frequencies of all the three codon positions, i.e.,

$$\begin{aligned}\sigma_F &= \sum_{i=1,2,3} \sum_{x=A,T,C,G} \left(F_{xi} - \frac{1}{12} \sum_{x=A,T,C,G} \sum_{j=1,2,3} F_{xj}\right)^2 \\ &= \sum_{i=1,2,3} \sum_{x=A,T,C,G} \left(F_{xi} - \frac{N}{12}\right)^2.\end{aligned}$$

Figure 6 is the plot of average power spectrum E and the frequency distributions F_c of different length DNA sequences. It demonstrates a correlation between the average power spectrum E and the nucleotide distributions F_c, which is a function of the length of a DNA sequence and the variance of nucleotide frequencies. The relationship between the average power spectrum E of a DNA sequence and the total nucleotide distribution at all three codon positions of the DNA sequence can be represented as

$$E \approx 0.8550\sqrt{F_c}. \tag{3.10}$$

Thus, the average power spectrum E can be simply calculated using the formula by frequency distributions in the DNA sequence. It should be noted that since the variance of the frequencies is much smaller than the length of the DNA sequence, the length of the DNA sequence is the dominant term in the formula (3.9); the longer a DNA sequence, the higher the noise signal in view of formulas (3.9) and (3.10). Thus, the average power spectrum of a DNA sequence depends on the length of the DNA sequence and the variance of nucleotide distributions on all the codon positions.

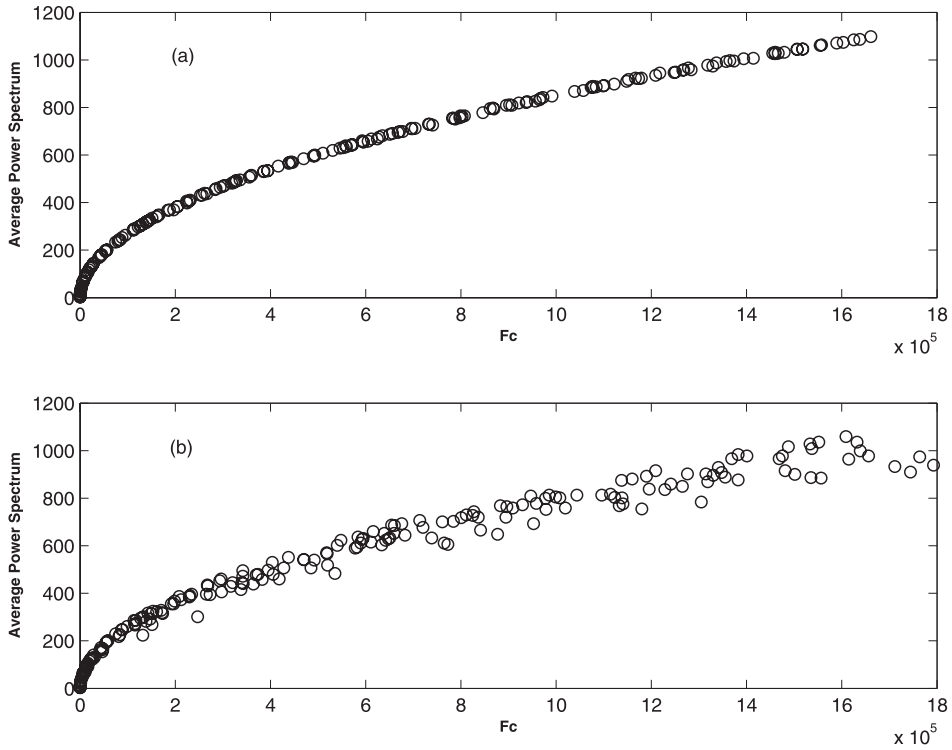

FIG. 6. The correlation between the average Fourier power spectrum and the nucleotide distributions. (**a**) DNA sequences of different lengths generated with the following occurrence probabilities used in Fig. 4, i.e., $P_{A1} = 0.10$, $P_{A2} = 0.25$, $P_{A3} = 0.25$, $P_{T1} = 0.25$, $P_{T2} = 0.40$, $P_{T3} = 0.10$, $P_{C1} = 0.25$, $P_{C2} = 0.25$, $P_{C3} = 0.25$, $P_{G1} = 0.40$, $P_{G2} = 0.10$, $P_{G3} = 0.40$. (**b**) The DNA sequences with different lengths and with any possible occurrence probabilities at the first, second, and third codon positions.

To find out whether if the nucleotide distribution determines the signal-to-noise ratio in a DNA sequence, the signal-to-noise ratios are calculated for different artificial DNA sequences with different lengths and with any possible occurrence probabilities at the first, second, and third codon positions. Figure 7 is the plot of the signal-to-noise ratio to the nucleotide distribution (F_3). This result indicates that there is a correlation between the signal-to-noise ratio and F_3: the higher F_3, the higher the signal-to-noise ratio. Thus, the nucleotide distribution not only determines the 3-base periodicity signal, but also basically determines the signal-to-noise ratio. Formula (3.1) obviously explains the reasons why the excess of a very low amount of a nucleotide in the first codon positions can cause the 3-base periodicity of a DNA sequence.

3.6. Case studies

We compute the Fourier power spectrum and nucleotide frequencies of the three codon positions for different test DNA sequences from different organisms. Table 1 lists the results for DNA sequences computed using the direct DFT method and the formulas in this study. The data in the table shows that $PS(N/3)$ is equal to $\frac{3}{2}F_3$. The signal-to-noise ratios, R_1 and R_2, are verified to be very similar. The signal-to-noise ratios in exons are usually larger than threshold value 2 while the ratios are less than 2 in introns. If R_1 or R_2 is larger than 2 in a given DNA sequence, the 3-base periodicity peak is visible among the signal noise, and then this DNA sequence can be a potential exon region. From these case studies, we can directly compute $PS(N/3)$ and the signal-to-noise ratio of a DNA sequence using the nucleotide distributions on the three codon positions. Thus, both simulation data and test DNA data indicate that an unbalanced frequency distribution on the three codon positions contributes to the 3-base periodicity in the DNA sequence. To study the feasibility of applying the formulas in identification of exon and intron sequences, we perform a statistics study of different exon and intron sequences of the human genome from the Xpro database.

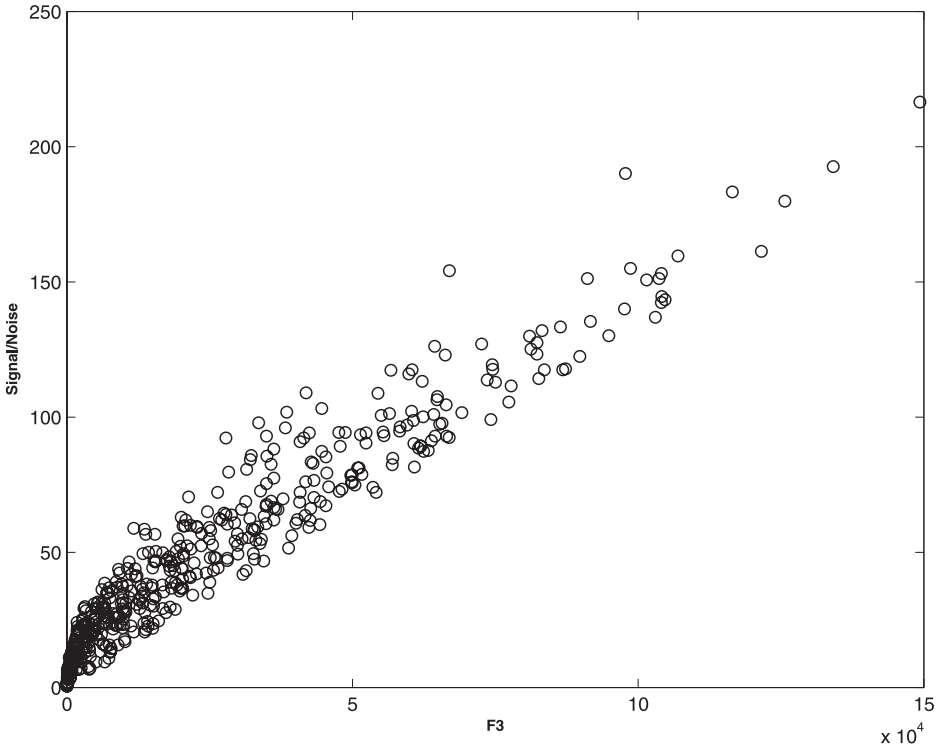

FIG. 7. The relationship of signal to noise and the nucleotide distribution (F_3) on the three codon positions.

TABLE 1. CASE STUDIES FOR THE 3-BASE PERIODICITY AND NUCLEOTIDE DISTRIBUTIONS[a]

Gene	*Organism*	*Location*	N	$PS(N/3)$	$\frac{3}{2}F_3$	E_1	E_2	R_1	R_2
1J942	*C. elegans*	Exon (114:374)	261	2481	2481	191.35	197.19	12.97	12.58
1J942	*C. elegans*	Exon (430:504)	75	144	144	56.03	56.32	2.57	2.56
1J942	*C. elegans*	Exon (560:844)	285	879	879	212.37	212.39	4.14	4.14
1J942	*C. elegans*	Exon (1328:1633)	306	1078	1078	226.9	228.33	4.75	4.72
1B727	*C. elegans*	Exon (1080:1280)	201	360	360	149.22	149.79	2.41	2.40
1B727	*C. elegans*	Exon (9553:9750)	198	488	488	147.58	147.62	3.31	3.31
TRK1	*Yeast*	Full gene	3708	49514	49514	2727.04	2758.90	18.16	17.95
UBP12	*Yeast*	Full gene	3765	99720	99720	2761.60	2806.80	36.11	35.53
GroES	*E. coli*	Full gene	294	1628	1628	219.48	219.68	7.42	7.38
GroEL	*E. coli*	Full gene	1647	49189	49189	1229.79	1230.33	40.00	39.98
DNA pol	*E. coli*	Full gene	2352	35287	35287	1758.04	1747.60	20.07	20.00
PhoP	*B. subtilis*	Full gene	723	4263	4263	529.40	538.93	8.00	7.86
1J942	*C. elegans*	Intron (845:1027)	183	42	42	129.56	136.95	0.32	0.31
1J942	*C. elegans*	Intron (1257:1328)	72	5	5	49.17	54.21	0.10	0.09
1B727	*C. elegans*	Intron (9750:11183)	1434	688	688	1015.96	1071.78	0.68	0.64
1B727	*C. elegans*	Intron (1279:2292)	7068	5048	5048	5175.78	5254.33	0.98	0.95

[a]The following abbreviations are used in the table: $PS(N/3)$ denotes the power spectrum at $N/3$ calculated by DFT; F_3 is described in formula (3.4); E_1 denotes the average power spectrum over all the frequencies calculated by DFT; E_2 denotes the average power spectrum calculated by the formula (3.10); R_1 and R_2 denote signal-to-noise ratios calculated using the formulas; $R_1 = PS(N/3)/E_1$ and $R_2 = \frac{3}{2}F_3/E_2$.

TABLE 2. STATISTICS OF 3-BASE PERIODICITY AND SIGNAL-TO-NOISE IN EXONS, INTRONS, AND RANDOM DNA SEQUENCES[a]

Type	*avg N*	*std N*	*avg* $\frac{3}{2}F_3$	*std* $\frac{3}{2}F_3$	*avg* E_2	*std* E_2	*avg* R_2	*std* R_2
Initial exons	59	85.38	144	371	44	64	2.10	1.39
Terminal exons	641	716.49	8910	36705	479	534	8.94	13.82
Introns	2882	12836.41	2230	11078	2138	9518	0.82	0.74
Random DNA	2882	12836.41	2310	11639	2134	9505	0.99	0.57

[a]Abbreviations: *avg* and *std* denote the average and standard deviation, respectively; N is the lengths of test sequences; E_2 and R_2 are same as those used in Table 1.

It has been observed that initial exons are relatively short (mostly <100 bp) and terminal exons are relatively long (mostly 300–500 bp) (Zhang, 1998). We measure the signal-to-noise ratios in 924 initial exons, 924 terminal exons, and 924 introns. In addition, we generate uniformly distributed random DNA sequences with same length as each test intron sequence. Table 2 lists the statistical results and shows the following facts: (1) Exon sequences have high signal-to-noise ratios which are larger than a threshold value 2, but intron sequences have low signal-to-noise ratios, which are usually less than 1. (2) Signal-to-noise ratios increase with the increase of lengths of exon sequence. The signal-to-noise ratio can be identified for small exon regions such as initial exons (about 60 bp). (3) Intron sequences show statistics features similar to those of uniformly random DNA sequences in terms of signal, noise, and signal-to-noise ratio.

We suggest that an intron which does not have the 3-base periodicity is a random DNA sequence with approximately uniform nucleotide distributions. To justify this assumption, we create different DNA sequences with nucleotides uniformly distributed and compute the signal-to-noise ratios based on formulas (3.6) and (3.10). From Fig. 8, it can be seen that uniformly distributed DNA sequences have low signal-to-noise ratios, with 94% of the test cases displaying a ratio smaller than 2. This result can be explained using formulas (3.1) and (3.2) since the uniform distribution of nucleotides generates very weak signal.

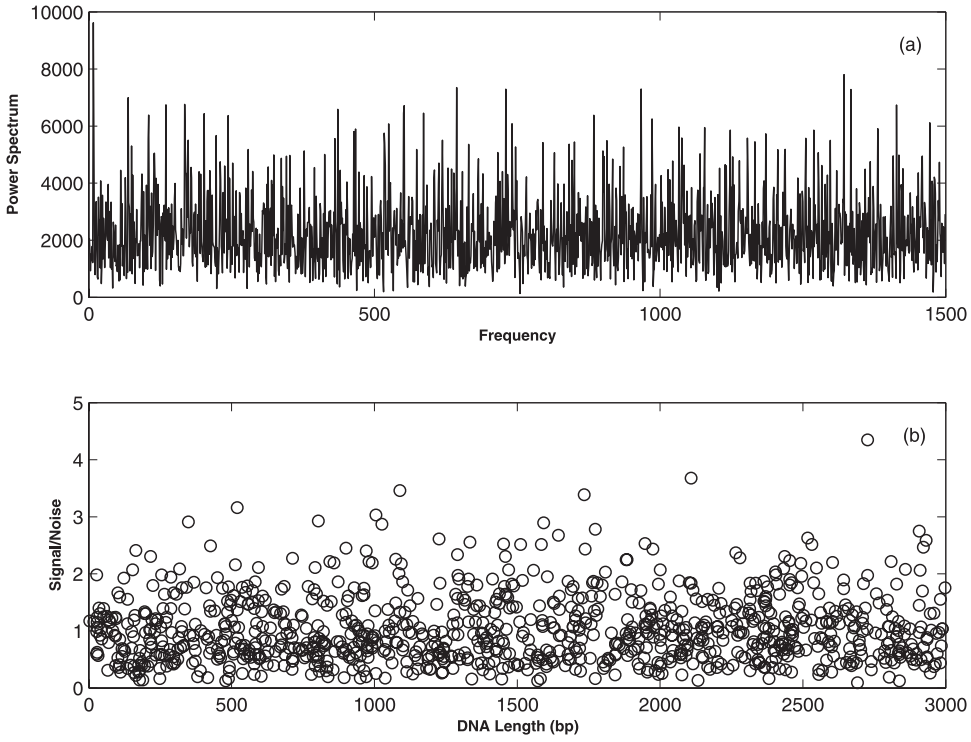

FIG. 8. (**a**) Power spectrum of a uniformly random DNA sequence of length 3,000 bp. (**b**) Simulation of the signal-to-noise ratio for uniformly random DNA sequences.

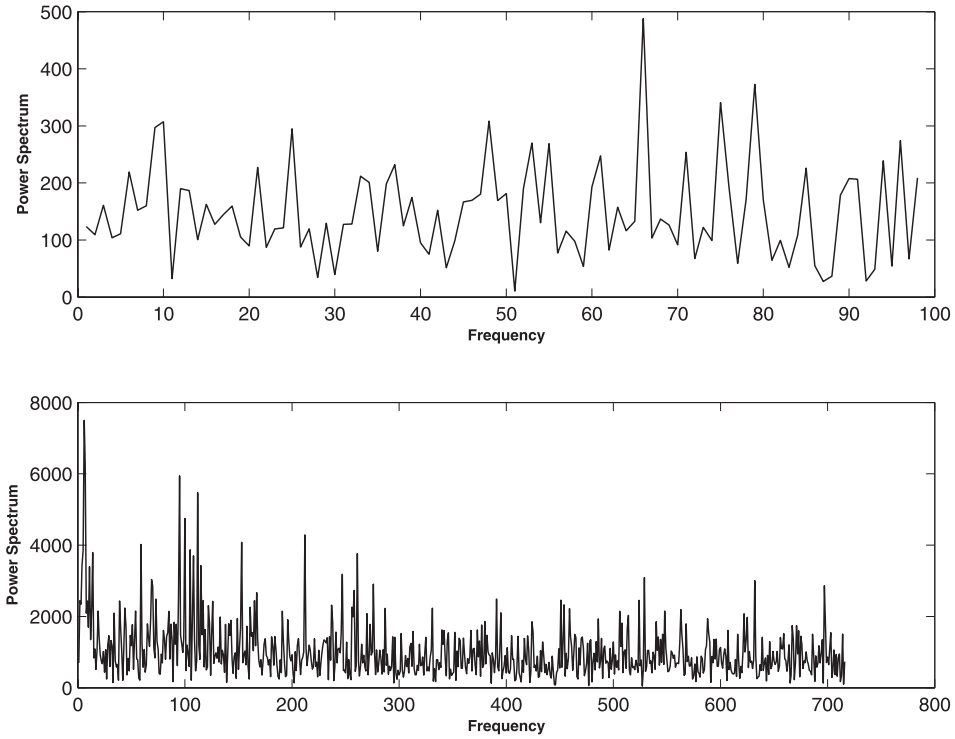

FIG. 9. (**a**) DFT power spectrum of an exon (9553:9750) of 1B727 of *C.elegans*. There is a clear peak at frequency $N/3$ ($N = 198$). (**b**) DFT power spectrum of an intron (9750:11183) of gene 1B727 of *C.elegans*. There is no peak at the frequency $N/3$ ($N = 1433$).

This result agrees with the suggestion that an intron is probably a uniformly distributed DNA sequence. The result also provides a threshold value of the signal-to-noise ratio to distinguish exons from introns. If a DNA segment has a signal-to-noise value in the Fourier spectrum greater than 2, the DNA fragment can be considered a putative exon; otherwise, it is an intron.

As an example to indicate the 3-base periodicity in a DNA sequence, Fig. 9 shows the power spectrum of an exon and intron of gene 1B727 described in Table 1. From the test experiments, it is also found that the nucleotide G is abundant on the first codon position of many test genes. This excess frequency of guanine on first codon positions in exons contributes to the unbalanced nucleotide distribution, which results in the 3-base periodicity in the exons. Thus, we may also infer that the total frequencies of amino acids, Val, Ala, Glu, Asp, and Gly, which are encoded by the genetic codons beginning with G, will be high in protein structures. Lobry (1997) showed that Ala, Asp, and Glu have high frequencies in 15 genomes. Other authors reported similar results (Bharanidharan *et al.*, 2004; Echols *et al.*, 2002). Thus, our results are consistent with other findings.

4. CONCLUSIONS

The results of this study offer deep insight into the origin of 3-base periodicity in DNA sequences and provide efficient ways to compute $PS(N/3)$ and noise/signal in a DNA sequence. The study shows that 3-base periodicity is due to a unique amino acid composition which leads to an unbalanced distribution of the four nucleotides on the three codon positions. The relationship between the power spectrum of $N/3$ and nucleotide distribution is linear. It also shows that the average power spectra have a correlation with nucleotide distribution and the lengths of DNA sequences. The signal-to-noise ratio in a DNA Fourier power spectrum can then be simply calculated using nucleotide distributions and the length of the DNA sequence. Without performing a Fourier transform, the approach presented in this paper reduces the computing time significantly and is very useful for gene-finding.

As a special case, excess of one of the four nucleotides, such as guanine (G), at the first codon position is a typical example of this unbalance, which contributes to the 3-base periodicity in many exons.

ACKNOWLEDGMENTS

We wish to thank Professor Jiangsong Wang at the Department of Mathematics of Nanjing University, China, for his helpful suggestions on this paper, Dr. Rong (Lucy) He at the Department of Pharmacology of University of Illinois at Chicago, for providing part of the test DNA sequences, and Professor Michael Waterman for his careful reading and many constructive suggestions to revise this paper.

REFERENCES

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

Bharanidharan, D., Bhargavi, G.R., Uthanumallian, K., and Gautham N. 2004. Correlations between nucleotide frequencies and amino acid composition in 115 bacterial species. *Biochem. Biophys. Res. Comm.* 315, 1097–1103.

Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C., and Marcourt, L. 2000. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J. Theor. Biol.* 206, 323–326.

Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N.M., Bertone, P., Zhang, Z., and Gerstein, M. 2002. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucl. Acids Res.* 30, 2515–2523.

Ficket, J.W., and Tung, C.-S. 1992. Assessment of protein coding measure. *Nucl. Acids Res.* 20, 6441–6450.

Gopalan, V., Tan, T.W., Lee., B.T., and Ranganathan, S. 2004. Xpro: Database of eukaryotic protein-encoding genes. *Nucl. Acids Res.* 32, D59–D63.

Lobry, J.R. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205, 309–316.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattachrya, S., and Ramaswamy, R. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS* 113, 263–270.

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

Yau, S.S.-T., Wang, J., Niknejad, A., Lu, C., Jin, N., and Ho, Y. 2003. DNA sequence representation without degeneracy. *Nucl. Acids Res.* 31, 3078–3080.

Zhang, M.Q. 1998. Statistical features of human exons and their flanking regions. *Human Mol. Genet.* 7, 919–932.

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

E-mail: yau@uic.edu