Distinguish Coding And Noncoding Sequences In A Complete Genome Using Fourier Transform

Yu Zhou[1], Li-Qian Zhou[1], Zu-Guo Yu[1,2*]
[1]School of Mathematics
and Computational Science,
Xiangtan University,
Hunan 411105, China.

Vo Anh[2]
[2]School of Mathematical Sciences,
Queensland University of Technology,
GPO Box 2434, Brisbane,
Q 4001, Australia.

Abstract

A Fourier transform method is proposed to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation of the DNA sequence proposed in our previous paper (Zhou et al., J. Theor. Biol. 2005) and the imperfect periodicity of 3 in protein coding sequences. The three parameters $P_{\mathbf{x}(\bar{\mathbf{s}})}(1)$, $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/3)$ *and* $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/36)$ *in the Fourier transform of the number sequence representation of DNA sequences are selected to form a three-dimensional parameter space. Each DNA sequence is then represented by a point in this space. The points corresponding to coding and non-coding sequences in the complete genome of prokaryotes are seen to be divided into different regions. If the point (*$P_{\mathbf{x}(\bar{\mathbf{s}})}(1)$, $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/3)$, $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/36)$*) for a DNA sequence is situated in the region corresponding to coding sequences, the sequence is distinguished as a coding sequence; otherwise, the sequence is classified as a noncoding one. Fisher's discriminant algorithm is used to study the discriminant accuracy. The average discriminant accuracies* p_c, p_{nc}, q_c *and* q_{nc} *of all 51 prokaryotes obtained by the present method reach 81.02%, 92.27%, 80.77% and 92.24% respectively.*

1. Introduction

The DNA sequence is formed from four different nucleotides, namely adenine (a), cytosine (c), guanine (g) and thymine (t). The complete genomes provide essential information for understanding gene functions and evolution. The determination of patterns in DNA and protein sequences is also useful for many important biological problems such as identifying new genes and discussing phylogenetic relationships among organisms. Accurate prediction of genes in genomes has always been a challenging task for bioinformaticians and computational biologists [1].

It is known that coding and non-coding sequences have different statistical and fractal behaviors. Li *et al.* [2] found that the spectral density of a DNA sequence containing mostly introns shows $1/f^{\beta}$ behavior, which indicates the presence of long-range correlation when $0 < \beta < 1$. The correlation properties of coding and non-coding DNA sequences were first studied by Peng *et al.* [3] in their fractal landscape or DNA walk model. They discovered that there exists long-range correlation in non-coding DNA sequences while the coding sequences correspond to a regular random walk. By undertaking a more detailed analysis, Chatzidimitriou-Dreismann and Larhammar [4] concluded that both coding and noncoding sequences exhibit long-range correlation. A subsequent work by Prabhu and Claverie [5] also substantially corroborates these results. If one considers more details by distinguishing c from t in pyrimidine, and a from g in purine (such as two or three-dimensional DNA walk models [6] and maps given by Yu and Chen [7], then the presence of base correlation has been found even in coding sequences. On the other hand, Buldyrev *et al.* [8] showed that long-range correlation appears mainly in noncoding DNA using all the DNA sequences available. Based on equal-symbol correlation, Voss [9,10] showed a power law behavior for the sequences studied regardless of the proportion of intron contents. The fractal methods for DNA sequence analysis were reviewed by Yu *et al.* [11]. Yu *et al.* [12] performed a multifractal analysis based on the chaos game representation of protein sequences from complete genome. The measure representation of linked protein sequence from a complete genome was proposed and its multifractal analysis was performed by Yu *et al.* [13]. Zhang *et al.* [14] used the parameters from root-mean-square fluctuation analysis to distinguish intron-containing and intronless genes based on the properties of Z curves [15]. Kulkarni *et al.* [1] proposed to

*Corresponding author Zu-Guo Yu, e-mail: yuzg1970@yahoo.com or z.yu@qut.edu.au

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00 © 2007 IEEE

use local Holder exponent formalism to identify coding and non-coding sequences.

In their review paper, Fickett and Tung [16] pointed out that future gene-finding algorithms should be Fourier, run, ORF and the in-phase hexamer [17]. Hence Yan *et al.* [17] proposed a new Fourier transform approach to distinguish coding sequences from noncoding sequences. The data set used in the above papers covers a large number of organisms.

In our previous paper [18], a number sequence representation of DNA sequences was proposed. Then a fractal method was used to distinguish coding and non-coding sequences in a complete genome based on their different statistical behaviors. In the present work, we propose to use the Fourier transform approach to distinguish coding and non-coding sequences in a complete genome based on the number representation of DNA sequences. The parameters $P_{\mathbf{x}(\bar{\mathbf{s}})}(1)$, $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/3)$ and $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/36)$ (to be elaborated below) in the Fourier transform of the number sequence representation of DNA sequences are selected to form a three-dimensional parameter space, and each DNA sequence is then represented by a point in this space. If the point $(P_{\mathbf{x}(\bar{\mathbf{s}})}(1),\ P_{\mathbf{x}(\bar{\mathbf{s}})}(1/3),\ P_{\mathbf{x}(\bar{\mathbf{s}})}(1/36))$ for a DNA sequence is situated in the region corresponding to coding sequences, the sequence is distinguished as a coding sequence; otherwise, the sequence is classified as a noncoding one. Fisher's discriminant algorithm is used to study the discriminant accuracy. The average discriminant accuracies of all 51 prokaryotes obtained by the present method will be reported.

2 Method

In this paper, we use a unique number sequence representation of each DNA sequence, which is proposed in our previous paper [18]. Here we briefly describe this representation.

Firstly, considering the properties of purine or pyrimidine, and strong or weak bonds, we define a map from the nucleotides to the numbers as

$$F: \quad \begin{cases} c & \mapsto & 1, \\ g & \mapsto & 3, \\ a & \mapsto & 5, \\ t & \mapsto & 7. \end{cases}$$

Secondly, we map each K nucleotides to a number. Any string made of K letters from the set $\{g, c, a, t\}$ is called a K-string. Denoting a K-string by $S = s_1 \cdots s_K,\ s_i \in \{c, g, a, t\}, i = 1, \cdots, K$, we define $x(S) = \sum_{i=1}^{K} F(s_i)/l^i$, where the base l can be any integer number which is larger than 7 to guarantee that $x(S)$ is unique for different K-string S. In this paper we set $l = 16$.

For each DNA sequence $\bar{S}$ and a fixed integer K, we construct a partition of $\bar{S}$ by dividing it into non-overlapping K-strings. If we denote the partition as $\bar{S} = S_1 S_2 \cdots S_N$, where each $S_i,\ i = 1, 2, \cdots, N-1$, is a K-string and S_N is a substring with length less than or equal to K, then the sequence $\mathbf{x}(\bar{S}) = (x(S_i), x(S_2), \cdots, x(S_N))$ is called the *number sequence representation* of the DNA sequence $\bar{S}$ corresponding to the given K. It can be proved that the number sequence representation is unique for each DNA sequence with any fixed K [18].

The power spectrum for a number sequence is defined as $P_{\mathbf{x}(\bar{\mathbf{S}})}(f) = \frac{1}{N}\left|\sum_{n=1}^{N} x(S_n) \exp\left(-2\pi i f n\right)\right|^2$, for a given frequency f.

Our idea is to select three parameters from the power spectrum $\{P_{\mathbf{x}(\bar{\mathbf{S}})}(f) : \quad f \in [0, 1]\}$ to form a three-dimensional parameter space, so that each DNA sequence can be represented by a point in this space.

3 The benchmark to evaluate the method

We use Fisher's linear discriminant algorithm [19,20] to calculate the discriminant accuracies of our method.

For all coding sequences of each genome, we randomly selected 80% of coding sequences to compose a training set, and the remaining 20% of coding sequences to form the test set. For all non-coding sequences of each genome, a similar selection is undertaken. We consider the three-dimensional space spanned by $\{P_{\mathbf{x}(\bar{\mathbf{S}}))}(f_1), P_{\mathbf{x}(\bar{\mathbf{S}}))}(f_2), P_{\mathbf{x}(\bar{\mathbf{S}}))}(f_3)\}$, where f_1, f_2, f_3 are three frequencies selected from the interval $[0, 1]$. Each coding or non-coding sequence can be represented as a point in this space.

We described Fisher's discriminant algorithm in [18]. We define p_c as the discriminant accuracy of coding sequences, p_{nc} as the discriminant accuracy of noncoding sequences, in the training set; q_c as the discriminant accuracy of coding sequences, q_{nc} as the discriminant accuracy of noncoding sequences, in the test set as in [18].

4 Results and Discussion

We selected the same 51 complete genomes of Archaea and Eubacteria available from the public database Genbank at the web site ftp://ncbi.nlm.nih.gov/ genbank/genomes/ as those in [18]. We use the abbreviations of these 51 prokaryotes in our figures and table in this paper. For full name and category of them, the reader can refer to Table 1 in [18].

We tried the cases $K = 1$ to 6, and found that $K = 1$ is the best length of non-overlapping substrings in the number sequence representation for the method. We then tried all combinations of 3 frequencies in the power spectra from the set {1, 1/2, 1/3, 1/4, 1/5, 1/6,

1/7 ,1/8 , 1/9, 1/10} and found the combination of parameters $\{P_{\mathbf{x}(\bar{\mathbf{S}})}(1), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/3),\ P_{\mathbf{x}(\bar{\mathbf{S}})}(1/9)\}$ is the best combination to distinguish coding and non-coding sequences in a genome. The average discriminant accuracies $p_c,\ p_{nc},\ q_c$ and q_{nc} of this combination reach 71.61%, 95.19%, 70.95%, 95.25%, respectively. We noticed that 3 and 9 are both multiples of 3. It is known that there exists imperfect periodicity of 3 in protein coding sequences [21-23]. Therefore we changed the frequency set to $\{1\} \cup \{1/(3m) :\ \ m \in N, 1 \leq m \leq 12\}$. After calculating the average prediction accuracies of all combinations of 3 frequencies from this set and considering that the number of coding sequences is much higher than the number of non-coding sequences, we found the combination of parameters $\{P_{\mathbf{x}(\bar{\mathbf{S}})}(1), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/3), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/36)\}$ is the best combination from this set. It seems that the reason for the situation is the imperfect periodicity of 3 in DNA sequences [24]. From the description in Section 2, we transform each DNA sequence to a unique point $\{P_{\mathbf{x}(\bar{\mathbf{S}})}(1), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/3), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/36)\}$ in the three-dimensional space spanned by these three parameters. After plotting all points of a genome in three-dimensional space, it is seen that the points corresponding to the coding sequences and those corresponding to the non-coding sequences assemble at different regions. Hence we suggest to select the three parameters $P_{\mathbf{x}(\bar{\mathbf{s}})}(1),\ P_{\mathbf{x}(\bar{\mathbf{s}})}(1/3)$ and $P_{\mathbf{x}(\bar{\mathbf{s}})}(1/36)$ in the power spectra of the number sequence representations of DNA sequences to form a parameter space to distinguish coding and noncoding sequences from one genome. As examples, the distributions of coding and noncoding sequences in the genomes of *Aquifex aeolicus* VF5 (Aquae), *Borrelia burgdorferi* B31 (Bbur), *Campylobacter jejuni* (Cjej) and *Pasteurella multocida* PM70 (Pmul) in this parameter space are shown in Figure 1. If the point $(P_{\mathbf{x}(\bar{\mathbf{s}})}(1),\ P_{\mathbf{x}(\bar{\mathbf{s}})}(1/3),\ P_{\mathbf{x}(\bar{\mathbf{s}})}(1/36))$ for a DNA sequence is situated in the region corresponding to coding sequences, the sequence is discriminated as a coding sequence; otherwise, the sequence is classified as a noncoding one. The Fisher's discriminant accuracies for the selected prokaryotes in this parameter space are shown in Table 1.

From Table 1, it is seen that our method worked well for a large part, nearly 90%, of all 51 prokaryotes. We list these prokaryotes in the top part of Table 1. The average discriminant accuracies $p_c,\ p_{nc},\ q_c$ and q_{nc} of these 46 prokatyotes reach 82.01%, 93.04%, 81.77% and 92.92%, respectively. The method was not that effective on the remaining 5 prokaryotes. The average discriminant accuracies $p_c,\ p_{nc},\ q_c$ and q_{nc} of these latter prokaryotes are only 71.94%, 85.23%, 71.61% and 85.95%, respectively. But for all 51 prokatyotes, the average discriminant accuracies $p_c,\ p_{nc},\ q_c$ and q_{nc} reach 81.02%, 92.27%, 80.77% and 92.24%, respectively.

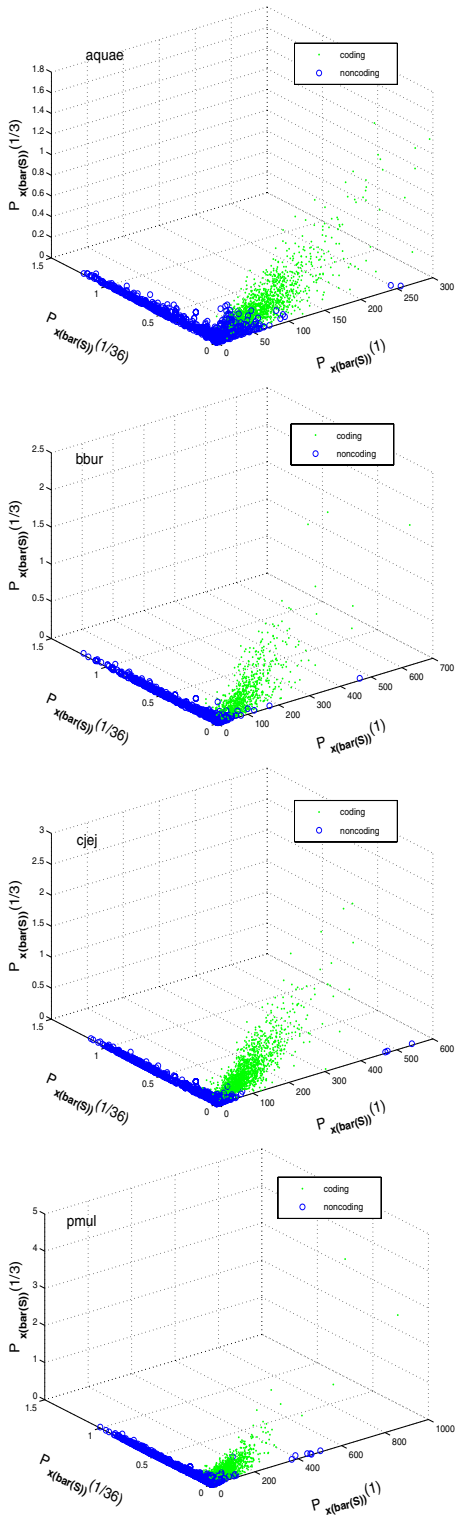

Figure 1. The distribution of all coding and noncoding sequences in the complete genomes of four bacteria in the parameter space generated by the three parameters $P_{\mathbf{x}(\bar{\mathbf{S}})}(1)$, $P_{\mathbf{x}(\bar{\mathbf{S}})}(1/3)$ and $P_{\mathbf{x}(\bar{\mathbf{S}})}(1/36)$ in power spectrum.

Table 1. The Fisher's discriminant accuracies for the 51 organisms.

Species	p_c	p_{nc}	q_c	q_{nc}
Aful	83.85%	94.65%	85.68%	92.33%
HaloNRC	73.89%	95.64%	75.97%	95.50%
Mthe	85.56%	96.93%	85.83%	96.36%
Mjan	81.25%	93.78%	78.93%	93.26%
Pabyssi	88.39%	90.77%	86.16%	94.81%
Phor	83.59%	79.83%	79.55%	78.63%
Ssol	77.37%	87.30%	77.01%	87.44%
Taci	78.28%	93.78%	81.42%	93.48%
Tvol	78.79%	88.39%	80.72%	90.56%
MtubC	83.97%	92.96%	83.29%	92.03%
MtubH	83.16%	98.29%	81.25%	98.16%
Aquae	93.76%	87.54%	96.39%	88.57%
Tmar	93.43%	91.17%	94.59%	88.89%
Bsub	80.49%	90.12%	81.12%	89.33%
Bhal	77.68%	96.65%	76.54%	97.57%
Llac	77.05%	96.15%	75.33%	95.64%
Mgen	96.26%	80.00%	95.74%	71.19%
Spne	80.79%	80.53%	77.33%	78.89%
CaceA	81.01%	95.39%	82.04%	94.97%
Mpneu	83.98%	84.86%	84.56%	86.02%
Mpul	83.07%	90.56%	82.80%	91.45%
Spyo	79.81%	94.33%	81.47%	93.91%
Uure	93.87%	88.45%	95.12%	89.22%
SaurM	76.70%	88.14%	76.61%	91.09%
SaurN	78.27%	84.61%	76.11%	83.48%
Cpneu	81.12%	96.47%	75.36%	97.65%
CpneuA	79.00%	95.52%	79.78%	95.17%
CpneuJ	81.19%	96.33%	76.64%	97.08%
Ctra	82.79%	98.05%	85.23%	94.37%
Nost	70.58%	98.11%	69.18%	98.60%
Synecho	74.80%	97.34%	77.60%	96.86%
Bbur	92.65%	92.20%	94.15%	92.86%
Tpal	80.36%	97.64%	82.13%	99.33%
Atum	81.07%	95.87%	81.47%	96.05%
Ccre	80.97%	97.67%	79.68%	98.96%
Rpro	77.54%	99.65%	82.04%	100%
Smel	76.62%	98.82%	76.08%	98.60%
NmenA	71.83%	94.05%	69.88%	93.58%
Buch	80.53%	97.55%	76.11%	93.20%
EcolKM	79.49%	95.90%	78.11%	96.52%
Paer	81.76%	99.52%	82.68%	99.46%
Pmul	88.03%	95.55%	86.10%	96.33%
EcolOH	77.41%	93.56%	77.10%	93.08%
Hinf	79.69%	94.50%	78.20%	95.89%
Cjej	92.22%	94.70%	96.68%	96.98%
Hpyl	88.51%	89.93%	85.35%	91.29%
Aero	74.86%	78.09%	72.91%	80.56%
Mlep	72.70%	85.12%	75.60%	85.84%
pNGR234	72.46%	89.47%	67.86%	89.55%
Xfas	71.08%	80.36%	72.20%	81.69%
Nmen	68.58%	93.12%	69.46%	92.10%

We also calculated average discriminant accuracies of the frequency combinations $(1, 1/3, 1/(3m))$, $12 < m < 39$. We found that the average p_c can be improved slightly, but p_{nc} becomes a little worse. Because we were not able to try all the combinations (f_1, f_2, f_3), $f_i \in [0, 1]$, we can only say that the combination (1,1/3,1/36) is satisfactory in the setting of the present method.

For the problem under study, we have shown the fractal method [18] is better than that proposed by Zhang *et al.* [14]. We now want to compare the present method with the fractal method in our previous paper [18]. Using the fractal method, the average discriminant accuracies p_c, p_{nc}, q_c and q_{nc} were 72.28%, 84.65%, 72.53% and 84.18% respectively. The average discriminant accuracies p_c, q_c, p_{nc} and q_{nc} by using the present method has been improved by 8.74%, 7.62%, 8.24% and 8.06%, respectively. Besides this improvement, the present method can be used to distinguish coding and non-coding sequences in the complete genome of each species without restricting their length, while the fractal method can only be used to distinguish long coding and non-coding sequences.

In order to compare the present method further, we performed the Fourier method based on the Z curve representation [15] of the 51 genomes, which is similar to the method proposed in Yan *et al.* [17]. We performed the Fourier transform directly on the Z curve, and fixed $f = 1/3$ in the power spectrum of $\{x_n\}, \{y_n\}$ and $\{z_n\}$ in the Z curve representation respectively of each coding and non-coding sequence to form a parameter space. We then calculated the discriminant accuracies by using Fisher's discriminant mentioned in Section 2. The average discriminant accuracies p_c, p_{nc}, q_c and q_{nc} are 65.61%, 98.56%, 65.42% and 98.62%, respectively. We notice the average discriminant accuracies p_{nc} and q_{nc} obtained by the Z curve method are 6.29% and 6.38% higher than those obtained by the present method, but the average discriminant accuracies p_c and q_c are 15.41% and 15.35% lower. Considering that the number of coding sequences is much larger than the number of non-coding sequences, we can conclude that the present method is more efficient than the Fourier transform approach based on the Z curve representation.

5 Conclusions

The number sequence representation proposed by Zhou *et al.* [18] is unique for each DNA sequence with any fixed K. We found $K = 1$ is the best length of non-overlapping substrings in the number sequence representation for the present method.

The combination of parameters $\{P_{\mathbf{x}(\bar{\mathbf{S}})}(1), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/3), P_{\mathbf{x}(\bar{\mathbf{S}})}(1/36)\}$ is a good combination to distinguish coding and non-coding sequences in each genome. Hence, we can transform each DNA sequence to a unique point $\{P_{\mathbf{x}(\bar{\mathbf{S}})}(1),$

$P_{\mathbf{x}(\tilde{\mathbf{S}})}(1/3)$, $P_{\mathbf{x}(\tilde{\mathbf{S}})}(1/36)\}$ in the three-dimensional space spanned by these three parameters. In this space, the points corresponding to the coding and noncoding DNA sequences can be divided into two different regions. Our method works well to distinguish coding and non-coding sequences in the 46 prokaryotes listed in the top of Table 1, but it does not work equally well for the remaining 5 prokaryotes. On the whole, the average discriminant accuracies p_c, p_{nc}, q_c and q_{nc} of all prokaryotes reach 81.02%, 92.27%, 80.77% and 92.24%, respectively.

Acknowledgement

Financial support was provided by the Chinese National Natural Science Foundation (no. 30570426), Fok Ying Tung Education Foundation (no. 101004) and the Youth foundation of Educational Department of Hunan province in China (no. 05B007) (Z.G. Yu), Australian Research Council (no. 0559807) (V.V. Anh), Scientific Research Fund of Hunan Provincial Education Department in China (no. 06C830) (L.Q. Zhou).

References

[1] O.C. Kulkarni, R. Vigneshwar, V.K. Jayaraman, B.D. Kulkarni, Identification of coding and non-coding sequences using local Holder exponent formalism, *Bioinformatics*, 21(20):3818-3823, 2005.

[2] W. Li, T. Marrand, K. Kaneko, Understanding long-range correlations in DNA sequences, *Physica D*, 75:392-416, 1994.

[3] C.K. Peng, S. Buldyrev, A.L. Goldberg, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, Long-range correlations in nucleotide sequences. *Nature*, 356:168-170, 1992.

[4] C.A. Chatzidimitriou-Dreismann and D. Larhammar, Long-range correlations in DNA, *Nature*, 361:212-213, 1993.

[5] V.V. Prabhu and J. M. Claverie, Correlations in intronless DNA, *Nature*, 359:782-782, 1992.

[6] L. Luo, W. Lee, L. Jia, F. Ji and L. Tsai, Statistical correlation of nucleotides in a DNA sequence, *Phys. Rev. E*, 58(1):861-871, 1998.

[7] Z.G. Yu and G.Y. Chen, Rescaled range and transition matrix analysis of DNA sequences. *Comm. Theor. Phys.*, 33(4):673-678, 2000.

[8] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons, and H. E. Stanley, Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, *Phys. Rev. E*, 51(5):5084-5091, 1995.

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

[10] R. Voss, Long-Range Fractal Correlations in DNA Introns and Exons, *Fractals*, 2:1-6, 1994.

[11] Z.G. Yu, V. Anh, Z.M. Gong and S.C. Long, Fractals in DNA sequence analysis, *Chin. Phys.*, 11(12):1313-1318, 2002.

[12] Z.G. Yu, V. Anh and K.S. Lau, Chaos game representation, and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model, *J. Theor. Biol.*, 226(3):341-348, 2004.

[13] Z.G. Yu, V. Anh and K.S. Lau, Multifractal and correlation analysis of protein sequences from complete genome, *Phys. Rev. E*, 68:021913, 2003.

[14] C.T. Zhang, Z.S. Lin, M. Yan and R. Zhang, A novel approach to distinguish between intron-containing and intronless genes based on the format of Z curves, *J. Theor. Biol.*, 192:467-473, 1997.

[15] C.T. Zhang and R. Zhang, Z curves, an intuitive tool for visualizating and analyzing the DNA sequences, *J. Biomolec. Struct. Dyn.*, 11:767-782, 1994.

[16] Fickett J.W. and Tung C.S., Assessment of protein coding measures. Nucleic Acids Res., **20**, (1992) 6441-6450.

[17] M. Yan, Z.S. Lin and C.T. Zhang, A new Fourier transform approach for protein coding measure based on the format of Z curve, *Bioinformatics*, 14:685-690, 1998.

[18] L.Q. Zhou, Z.G. Yu, J.Q. Deng, V. Anh and S.C. Long, A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation, *J. Theor. Biol.*, 232:559-567, 2004.

[19] K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis*, London: Academic Press, 1979.

[20] Duda R.O., Hart P.E. and Stork D.G., Pattern Classification (Second Edition). New York: JOHN WILEY & SONS, INC, 2001.

[21] B.D. Silverman and R. Linsker, Periodicity of DNA sequences. *J. Theor. Biol.*, 118:295-300, 1986.

[22] E.N. Trifonov, Translation framing code and frame-monitoring mechanism as suggested by analysis of mRNA and 16S rRNA nucleotide sequences, *J. Mol. Biol.*, 194:643-652, 1987.

[23] P. Lio, S. Ruffo and M. Buiatti, Third codon G+C periodicity as possible signal for an internal selective constraint, *J. Theor. Biol.*, 171:215-223, 1994.

[24] Fickett J.W., Finding genes by computer: the state of the art. *Trends Genetics*, 12:316-320, 1996.