QUT Digital Repository:
http://eprints.qut.edu.au/

Yu, Zuguo and Anh, Vo V. and Zhou, Yu and Zhou, Li-Qian (2007) Numerical Sequence Representation of DNA Sequences and Methods To Distinguish Coding And Non-Coding Sequences in a Complete Genome. In Callaos, N. and Lesso, W. and Zinn, C. and Zmazek, B., Eds. *Proceedings 11th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2007*, pages pp. 171-176, Florida, USA.

© Copyright 2007 (please consult author)

Numerical sequence representation of DNA sequences and methods to distinguish coding and non-coding sequences in a complete genome

Zu-Guo Yu
School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia. Email: z.yu@qut.edu.au
School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China.

Vo Anh
School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia. Email: v.anh@qut.edu.au

Yu Zhou and Li-Qian Zhou
School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China.
Email: zynova@hotmail.com (for Yu Zhou), zhoulq@xtu.edu.cn (for Li-Qian Zhou).

Abstract

In this presentation we introduce two methods to distinguish coding and non-coding sequences in a complete genome. A numerical sequence representation of DNA sequences is introduced first. There exists a one-to-one correspondence between a DNA sequence and its numerical sequence representation. In the first method, three exponents from a multifractal analysis are selected to construct the parameter space. In the second method, which is based on a Fourier transform approach, three parameters from the power spectrum of the numerical sequence representation are selected to construct the parameter space. Each DNA may be represented by a point in these three-dimensional spaces. We found that the points corresponding to coding and non-coding sequences in the complete genomes of prokaryotes are divided into different regions in both parameter spaces. If the point for a DNA sequence is situated in the region corresponding to coding sequences, the sequence is recognized as a coding sequence; otherwise, the sequence is classified as a non-coding one. The average accuracies using Fisher's discriminant algorithm for coding and non-coding sequences are satisfactory.

Keywords: complete genome; coding/non-coding sequences; fractal analysis; Fourier transform.

1. Introduction

The complete genomes provide essential information for understanding gene functions and evolution. Retrieval of biological information from complete genomes and finding the appropriate proteins or coding/non-coding regions of a complete genome for a specific biological problem are some of the challenges for researchers in the bioinformatical field. Information from complete genomes has been used to discuss the phylogenetic relationship of organisms (Sankoff *et al.* 1992; Fitz-Gibbon and House 1999; Tekaia *et al.* 1999; Lin and Gerstein 2000; Li *et al.* 2001; Yu and Jiang 2001; Stuart *et al.* 2002; Yu *et al.* 2004, 2005; Qi *et al.* 2004). Accurate prediction of genes in genomes has always been a challenging task for bioinformaticians and computational biologists (Kulkarni *et al.* 2005). Computer-aided gene finding in uncharacterized DNA sequences is one of the most important problems of bioinformatics. For most prokaryotic genomes, the problem is to determine which open reading frames (ORF) in a given genome are really those coding for proteins (Yan *et al.* 1998).

Many works have been done to study a range of different statistical and fractal behaviors of coding and non-coding sequences. Li *et al.* (1994) found that the spectral density of a DNA sequence containing mostly introns shows a power law behavior. Peng *et al.* (1992) proposed the fractal landscape or DNA walk model and 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 (1993) concluded that both coding and non-coding sequences exhibit long-range correlation. Using two or three-dimensional DNA walk models (Luo *et al.* 1998) and maps given by Yu and Chen (2000), the presence of base correlation has been found even in coding sequences. Zhang *et al.* (1997) used the parameters from root-mean-square fluctuation analysis to distinguish intron-containing and intronless genes based on the properties of Z curves (Zhang *et al.* 1994). A multifractal analysis based on the chaos game representation of DNA sequences was given in Gutierrez *et al.* (1998, 2001). Yu *et al.* (2004)

performed a multifractal analysis based on the chaos game representation of protein sequences from complete genomes. The measure representation of linked protein sequences from complete genomes was proposed and its multifractal analysis was performed in Yu et al. (2003).

In their review paper, Fickett and Tung (1992) pointed out that future gene-finding algorithms should be Fourier transform-based. Hence Yan *et al.* (1998) proposed a new Fourier transform approach to distinguish coding sequences from non-coding sequences. The data set used in the above papers covers a large number of organisms.

We are interested in the problem of distinguishing coding and non-coding sequences in the complete genome of one organism. In Zhou *et al.* (2005), we proposed a numerical sequence representation of DNA sequences. Multifractal analysis was then performed on the measure representation of the obtained numerical sequence (this technique appeared first in Yu *et al.* (2001)). Based on our numerical sequence representation, Kulkarni *et al.* (2005) proposed to use local Holder exponent formalism to distinguish coding and non-coding sequences. In this presentation we introduce two methods to distinguish coding and non-coding sequences in a complete genome based on different statistical behaviors of these two kinds of sequences: One is a fractal method proposed in Zhou *et al.* (2005), the other is a Fourier transform approach (Zhou *et al.* 2006).

2. Numerical sequence representation of DNA sequences

Luo (1998) considered the purine/pyrimidine and strong/weak bond properties of the four kinds of nucleotides to give their two-dimensional DNA walk representation. Here we also consider these two kinds of properties. We use the point (1,1) to represent nucleotide *c* corresponding to its pyrimidine and strong bond properties; the point (-1,1) to represent nucleotide *g* corresponding to its purine and strong bond properties; the point (-1,-1) to represent nucleotide *a* corresponding to its purine and weak bond properties; and the point (1,-1) to represent nucleotide *t* corresponding to its pyrimidine and weak bond properties. Then the vectors connecting the origin to the four points (1,1), (-1,1), (-1,-1) and (1,-1) have the rotational angles π /4, 3π/4, 5π /4, 7π/4 with the *x*-axis. We accordingly define the map

$$f:\ \begin{cases} c & \mapsto & 1, \\ g & \mapsto & 3, \\ a & \mapsto & 5, \\ t & \mapsto & 7. \end{cases} \qquad (1)$$

We call any string made of *K* letters from the set {*g*,*c*,*a*,*t*} a *K-string*. Letting $S = s_1 \ldots s_K$, ($s_i \in \{a,c,g,t\},\ i = 1,\ldots,K$), be a *K*-string, we define

$$x(S) = \sum_{i=1}^{K} f(s_i)/l^i, \qquad (2)$$

where the base *l* can be any integer number which is larger than 7 to guarantee that *x*(*S*) is unique for different *K*-strings *S*. In this presentation we set *l*=16. It can be proved that different substrings *S* have different representative values *x*(*S*).

Now, for a DNA sequence $\overline{S}$ and a fixed integer *K*, we can construct a partition of $\overline{S}$ by dividing it into non-overlapping *K*-strings. If we denote the partition as $\overline{S} = S_1 S_2 \ldots S_N$, with $S_i,\ i = 1,2,\ldots,N-1,$ being *K*-strings and S_N a substring with length less than or equal to *K*, then the numerical sequence $x(\overline{S}) = (x(S_1), x(S_2), \ldots, x(S_N))$ is called the *numerical sequence representation* of the DNA sequence $\overline{S}$ corresponding to the given *K* (Zhou *et al.* 2005). We have mentioned that different substrings *S* have different representative values *x*(*S*); so for any fixed *K*, different DNA sequences will have different numerical sequence representations. Hence there exists a one-to-one correspondence between a DNA sequence and its numerical sequence representation.

3. Fractal method

3.1. *A measure for the numerical sequence representation of a DNA sequence*

Let $x(S_1), x(S_2), \ldots, x(S_N)$ be the numerical sequence representation of a DNA sequence. First we define $F_t = x(S_t)/(\sum_{j=1}^{N} x(S_j))$ to be the frequency of $x(S_t), t = 1,2,\ldots,N.$ It follows that $\sum_t F_t = 1.$ We denote by I_t the interval $[(t-1)/N, t/N).$ Now we can define a measure μ on the interval [0,1) by $d\mu(r) = Y(r)dr,$ where

$$Y(r) = N \times F_t = x(S_t)/(\frac{1}{N}\sum_{j=1}^{N} x(S_j)), \text{for } r \in I_t \qquad (3)$$

It is seen that $\int_0^1 d\mu(r) = 1$ and $\mu(I_t) = F_t.$ The way to define the measure μ for the numerical sequence representation is the same as that of the measure for the length sequence from a complete genome (Yu *et al.* 2001).

3.2. *Multifractal analysis*

The most common numerical implementations of multifractal analysis are based on the *fixed-size box-counting algorithms* (Halsy *et al.* 1986). In the one-dimensional case, for a given measure μ with support $E \subset \boldsymbol{R}$, we consider the *partition sum*

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q, \; q \in \boldsymbol{R}, \qquad (4)$$

where the sum runs over all different nonempty boxes B of a given side ε in a grid covering of the support E, that is, $B = [k\varepsilon, (k+1)\varepsilon]$. The scaling exponent $\tau(q)$ is defined as

$$\tau(q) = \lim_{\varepsilon \to 0} \frac{\log Z_\varepsilon(q)}{\log \varepsilon}. \qquad (5)$$

The scaling exponent $\tau(q)$ is numerically estimated through a linear regression of $\log Z_\varepsilon(q)$ against $\log\varepsilon$ for any real number q. The relationship between the exponent $\tau(q)$ and the generalized fractal dimension D_q is $\tau(q) = D_q(q-1), q \in \boldsymbol{R}$. For example, we give the generalized dimension D_q for two coding regions and two non-coding regions in the genome of *Escherichia coli* K-12 MG1655 (EcoliKM) in Figure 1.

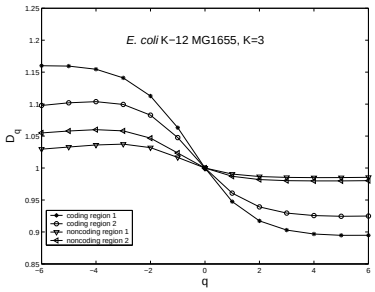

Figure 1. The generalized dimension D_q for two coding regions and two non-coding regions in the genome of *Escherichia coli* K-12 MG1655 (EcoliKM).

By following the thermodynamic formulation of multifractal measures, Canessa (2000) derived an expression for the `analogous' specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1) \qquad (6)$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. In the following we calculate the analogous specific heat of numerical sequence representations. For example, the analogous specific heat C_q for the same regions as those in Figure 1 is shown in Figure 2.

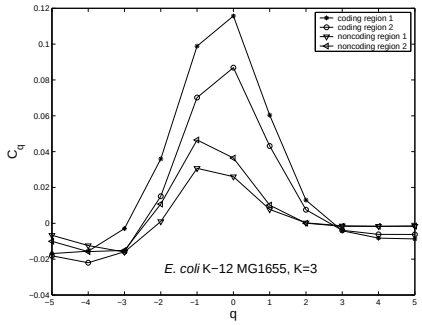

Figure 2. The `analogous' specific heat C_q for the same regions as those in Figure 1.

Our idea is to select three values of C_q to form a three-dimensional parameter space to distinguish coding and non-coding sequences from one complete genome.

4. Fourier transform approach

Let $x(S_1), x(S_2), ..., x(S_N)$ be the numerical sequence representation of a DNA sequence. The power spectrum for a numerical sequence is defined as

$$P_{x(\overline{S})}(f) = \frac{1}{N}\left|\sum_{n=1}^{N} x(S_n)\exp(-2\pi i f n)\right|^2,$$

for a given frequency f. (7)

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

5. Results and discussion

We selected 51 complete genomes of Archaea and Eubacteria available from the public database Genbank at the web site ftp://ncbi.nlm.nih.gov/genbank/genomes/. They are ten **Archaebacteria**: *Aeropyrum pernix* (Aero), *Archaeoglobus fulgidus* DSM4304 (Aful), *Halobacterium* sp NRC-1 (HaloNRC), *M. jannaschii* DSM2661 (Mjan), *M. thermoautotrophicum* deltaH (Mthe), *Pyrococcus abyssi* (Pabyssi), *Pyrococcus horikoshii* OT3 (Phor), *Sulfolobus solfataricus* (Ssol), *Thermoplasma acidophilum* (Taci), *Thermoplasma volcanium* GSS1 (Tvol); three **Gram-positive Eubacteria (high G+C)**: *Mycobacterium leprae* TN (Mlep), *Mycobacterium tuberculosis* CDC1551 (MtubC), *Mycobacterium tuberculosis* H37Rv (MtubH); twelve **Gram-positive Eubacteria (low G+C)**: *Bacillus halodurans* C-125 (Bhal), *Bacillus subtilis* 168 (Bsub), *Clostridium acetobutylicum* ATCC824 (CaceA), *Lactococcus lactis* IL 1403 (Llac), *Mycoplasma genitalium* G37 (Mgen), *Mycoplasma pneumoniae* M129 (Mpneu), *Mycoplasma pulmonis* (Mpul), *Staphylococcus aureus*

Mu50 (SaurM), *Staphylococcus aureus* N315 (SaurN), *Streptococcus pneumoniae* (Spne), *Streptococcus pyogenes* M1 (Spyo), *Ureaplasma urealyticum* (serovar 3) (Uure); two **Hyperthermophilic bacteria**: *Aquifex aeolicus* VF5 (Aquae) and *Thermotoga maritima* MSB8 (Tmar); four **Chlamydia**: *Chlamydia pneumoniae* CWL029 (Cpneu), *Chlamydia pneumoniae* AR39 (CpneuA), *Chlamydia pneumoniae* J138 (CpneuJ), *Chlamydia trachomatis* (serovar D) (Ctra); two **Cyanobacteria**: *Nostoc sp.* PCC6803 (Nost), *Synechocystis* sp. PCC6803 (Synecho); two **Spirochaete**: *Borrelia burgdorferi* B31 (Bbur) and *Treponema pallidum* Nichols (Tpal); five **Proteobacteria alpha subdivision**: *Agrobacterium tumefaciens* (Atum), *Caulobacter crescentus* (Ccre), *Rhizobium* sp. NGR234 (pNGR234), *Rickettsia prowazekii* Madrid (Rpro), *Sinorhizobium meliloti* (Smel); two **Proteobacteria beta subdivision**: *Neisseria meningitidis* MC58 (Nmen) and *Neisseria meningitidis* Z2491 **(**NmenA); seven **Proteobacteria gamma subdivision**: *Buchnera* sp. APS (Buch), *Escherichia coli* K-12 MG1655 (EcolKM), *Escherichia coli* O157:H7 EDL933 (EcolOH), *Haemophilus influenzae* Rd (Hinf), *Pseudomonas aeruginosa* PA01 (Paer), *Pasteurella multocida* PM70 (Pmul), *Xylella fastidiosa* 9a5c (Xfas); and two **Proteobacteria epsilon subdivision**: *Campylobacter jejuni* (Cjej) and *Helicobacter pylori* 26695 (Hpyl).

For the fractal method, we found that C_{-1}, C_1, C_2 for *K*=3 are good parameters to form a three-dimensional parameter space to distinguish coding and non-coding sequences from one complete genome (Zhou *et al*. 2005). Each DNA may be represented by a point in this space. From the three-dimensional plots, we found that points corresponding to coding and non-coding sequences in the complete genomes of many prokaryotes are roughly distributed in different regions. For example, the results for *Archaeoglobus fulgidus* DSM4304 (Aful) and *Thermoplasma acidophilum* (Taci) are shown in Figure 3. We found that a clear pattern similar to that shown in Figure 3 exists in the plots for 31 prokaryotes which include **Archaebacteria**, **Hyperthermophilic bacteria**, **Chlamydia** and **Proteobacteria** (alpha, beta and gamma subdivisions). For left prokaryotes this method does not seem to work well (their plots are similar to those shown in Figure 4.

For the Fourier transform approach, we found that the three parameters $P_{x(\bar{S})}(1)$, $P_{x(\bar{S})}(1/3)$ and $P_{x(\bar{S})}(1/36)$ for *K*=1 in the power spectra of the numerical sequence representations of DNA sequences are good parameters to form a parameter space (Zhou *et al*. 2006). As examples, the distributions of coding and non-coding sequences in the genomes of *Campylobacter jejuni* (Cjej) and *Pasteurella multocida* PM70 (Pmul) in this parameter space are shown in Figure 5. If the point ($P_{x(\bar{S})}(1)$, $P_{x(\bar{S})}(1/3)$, $P_{x(\bar{S})}(1/36)$) for a DNA sequence is situated in the region corresponding to coding sequences, the sequence is recognized as a coding sequence; otherwise, the sequence is classified as a non-coding one. This method works well for a large portion, nearly 90%, of all 51 prokaryotes (Zhou *et al*. 2006).

In order to quantitatively evaluate the performance of our methods and compare them with methods proposed by other groups (for example, Yan *et al*. 1998, Zhang *et al*. 1997). We use Fisher's discriminant algorithm (Duda *et a*l. 2001). We denote by p_c, p_{nc}, q_c, q_{nc} the discriminant accuracies of

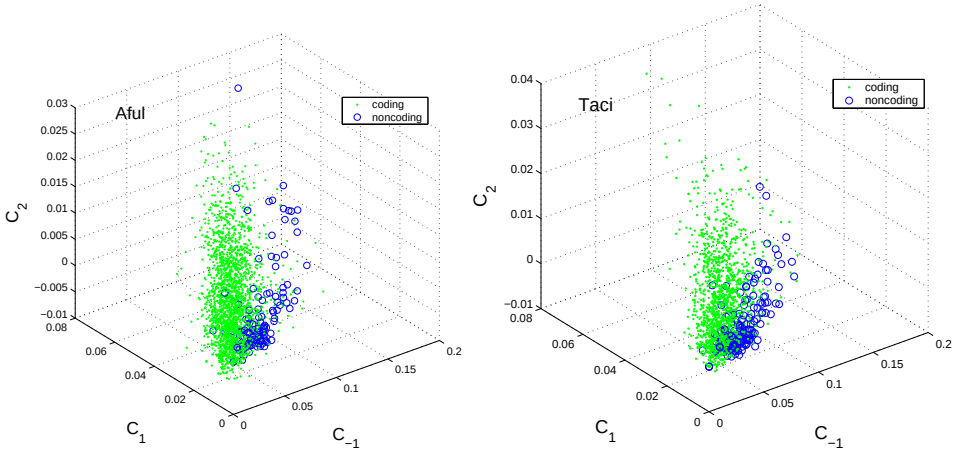

Figure 3. The distribution of the three exponents C_{-1}, C_1, C_2 of all coding and non-coding sequences in the complete genomes of *Archaeoglobus fulgidus* DSM4304 (Aful) and *Thermoplasma acidophilum* (Taci).

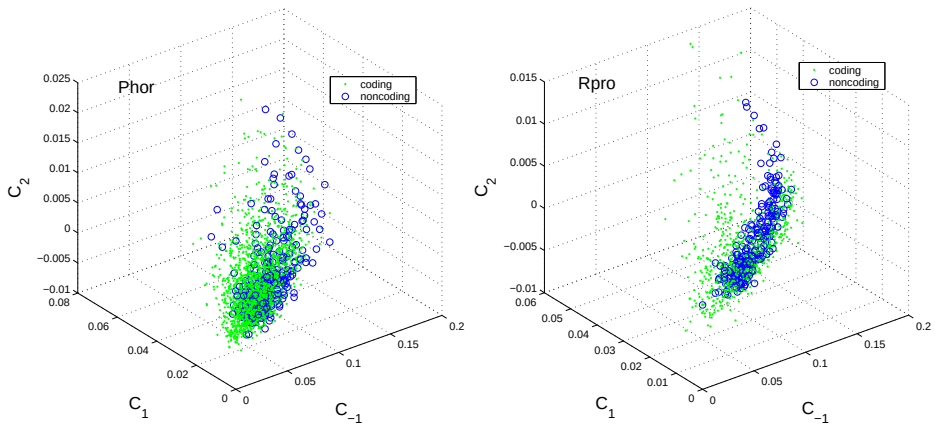

Figure 4. The distribution of the three exponents C_{-1}, C_1, C_2 of all coding and non-coding sequences in the complete genomes of *Pyrococcus horikoshii* OT3 (Phor) and *Rickettsia prowazekii* Madrid (Rpro).

coding and non-coding sequences in the training and test sets from one complete genome respectively. Here we randomly select 80% of coding and non-coding sequences as training sets and the remaining 20% of sequences are left as test sets. In the fractal method, for all 51 prokaryotes considered, the average discriminant accuracies p_c, p_{nc}, q_c and q_{nc} reach 72.28%, 84.65%, 72.53% and 84.18% respectively. In the Fourier transform approach, these average discriminant accuracies of all 51 prokaryotes reach 81.02%, 92.27%, 80.77% and 92.24% respectively. Based on these discriminant accuracies, the fractal method is seen to outperform that proposed by Zhang *et al.* (1997) and our Fourier transform approach (Zhou *et al.*, 2006) is superior to the fractal method (Zhou *et al*. 2005) and the Fourier transform method proposed in Yan *et al*. (1998).

6. Conclusions

The numerical sequence representation proposed by Zhou *et al.* (2005) is unique for each DNA sequence with any fixed *K*.

The parameters $P_{x(\bar{S})}(1)$, $P_{x(\bar{S})}(1/3)$, $P_{x(\bar{S})}(1/36)$ form a good combination to distinguish coding and non-coding sequences in each genome (Zhou *et al*. 2006).

These methods achieve satisfactory average accuracies based on Fisher's discriminant algorithm for coding and non-coding sequences.

Acknowledgments

This work is partially supported by the Australian Research Council grant DP0559807, Natural Science Foundation of China grant No. 30570426, Fok Ying

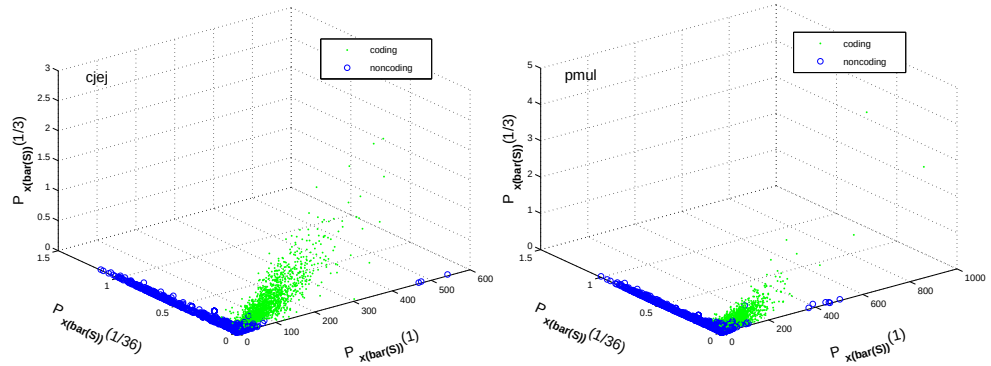

Figure 5. The distribution of all coding and non-coding sequences in the complete genomes of *Campylobacter jejuni* (Cjej) and *Pasteurella multocida* PM70 (Pmul) in the parameter space generated by the three parameters $P_{x(\bar{S})}(1), P_{x(\bar{S})}(1/3), P_{x(\bar{S})}(1/36)$ in the power spectrum.

Tung Education Foundation grant No. 101004 and the Youth Foundation of Educational Department of Hunan Province grant No. 05B007.

REFERENCES

[1] E. Canessa, Multifractality in time series, **J. Phys. A: Math. Gen.** vol 33, 2000, pp3637-3651.

[2] C.A. Chatzidimitriou-Dreismann and D. Larhammar, Long-range correlations in DNA, **Nature**, vol 361, 1993, pp212-213.

[3] J.W. Fickett and C.S. Tung, Assessment of protein coding measures. **Nucleic Acids Res.,** vol 20, 1992, pp6441-6450.

[4] S.T. Fitz-Gibbon and C.H. House, Whole genome-based phylogenetic analysis of free-living microorganisms. **Nucleic Acids Res.**, vol 27, 1999, pp4218-4222.

[5] J. M. Gutierrez, A. Iglesias and M.A. Rodriguez, Analyzing the multifractal structure of DNA nucleotide sequences, in: *Chaos and Noise in Biology and Medicine* (Eds.: M. Barbi and S. Chillemi) World Scientific Publishing, Singapore, 1998, pp. 315-319.

[6] T. Halsy, M. Jensen, L. Kadanoff, I. Procaccia and B. Schraiman, Fractal measures and their singularities: the characterization of strange set, **Phys. Rev. A**, vol 33, 1986, pp1141-1151.

[7] J.M. Gutierrez, M.A. Rodriguez and G. Abramson, Multifractal analysis of DNA sequences using novel chaos-game representation, **Physica A**, vol 300, 2001, pp271-284.

[8] O.C. Kulkarni, R. Vigneshwar, V.K. Jayaraman, B.D. Kulkarni, Identification of coding and non-coding sequences using local Holder exponent formalism, **Bioinformatics**, vol 21(20)}, 2005, pp3818-3823.

[9] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny. **Bioinformatics**, vol 17, 2001, pp149-154.

[10] W. Li, T. Marr, and K. Kaneko, Understanding long-range correlations in DNA sequences, **Physica D**, vol 75, 1994, pp392-416.

[11] J. Lin, and M. Gerstein, Whole-genome trees based on the occurrence of folds and orthologs, implications for comparing genomes at different levels. **Genome Res.**, vol 10, 2000, pp808-818.

[12] L. Luo, W. Lee, L. Jia, F. Ji and L. Tsai, Statistical correlation of nucleotides in a DNA sequence, **Phys. Rev. E**, vol 58(1), 1998, pp861-871.

[13 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**, vol 356, 1992, pp168-170.

[14] J. Qi, B. Wang, and B.L. Hao. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. **J. Mol. Evol**., vol 58(1), 2004, pp1-11.

[15] D. Sankoff, G. Leaduc, N. Antoine, B. Paquin, B.F. Lang and R. Cedergren, Gene order omparisons for phylogenetic inference: Evolution of the mitochondrial genome. **Proc. Natl. Acad. Sci. U.S.A.,** vol 89, 1992, pp6575-6579.

[16] G.W. Stuart, K. Moffet, and S. Baker, Integrated gene species phylogenies from unaligned whole genome protein sequences, **Bioinformatics**, vol 18, 2002, pp100-108.

[17] F. Tekaia, A. Lazcano, and B. Dujon, The genomic tree as revealed from whole proteome comparisons. **Genome Res**., vol 9, 1999, pp550-557.

[18] 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,** vol 14, 1998, pp685-690.

[19] Z.G. Yu, V.V. Anh and K.S. Lau, Multifractal characterisation of length sequences of coding and non-coding segments in a complete genome, **Physica A**, vol 301(1-4), 2001, pp351-361.

[20] Z.G. Yu, V.V. Anh and K.S. Lau, Multifractal and correlation analysis of protein sequences from complete genome, **Phys. Rev. E**, vol 68}, 2003, pp021913.

[21] Z.G. Yu, V.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**., vol 226(3), 2004, pp341-348.

[22] Z.G. Yu and G.Y. Chen, Rescaled range and transition matrix analysis of DNA sequences, **Comm. Theor. Phys.**, vol 33(4), 2000, pp673-678.

[23] Z.G. Yu, and P. Jiang, Distance, correlation and mutual information among portraits of organisms based on complete genomes, **Phys. Lett. A**, vol 286, 2001, pp34-46.

[24] Z.-G. Yu, L.-Q. Zhou, V. Anh, K.H. Chu, S.-C. Long and J.-Q. Deng, Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment, **J. Mol. Evol**., vol 60, 2005, pp 538-545.

[25] 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**., vol 192, 1997, pp467-473.

[26] C.T. Zhang and R., Zhang, Z curves, an intuitive tool for visualizating and analyzing the DNA sequences, **J. Biomolec. Struct. Dyn**., vol 11, 1994, pp767-782.

[27] 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 numerical sequence representation, **J. Theor. Biol**., vol 232, 2004, pp559-567.

[28] Y. Zhou, L.Q. Zhou, Z.G. Yu and V. Anh, Distinguish coding and non-coding sequences in a complete genome using Fourier transform. 2006, submitted.