2001 Proceedings of the 23rd Annual EMBS International Conference, October 25-28, Istanbul, Turkey

FEATURE EXTRACTION FROM DNA SEQUENCES BY MULTIFRACTAL ANALYSIS

H. Zhang and W. Kinsner
Department of Electrical and Computer Engineering, Signal and Data Compression Laboratory
University of Manitoba, Winnipeg, Manitoba R3T 5V6, Canada
e-mail: {kinsnerlhgzhang}@ee.umanitoba.ca

Abstract-**This paper presents feature extraction and estimation of multifractal measures of DNA sequences, using a multifractal methodology, and demonstrates a new scheme for identifying biological functionality, using information contained within the DNA sequences. It shows that the Rényi and Mandelbrot fractal dimension spectra may be useful techniques for extracting the information contained in the DNA sequences.**
Keywords - **DNA, power spectrum density, fractal**

I. INTRODUCTION

The genome of an organism is a total set of the deoxyribonucleic acid (DNA) molecules, which are composed of the sugar-phosphate backbone and four nitrogenous bases (adenine, A, thymine, T, cytosine, C, and guanine, G), repeated millions of times throughout the genome. The genome contains the master blueprint, which is mostly stored in its genes, each responsible for making a single protein, for all cellular structures and activities for the lifetimes of the cell or the organism. A gene is a piece of DNA sequence which is composed of exons and introns in higher eukaryotic organisms. It has been long known that the coding regions (also known as exons) of the genes carry information which instructs the cellular process in the way of leading from DNA sequences to amino acid sequences or proteins, while the non-coding regions (the introns of the genes and the intergenic regions) contain no information about the proteins in the organism. The proteins in the organism determine, among other things, how the organism looks, how well its body metabolizes food or fights infection, and sometimes even how it behaves. Therefore, accurate localization of genes and other parts of our genome may lead to an understanding of the genome and to the understanding of life.

Recently a draft sequence of the human genome, which covers 96% of the entire human genome containing 3×10^9 base pairs, has been published by the Human Genome Project (HGP) and Celera Genomics. However, the rate of locating genes is relatively low, and about 35% of human genes remains unknown. Therefore, developing computational tools for locating genes and elucidating the structure of genes is becoming essential for molecular biology. In addition, new computational methods can provide complementary information which can be of benefit to locating genes by the traditional experimental methods.

Most of the current research in the deciphering the meaning of DNA sequences is approached from the low base-pair level. Its main objective is to search for patterns or correlations existing in the DNA sequence related to codons (three-base sequences), amino acids, and proteins. A number of gene prediction systems have been developed in recent years. These systems use a variety of sophisticated computational techniques, including neural networks [1], dynamic programming [2], rule-based methods [3], decision trees [4], probability reasoning [5] and hidden Markov chains [6]. Most of these techniques rely on the statistical qualities of exons in the genome and therefore, the fundamental limitation of them is the use of a known gene data pool as a training set for their classification. Consequently, they are capable of finding only the genes that are homologous with those in the training data set.

Kinsner *et al.* have demonstrated that fractal techniques can be useful in the classification of stationary and nonstationary signals such as speech, image, and radio transmitter transients [7]. Assuming that DNA sequences have a fractal structure, Karlin *et. al.* [8] have demonstrated a long-range power-law relations on the DNA sequences, spanning 10^4 nucleotides. Peng *et al.* [9] demonstrated the correlation properties of coding and non-coding regions of DNA sequences, using Lévy walk method to map the DNA alphabet sequence into a numerical sequence. Other researches have shown that the long-range fractal correlations appear in the coding region of the DNA sequences, with different values in different regions of the sequence [10], [11], [12]. Alternatively, Barral *et. al.* [13] reported that coding regions behave statistically as random chains, as compared to non-coding regions. Yu *et al.* [14], [15] proposed a time series model based on the global structure of the complete genome, and showed long-range correlations in the bacteria DNA sequences. Although those papers present various algorithms in DNA research, the Lévy walk or a modified Lévy walk are often used for translating a DNA alphabet sequence into a numerical sequence. In the Lévy walk, a walker either descends or rises one step at the position i along a DNA sequence chain if a pyrimidine (C/T) or a purine (A/G) occurs, respectively. Therefore, an artificial error is introduced by giving specific values to the sequence at each base pair.

Unequal usage of codons in the coding regions appears to be a universal feature of the genomes across a wide range of species. The bias is mainly the result of two forces, the amino acid usage bias and the unequal usage of synonymous

0-7803-7211-5/01/$17.00 © 2001 IEEE

codons [16]. The latter could be correlated with the mutational biases and natural selection acting at the levels of replication, transcription, and translation [17], [18]. In other words, each organism has its own synonymous codon preferences.

In this paper, we introduce a new algorithm which is using multifractal techniques and the uneven codon usage between coding and non-coding regions in the DNA sequences. Based only on the structural information given by the DNA sequences, a significant difference between the coding and non-coding regions can be demonstrated by using our algorithm without pre-training data sets. Therefore, we demonstrate a new way of locating genes in an genome. We have developed a technique that: (i) transforms a DNA alphabet sequence to a DNA numerical signal sequence using a specially constructed transform matrix; and (ii) estimates the Rényi and Mandelbrot dimension spectra of the DNA numerical signal sequence.

II. BACKGROUND

2.1 Transcription, Translation, and Open Reading Frame

As shown in Fig. 1, a genome is composed of genes and intergenic regions. The structure of the genes in higher eukaryotic organisms usually consists of a number of small exons (coding portions) separated by larger introns (non-coding portions). During a transcription process, a messenger ribonucleic acid (mRNA) is produced based on the corresponding individual base pairs of the coding portions of a gene (with T substituted by uracil, U), and the non-coding portions of the gene are removed. Within a translation process, each specific codon from the mRNA template is responsible for the selection of a corresponding amino acids. In addition, the sequence of codons, terminated by stop codons and the poly-A tail in the mRNA template, is responsible for the orderly assembly into an amino acid sequence. The resulting amino acid sequence is a protein. There are 64 possible codons corresponding to only 20 amino acids and a stop signal. An open reading frame (ORF) is the decoding sequence which is translated from an mRNA template. An ORF contains not only amino acids but stop signals. There are three frames according to an mRNA template since three-base DNA sequence represents one amino acid. Only one frame, which is the ORF, represents the amino acid sequence. The other frames are shifted by either one or two bases of the DNA sequence with respect to the ORF. To locate a gene's position is to determine the positions of its exons.

2.2 DNA Signal

For the DNA multifractal analysis, we first translate a DNA sequence into a corresponding DNA numerical signal sequence, using a specific character-to-number translation matrix which is constructed based on the general assumption that all the ORFs of the coding sequences within a genome have a common feature of codon usage bias, while the non-coding regions and all the shifted frames of the coding sequences in the same genome do not have a codon usage bias. Then we treat the DNA numerical signal sequence as a spatial series.

2.3 Rényi and Mandelbrot Fractal Dimension Spectra

Fractals have been studied extensively in physics and mathematics. A fractal dimension demonstrates the degree of complexity (or roughness, brokenness, and irregularity) of an object which is statistically self-similar to some extent [19]. There are many distinct definitions of fractal dimen-

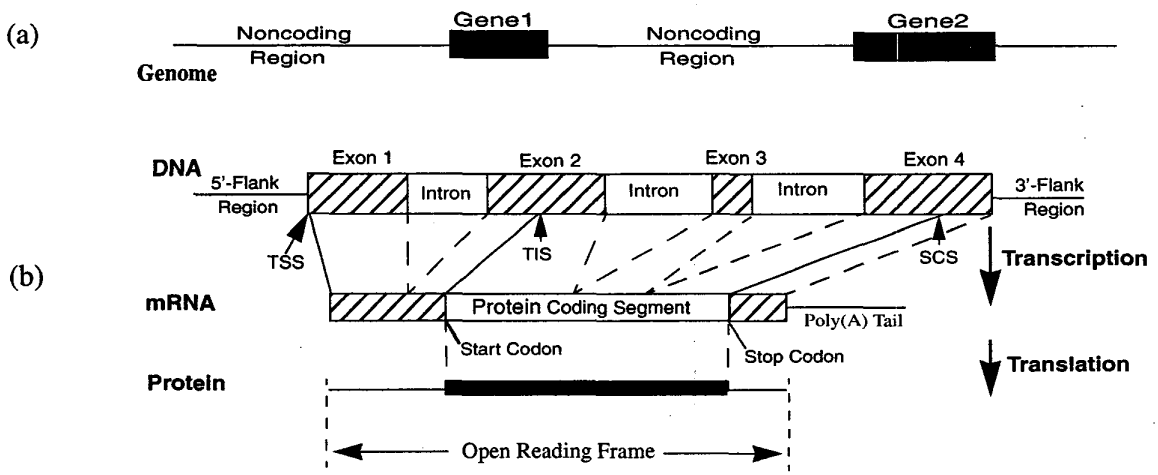

Fig. 1. From a genome to proteins. (a) The structure of the higher eukaryotic genomic DNA and (b) A schematic chart of the transcription and translation process. (TSS, transcription start site; TIS, translation initiation site; SCS, stop coding site.)

sions in order to reflect the different properties of self-similar and self-affine signals. Morphological fractal dimensions reveal the dominant fractal properties of a multifractal signal. Using multifractal analysis, such as the Rényi and Mandelbrot fractal dimension spectra, more information in the multifractal signal structure can be revealed. Therefore, this approach can determine if the signal is a single fractal (with a single-valued fractal dimension), or a multifractal (with a spectrum of fractal dimensions).

The Rényi dimension D_q is defined as [19]

$$D_q = \lim_{r \to 0} \frac{1}{q-1} \frac{\log \sum_{j=1}^{Nr} \left(p_j^q\right)}{\log(r)} \qquad -\infty \le q \le \infty \qquad (1)$$

where r stands for the size of a volume element (vel); N_r is the number of vels for a given vel size that cover the fractal object, p_j denotes the frequency of occurrence in a given vel, and q is the moment order.

The Mandelbrot dimension D_{Man} and the Hölder exponent α_q are related to the Rényi dimension by [19]

$$\alpha_q = \frac{d}{dq}[(q-1)D_q] \qquad (2)$$

$$D_{Man} = q\alpha_q - (q-1)D_q \qquad (3)$$

The Mandelbrot spectrum reveals the distribution of singularities in the fractal object, and is a useful tool for an "online" analysis. Since the Mandelbrot fractal dimension spectrum is derived from the Rényi fractal dimension spectrum, there should be similarities between them. However, there are also differences that may be useful in classification.

III. Results and Discussion

3.1 DNA Samples

The data of human codon usage is originally from Ikemura *et al.* [20]. For the testing, we construct (i) a random DNA sequence which is generated from a uniform white noise and (ii) a Cantor DNA sequence which has a Cantor set property. A piece of the human genomic DNA ph-20, the Hyal1, and Hyal2 cDNA sequences have been obtained from the GenBank. The genomic DNA sequence contains the human gene of the Hyal2 and the exon1 of the Hyal1. For the cDNA sequences, we remove the 5' end and 3' end of the non-coding regions before testing.

3.2 Rényi Fractal Dimension Spectrum

We have calculated the Rényi dimension spectra of the Cantor DNA sequence, the random DNA sequence, and the coding regions (cDNAs and exons) and non-coding regions (introns) of the real DNA sequences. The results are shown in Fig. 2. Figure 2(a) shows that one of the three frames of the Cantor DNA sequence has a single fractal dimension of 0.6288 (the solid line) which is very close to the theoretical value (log2/log3) of the Cantor set fractal dimension. The other two frames are strictly not the Cantor set since they are shifted by one or two bases and, hence, they show a slight multifractality. The three frames of the random DNA sequence (Fig. 2(b)) have exactly the same Rényi dimension spectrum since they have the same statistical property. For the real DNA sequences (Figs. 2(c) to 2(f)), our results support Kinsner and Rifaat's conclusion that DNA sequences are multifractal [21]. Only the ORFs have a significant difference of the Rényi dimension, as compared with that of the shifted frames due to the uneven codon usage on the ORF. As shown in Fig. 2(c) and 2(d), the non-ORFs have shapes statistically similar to the white noise DNA sequence. For the genomic DNA and the intron (Fig. 2(e) and 2(f)), the codon usage is even in general on the three frames, although the genomic DNA contains some coding regions. Therefore, the three frames of the genomic sequence have the same Rényi dimension spectrum.

We also tested different flanking regions of the DNA sequences and there is a slight, but not significant, difference among the Rényi dimension spectra of the three frames (data not shown), which reflects the GC-rich phenomena in some areas of the flank regions as the appearance of regulatory elements. Our algorithm can even show the difference of the Rényi dimension spectra between the non-coding regions (including intron and flanking region) and the non-coding genes (data not shown), although the differences are not significant.

3.3 The Mandelbrot Fractal Dimension Spectrum

Figures 3(a) to 3(f) show the Cantor DNA sequence, the white noise random DNA sequence, the coding region of the Hyal1 cDNA, the exon 3 of the hyal2, the human genomic DNA sequence (ph-20), and the intron 1 of the Hyal2, respectively. The solid lines in Figs. 3(b) to 3(f) and the point in Fig. 3(a) represent the ORFs of the coding regions and frame 1 of the non-coding regions. The dashed lines and dash-dot lines represent the shifted frames in both the coding and non-coding regions. Since one of the frames of the Cantor DNA sequence demonstrates a single fractal property, its Mandelbrot dimension and Hölder exponent are constant and therefore, the Mandelbrot spectrum is degraded to a single point. Similar to the Rényi dimension spectrum, the ORFs have different Mandelbrot spectra compared to that of the shifted frames. The results also show that the non-ORFs have a similar Mandelbrot spectrum with the white noise, indicating that the non-ORF and the white noise have the same multifractal behaviour. Therefore, our results demonstrate a scheme for locating coding regions within the genomic DNA sequences.

IV. Conclusions

In the DNA sequences, there is a significant difference

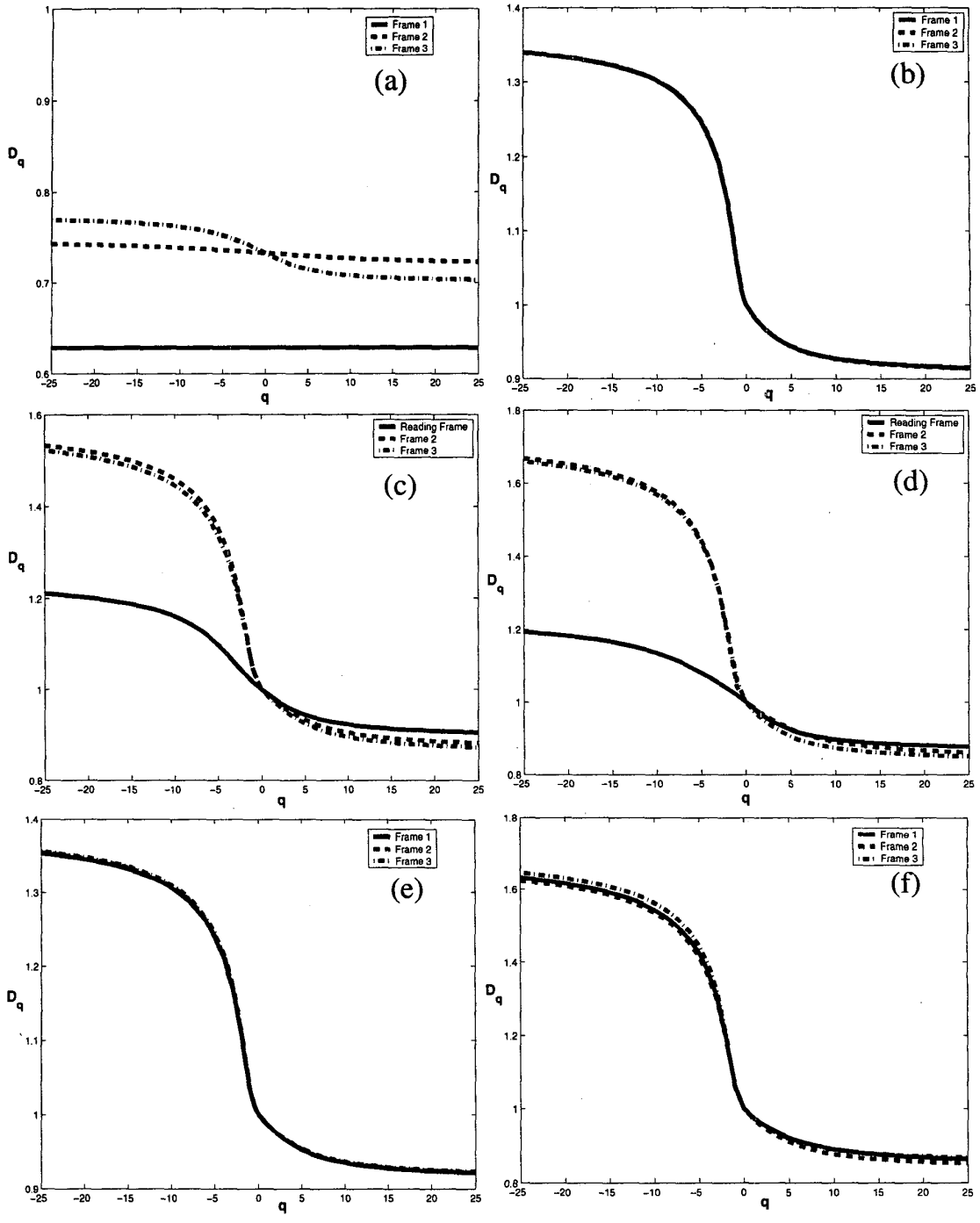

Fig. 2. Rényi dimension spectrum of the DNA sequences: (a) a Cantor DNA sequence, (b) a white noise random DNA sequence, (c) the coding region of the Hyal1 cDNA, (d) the exon 3 of the Hyal2, (e) the human genomic DNA sequence (ph-20), and (f) the intron 1 of the Hyal2. The solid lines represent the ORF (frame 1) and the dash and dot-dash lines are shifted frames.

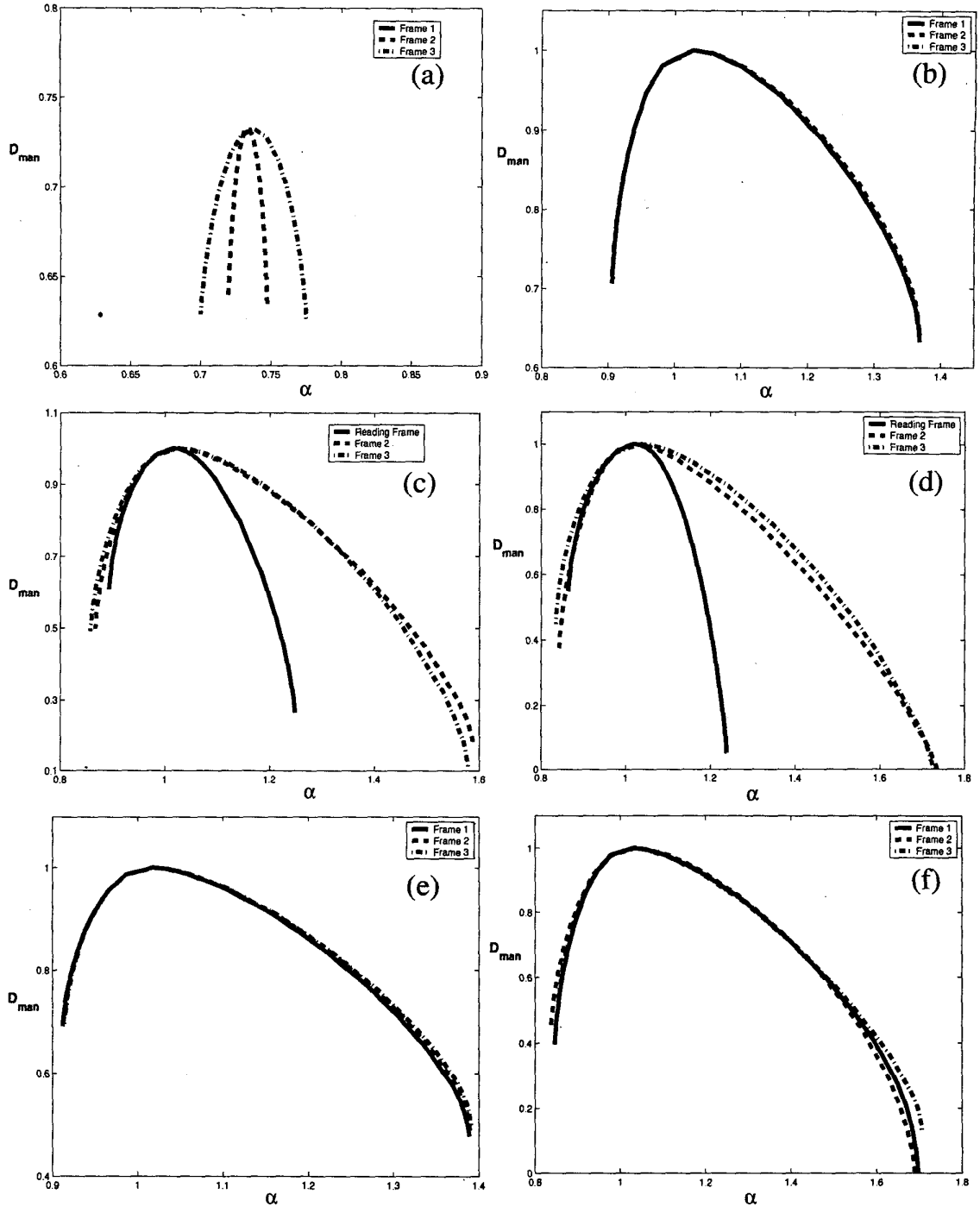

Fig. 3. The Mandelbrot dimension of the DNA sequences: (a) a Cantor DNA sequence, (b) a white noise random DNA sequence, (c) the coding region of the Hyal1 cDNA, (d) the exon 3 of the Hyal2, (e) the human genomic DNA sequence (ph-20), and (f) the intron 1 of the Hyal2. The solid lines represent the ORF (frame 1) and the dash and dot-dash lines are shifted frames. In Fig. 3(a), the solid line of Fig. 2(a) has collapsed to a dot.

of the Rényi dimensions between the ORF and the shifted frames of the coding region. The three frames of the non-coding region have similar multifractality, with a shape similar to the uniform white noise. Unlike the current gene prediction algorithms, our multifractal algorithm is carried out based exclusively on the multifractal structure and entropy properties of the DNA sequences, and does not need pre-training data sets for program training. Therefore, it opens up a useful way of classifying coding and non-coding regions of the DNA sequences.

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

[1] E. Uberbacher and R. Mural, "Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach," *Proc. Natl. Acad. Sci USA*, vol. 88, p11261-11265, 1991.

[2] E. E. Snyder and G. D. Stormo, "Identification of coding regions in genomic DNA sequences: An applocation of dynamic programming and neural networks," *Nuclei Acids Res.*, vol. 21, p607-613, 1993.

[3] V. Solovyev, A. Salamov, and C. Lawrence, " Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames," *Nucleic Acids Res.*, vol. 22, p5156-5163, 1994.

[4] G. Hutchinson and M. Hayden, "The prediction of exons through an analysis of spliceable open reading frames," *Nucleic Acids Res.*, vol. 20, p3453-3462, 1992.

[5] R. Guigo, S. Knudsen, N. Drake, and T. Smith, "Prediction of gene structure," *J. Molecular Biology*, vol. 226, p141-157, 1992.

[6] J. Henderson, S. Salzberg, and K. H. Fasman, " Finding genes in DNA with a hidden Markov model," *J. Computational Biology*, vol. 4, p127-141, 1997.

[7] W. Kinsner, "Fractal dimensions: Morphological, entropy, spectrum, and variance classes," *Technical Report*, DEL94-5, University of Manitoba, May 1994, 146 pp.

[8] S. Karlin and V. Brendel, "Patchiness and correlations in DNA sequences," *Science*, vol. 259, pp. 677-680, 1993.

[9] 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, pp. 168-171, 1992.

[10] P. Bernaola-Galvan, R. Roman-Roldan, and J. L. Oliver, "Compositional segmentation and long-range fractal correlations in DNA sequences," *Phys. Rev.* E, vol. 53, pp. 5181-5189, 1996.

[11] Y. Xiao, R. Chen, R. Shen, J. Sun, and J. Xu, "Fractal dimension of exon and intron sequences," *J. Theor. Biol.*, vol. 175, pp. 23-26, 1995.

[12] D. Larhammar and C. A. Chatzidimitriou-Dreismann, "Biological origins of long-range correlations and compositional variations in DNA," *Nucl. Acids Res.*, vol. 21, pp. 5167-5170, 1993.

[13] J. Barral P., A. Hasmy, J. Jiménez, and A. Marcano, "Nonlinear modelling technique for the analysis of DNA chains," *Phys. Rev. E*, vol. 61, pp.1812-1815, 2000.

[14] Z. Yu, V. V. Anh, and B. Wang, "Correlation property of length sequences based on global structure of the complete genome, " *Phys. Rev. E*, vol. 63, pp. 1-8, 2001.

[15] Z. Yu and V. Anh, "Time series model based on global structure of complete genome, " *Chaos, Solitons & Fractals*, vol. 12, pp. 1827-1834, 2001.

[16] R. Grantham, C. Gautier, M. Gouy, R. Mercier and A. Pave, "Codon catalog usage and the genome hypothesis," *Nucl. Acids Res.*, vol. 8, pp. r49-r62, 1980.

[17] R. Staden and A. D. McLachlan, "Codon preference and its use in identifying protein coding regions in long DNA sequences," *Nucl. Acids Res.*, vol. 9, pp. 141-156, 1981.

[18] H. Romero, A. Zavala, and H. Musto, "Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biased and a complex pattern of selective forces," *Nucl. Acids Res.*, vol. 28, pp. 2084-2090, 2000.

[19] W. Kinsner, "Fractal dimensions: Morphological, entropy, spectrum, and variance classes," *Technical Report, DEL 94-4,* Dept. of Electrical and Computer Engineering, University of Manitoba, 146 pp., May 1994.

[20] Codon Usage Database is developed and maintained by the First Laboratory for Plant Gene Research, Kazusa DNA Research Institute. Available from (as of June 2001) www.kazusa.or.jp/codon.

[21] R. Rifaat and W. Kinsner, "Multifractal analysis of DNA sequences," *Proc. of IEEE CCECE'99*, Edmonton, Canada, pp. 801-804, 1999.