Chinese Chemical Letters Vol. 14, No. 5, pp 503 – 504, 2003
http://www.imm.ac.cn/journal/ccl.html

Characterizing Long-range Correlation Properties in Nucleotide Sequences

Xiao Yan CHEN[1], Lun Jun BAO[2]*, Jin Yuan MO[1], Ying WANG[1]

[1]Department of chemistry, Zhongshan University, Guangzhou 510275
[2]Guangzhou Entry-Exit Inspection and Quarantine Bureau, Guangzhou 510623

Abstract: Using continuous wavelet transform as the analytical tool, the fractal characteristic of nucleotide sequences was studied. The fractal dimension of the exon and intron sequences for different species was calculated. We use the Mexican hat wavelet function as the mother wavelet and Hurst exponent to describe the long-range correlation. It is found that the Hurst exponent of intron sequence is larger than that of exon sequence for the same gene.

Keywords: Nucleotide sequences, long-range correlation, wavelet transform.

In recent years, it becomes a hotspot in analyzing and mining the information that contain in DNA sequences[1]. There has been intense discussion about the existence, the nature and the origin of long-range correlations in DNA sequences. Different techniques including mutual information functions[2], autocorrelation functions[3,4], power spectra[5], Zipf analysis[6] and Fourier analysis[7] were used for statistical analysis of DNA sequences. This letter uses wavelet transform[8], associated with fractal formalism to show the specialty of DNA sequences under multiple scales. This research focuses on the deeper understanding of the mechanisms of the genes.

A DNA string is arranged by adenosine (A), guanine (G), cytosine (C), and thymine (T). When searching patterns for a DNA string, a critical issue is how to represent the string. Standard techniques in signal analysis and temporal series analysis typically take a real or complex valued signal as input, but in this case the signal is built up from the symbols {A, C, G, T}. Numerous possibilities of representing a sequence have been proposed. In this case we define the DNA walk analysis as:

$$X_i = \begin{cases} 1 & \text{where these are two any nucleotides} \\ -1 & \text{where these are the other two nucleotides} \end{cases}$$

Here i=1,2,3,⋯,L. L is the length of that DNA string, and $X=X_1, X_2, X_3, \cdots, X_L$, then X can be accumulated to be a single dimension curve as $f(k)=\sum_{i=0}^{k} X(i)$

By means of combining nucleotide A with the other three nucleotides, we got three maps by: AG walk curve, AC walk curve and AT walk curve.

We use the nucleotide sequences in EMBL database. Choose 300 gene sequences,

*E-mail: ljbao98@163.net

100 from plant, 100 from human and 100 from animal thereinto. We have taken out the intron sequences and exon sequences from each gene. Then string the intron and the exon sequences into two sequences. So we have 600 sequences that can be separated in three classes. We chose 80 sequences as samples from each class at random respectively, and compute the Hurst exponent for each sample. The Hurst exponent is used to describe the long-range correlation. All the results are too great to enumerate. As an example, **Table 1** shows the result of 6 sequences for 3 DNA sequences.

Table 1 The fractal dimensions of different species between exon and intron sequences

Species	Dimension	Exon sequences			Intron sequences		
		AG walk	AC walk	AT walk	AG walk	AC walk	AT walk
Plant	H	0.5240	0.4589	0.5731	0.6789	0.6225	0.7341
	D	1.4760	1.5411	1.4269	1.3211	1.3775	1.2659
Huaman	H	0.5508	0.4694	0.6231	0.5864	0.5853	0.7317
	D	1.4492	1.5306	1.3769	1.4136	1.4147	1.2683
Animal	H	0.5132	0.4382	0.5404	0.6338	0.6033	0.7050
	D	1.4868	1.5618	1.4596	1.3662	1.3967	1.2950

Here H is Hurst exponent, D is fractal dimension, D=2-H.

From the results we find that the dimension of exon sequence is bigger than that of intron sequence for the same DNA sequence. The different nucleotide sequences representations will get different results. In general, the fractal dimension's value is ordered by representing the sequences as AC walk, AG walk and AT walk in turn from great to little. Very few sequences make an exception. It must be indicated that the most important characteristic of fractal objects is fractal dimension. It shows the self-similarity that is present in the DNA sequences. We can speculate that it arises from a dynamic process that controls its evolution. We are getting on the forward research.

Acknowledgments

This work was supported by the Provincial Natural Science Foundation of Guangdong (Contract 990944), and the National Natural Science Foundation of China (Contract 20205003, 29975033).

References

1. C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simon, H. E. Stanley, *Nature*, **1992**, *356*, 168.
2. W. Li, *Int. J. Bifurcation Chaos*, **1993**, *2*, 137.
3. M. Y. Azbel, *Phys. Rev. Lett.*, **1995**, *75* , 168.
4. H. Herzel, I. Groe, *Physica A*, **1995**, *216*, 518.
5. R. F. Voss, *Phys. Rev. Lett.*, **1992**, *68*, 3805.
6. R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, H. E. Stanley, *Phys. Rev. E*, **1995**, *52*, 2939.
7. G. Dodin, P. Vandergheynst, *J. theor. Biol.*, **2000**, *206*, 323.
8. L. J. Bao, J. Y. Mo, Z. Y. Tang, *Anal. Chem.*, **1997**, *69* (15), 3053.

Received 24 June, 2002