Local Scaling Analysis of Nucleotide Sequences in DNA—GenBank Data Analysis

Zhi-Yuan Su[1], Tzuyin Wu[2], Shu-Yin Wang[3]
[1] National Center for High-Performance Computing
[2] Department of Mechanical Engineering, National Taiwan University
[3] Department of Animal Science, Chinese Culture University

Abstract

In our previous study [15], we have proposed a local scaling analysis and exploited to study the myosin heavy chain gene family of different species. The result shows a tendency of clustering of base distribution in myosin heavy chain genes for higher-level species, and an increasing correlation between the coding segments (exons) of the gene and the value of the local scaling exponent (Hölder exponent) α with evolutionary order. Basically, our analysis suggests that the exon parts of the gene of more complicated species are more likely to fall into regions where the values of α are less than one. In this study, we are planning to apply the local scaling analysis to massive amounts of DNA data collected and organized from the GenBank. By comparing our calculation results with known exon locations recorded in the GenBank, it is hoped that a simple method for locating protein-coding regions in a long DNA strand can be devised.

Keywords: DNA sequence, local scaling analysis, Hölder exponent, protein-coding region

1. Introduction

DNA is a long double helical chain composed of a large number of nucleotides, each carrying one of the four bases conventionally symbolized by the four letters: A (adenine), T (thymine), C (cytosine) and G (guanine). The sequential order of these four bases along the DNA chain encodes important genetic information concerning instructions of critical life activities and inheritable features of a living organism.

In the genomes of high eukaryotic organisms, only a small portion of the total genome length is used for protein coding. The segments of the chromosomal DNA that are spliced out during the formation of a mature mRNA are called introns. The coding sequences are called exons.

Recent studies [1-6] have discovered that the nucleotide sequences in DNA exhibit the ubiquity of long-range correlations extending over many decades of base positions. Typically, such correlation feature was analyzed either by Fourier transforming the repetition of the appearance of a particular base along the DNA strand [1-3], or by converting the base sequence into a random-walk process commonly applied to the study of a fluctuating time-series [4-6]. In either approaches, the resulting power spectra measurements exhibit the trend of a power-law behavior. These fluctuations are self-similarity and are now categorized as ‘fractals’ [7].

From a geometric point of view, the sequence of a particular base in the DNA strand can be viewed as a non-uniform distribution of a set of points along a line. It would be better to characterize different parts of this point distribution using different scales. Since the production of a polypeptide chain (protein) depends only on the linear order of bases along the DNA strand, spatial distribution patterns of bases are most naturally scrutinized using the multifractal formalism [7-8]. One of the analytic methods in multifractal formalism is the so-called local scaling analysis, as described in the following section.

2. The Local Scaling Analysis

In this study, a local scaling exponent is introduced to characterize non-uniformity of a fractal distribution. Let l be the size of the covering boxes and $P_i(l)$ be the fraction of points (mass density or probability measure) at the i^{th} – position, then in the limit $l \to 0$ we can define an exponent (singularity strength, or Hölder exponent) α by

$$P_i(l) \propto l^{\alpha} \tag{1}$$

In general, α is not uniformly distributed and hence can be taken as a crowding index for local cluster. α less than one denotes a densely occupied region surrounded by sparse vicinity, while α greater than one represents a less populated region surrounded by dense vicinity.

We begin our analysis by forming a subsequence from DNA strand. There are several rules of forming them, for examples, we can form subsequence of each different base A, T, C, G separately (single base rule); subsequence containing A and G (or C and T) only (purine-pyrimidine rule); subsequence containing A and T (or C and G) only (hydrogen bond rule); etc. Basically, a purine-pyrimidine rule has been adopted in our present study. That is, starting from the beginning of a gene and reading down along the strand, each base position is either filled by a point whenever a purine (A or G) is encountered, or left empty when a pyrimidine (C or T) is met. The resulting sequence can then be treated as a distribution of a set of points (purines) in a one-dimensional line.

The Hölder exponent defined in Eq.(1) is obtained by first calculating the probabilities that a purine falls into different sizes of the covering boxes, $P_i(l)$, and then finding the slope of the log-log plot of $P_i(l)$ vs. l. This slope reflects the invariant scaling nature of the population density of purine bases in a small region centered at position i with that in the vicinities of increasing sizes. Variations in α values with base position i signify changes in the local clustering pattern of purine bases along the DNA strand.

3. Application to Myosin Heavy Chain Gene

First, the type II myosin heavy chain genes belonging to seven different species are analyzed for preliminary testing purpose. The reason for choosing this particular gene is that it represents one of the few gene families whose complete sequences are well documented in the GenBank for a phylogenetically diverse group of organisms, thus providing us good opportunity to look into changes in fractal properties of spatial organization of their components with evolution.

Detail spatial organization and local scaling properties of coding and non-coding segments of this myosin heavy chain gene family are investigated by inspecting the distribution of Hölder exponent α—a crowding index that quantifies the local clustering of base distributions. Figure 1 presents the variation of α from base position 20000 to 25594 along the strand of human myosin heavy chain gene obtained by using different subsequences forming rules. The texts ‘A’, ‘T’, ‘C’ and ‘G’ in Fig.1 denote the resulting subsequences containing nucleotide A, T, C or G, and ‘A+G’ and ‘C+G’ denote the subsequences are formed using the purine-pyrimidine and hydrogen bond rules, respectively. In these plots, irregular fluctuation of the curve is apparent, suggesting once again non-uniformity in base distributions.

Also overlaid on these α curves calculated from the human myosin heavy chain gene are the known positions of exon segments (extracted from the GenBank), just for the purpose of comparing the exon locations and distribution of α value. It is surprising to find the striking feature that most exons appear to be at the locations where the local Hölder exponent α of purine-pyrimidine rule (A+G) is less than one (see Fig.1). If we define a ‘matching ratio’ as the percentage of exons that fall in the region with $\alpha < 1$, then in this case, the matching ratio is $(86.8^{+3.2}_{-4.1})\%$. The $\pm$ errors are estimated from the uncertainties in obtaining the α values from linear fitting. The matching ratios

calculated from different rules are summarized in Table 1.

The biological explanations of this phenomenon are probably those due to the difference in the molecular structures between purines and pyrimidines, a chemically complementary role of purine-pyrimidine or the phenomenon about pyrimidine clusters of intron sequences; the latter is proposed by [9-10] who studied the cluster-size distributions in coding and non-coding DNA sequences. Notice that in the purine-pyrimidine rule of Fig.1, large peaks in α curve are normally found in between exon segments, indicating the existence of large clusters (either pyrimidine or purine) in the non-coding regions.

Table 2 summarizes the results of our local scaling analysis on the myosin heavy chain gene family chosen from seven different species, ranging from yeast to human. It is noticed from all these documentary data that myosin heavy chain gene of higher species has longer total length and more fragmented coding regions that take up less portion of the whole sequence. Apparently, there is also a tendency of increasing correlation between the exon locations and regions in the sequence where $\alpha < 1$ with phylogenetic order. The last column lists the number of exon segments whose matching ratios are greater than 70%. There are 33 out of total 39 exon segments in the human myosin heavy chain gene having good matching ratio. Again, the goodness of matching increases with evolutionary order.

4. GenBank Analysis and Discussion

Besides myosin heavy chain family, we also analyze other DNA sequences of genes from different species. We have analyzed a total of 323 nucleotide sequences of genes from four different species by using the purine-pyrimidine rule and summarized the results in Table 3. In the first row, the number of sequences for four different species is listed. The second and third rows of the table list the total length (measured in bp) of each organism and the length that was actually analyzed. The forth and fifth rows show the number of exon segments in the genes and analyzed ranges. The sixth row lists the average matching ratio of each organism , and the seventh row lists the number of exon segments whose matching ratios are greater than 70. Though the total 176 human's gene have a highest average matching ratio 70.2%，some of them may have low matching values, say about 50%. In Fig.2, for example, we see that some of the exon segments of Human heart/skeletal muscle ATP/ADP translocator (ANT1) gene do not fit well into the valley of $\alpha < 1$. Obviously, different structural properties in coding and non-coding parts of genomic sequences other than the simple purine-pyrimidine rule must also be consulted in the strategy of identifying protein-coding regions.

Recent literature has shown that coding regions tend to be more GC-rich than non-coding ones [11-13], i.e., the exon parts of genes are more likely to fall into regions where the local Hölder exponent α values calculated with hydrogen bond rule (C+G) are less than one. This property can then serve as another useful criterion in locating the exon regions, see e.g., Fig.3. Note that using either purine-pyrimidine rule or hydrogen bond rule to predict protein-coding regions still has the situation of false prediction. However, as shown in Fig.4, it is observed that the exon segments that do not fit into regions of $\alpha < 1$ as calculated with purine-pyrimidine rule could fall into $\alpha < 1$ regions obtained by using the hydrogen bond rule, and vice versa. It is then suggested to combine both purine-pyrimidine and hydrogen bond rules to locate the exon segments. As a result, we found that there are 996 out of total 1142 exon segments in 176 human genes whose matching ratios are greater than 70%, which is more than the original 644 exon segments as predicted by using the purine-pyrimidine rule only. Clearly, the prediction has been improved.

Because looking for regions with $\alpha < 1$ in the sequence formed by either using the purine-pyrimidine (A+G) rule or hydrogen bond rule (C+G) is somewhat equivalent to inspecting the locations where the local

Hölder exponent α of 'T' subsequence is 'greater' than one, this then gives us yet another potentially promising alternative to predict the coding segments in a genomic sequence.

5. Concluding Remarks

Systematic study of spatial organizations of base distributions in genetic sequences chosen from GenBank has been performed using a local scaling analysis technique. It is observed that in higher-order species, the coding segments of genes fit well with $\alpha < 1$ sites of sequences constructed by using the purine-pyrimidine rule, suggesting the formation of pyrimidine clusters in the evolution of intron sequences. The coding segments also fit with the $\alpha < 1$ sites calculated by the hydrogen bond rule. The possible explanation of this phenomenon is either the result of mutation bias, biased gene conversion (BGC), an adaptation to elevated temperature or active gene transcription [14].

Because different species or gene families may have significant differences in structure, using Hölder exponent calculated according to single rule to identify protein-coding regions is not adequate in general. In this paper, we have shown that combination of the purine-pyrimidine (A+G) rule and hydrogen bond (C+G) rule results in a better prediction of the coding segments locations. The local scaling analysis proposed in the present study has the merits of being scale invariant, easy computing, and free of 'window-size' determination that is often encountered in the traditional local-property analyses. When implemented with other statistical techniques, the present local scaling analysis does have the potential to become one of the effective tools to assist bio-informatics for rapid location of possible coding sites in genomic sequences.

Acknowledgment

This study is sponsored by National Science Council of Taiwan under contract no. NSC90-2213-E-002-082

References

1. Voss, R. F., Evolution of Long-Range Fractal Correlations and 1/*f* Noise un DNA Base Sequences, Phys. Rev. Lett., Vol.68, pp.3805-3808, 1992.
2. Voss, R. F., Long-Range Fractal Correlations in DNA Introns and Exons, Fractals, Vol.2, pp.1-6, 1994.
3. Li, W., Marr, T. G., Kaneko,K, Understanding Long-Range Correlations in DNA Sequences, Physica D, Vol.75, pp. 392-416.
4. Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H. E., Long-Range Correlations in Nucleotide Sequences, Nature, Vol.356, pp.168-170, 1992.
5. Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Hausdorff, J. M., Havlin, S., Mietus, J., Peng, C.-K., Sciortino, F. and Simons, M., Fractal Landscapes in Biological Systems: Long-Range Correlations in DNA and Interbeat Heart Intervals, Physica, A, Vol.191, pp.1-12. 1992.
6. Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Ossadnik, S. M., Peng, C.-K. and Simons, M., Fractal Landscapes in Biological Systems, Fractals, Vol.1, pp.283-301. 1993.
7. Mandelbrot, B. B., The Fractal Geometry of Nature, Freeman, New York, 1983.
8. Halsey, T. C., Jensen, M. H., Kadanoff, L. P., Procaccia, I. and Shraiman, B. I., Fractal Measures and Their Singularities: The Characterization of Strange Sets, Phys. Rev. A, Vol.33, No.2, pp.1141-1151, 1986.
9. Provata, A. and Almirantis, Y., Scaling Properties of Coding and Non-coding DNA Sequences, Physica A, Vol.247, pp.482-496, 1997.
10. Provata, A., Random Aggregation Models for the Formation and Evolution of Coding and Non-coding DNA, Physica A, Vol. 264, pp.570-580, 1999.
11. Burset, M. and Guigo, R., Evaluation of Gene Structure Prediction Programs, Genomics, Vol. 34, pp. 353-367, 1996.
12. Bernardi, G. and Bernardi, G., Compositional

Transitions in the Nuclear Genomes of Cold-Blooded Vertebrates, J. Mol. Evol., Vol. 31, pp. 282-293, 1990.

13. Vinogradov, A. E., DNA Helix: The Importance of Being GC-Rich, Nucleic Acids Research, Vol. 31, pp. 1838-1844, 2003.
14. Vinogradov, A. E, Isochores and Tissue-Specificity, Nucleic Acids Research, Vol. 31, pp.5212-5220, 2003.
15. Wu, T., Su, Z.-Y. and Wang, S.-Y., Multifractal and Local Scaling Analyses of Nucleotide Sequences in DNA, The 20th National Conference on Mechanical Engineering, the Chinese Society of Mechanical Engineers, National Taiwan Univ., Taipei., 2003.

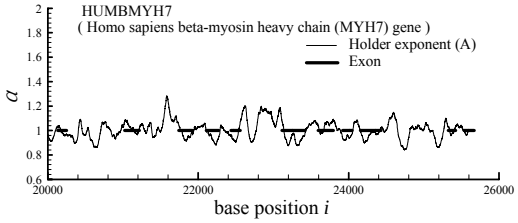
DNA 序列之局部碎形分布比例指數分析－GenBank 基因資料庫資料分析

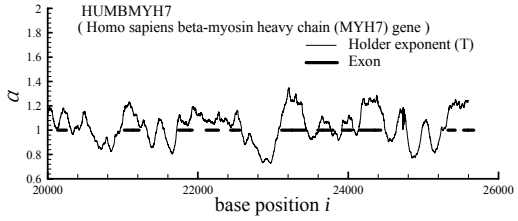
蘇致遠[1]　伍次寅[2]　王淑音[3]
[1]國家高速網路與計算中心
[2]國立台灣大學機械工程學系
[3]中國文化大學畜產學系

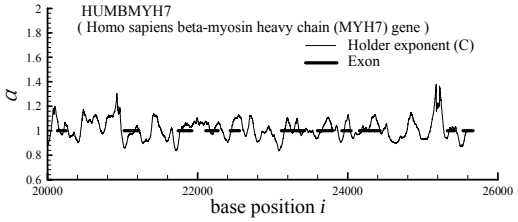
摘要

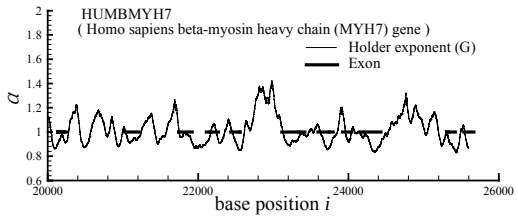
在先前的研究我們從近代發展出來的碎形理論(Fractal Theory)出發，針對 Myosin Heavy Chain 基因家族裡屬於各個不同等級生物體的一系列 DNA 序列進行局部碎形尺度比例指數分析，研究結果發現 DNA 長鏈中帶遺傳訊息片段(Exon)有落在 DNA 序列 A、G 鹼基分布之局部碎形尺度比例指數(Hölder Exponent)小於 1 地方的趨勢，因此計算局部碎形尺度比例指數並尋找其小於 1 區域的方式將可以用來辨認 DNA 序列中會譯製成蛋白質的片段。在接續的研究中，我們更進一步發現不只 Myosin Heavy Chain 基因家族具有這樣的特性。若是如此，運用簡單、所需計算時間短以及具備尺度不變性等多項優點的局部碎形尺度比例指數來辨認 DNA 序列中會譯製成蛋白質的片段將是值得深入研究的課題。因此在本文中，我們將以先前的研究成果為基礎，預計從基因資料庫 GenBank 中收集整理出分屬不同演化層次，總數可能達數百的基因 DNA 序列，對其進行大量且有系統的碎形分布尺度比例指數分析，希望能從分析結果中歸納整理出可以簡易辨識 DNA 序列中會譯製成蛋白質片段的模式。

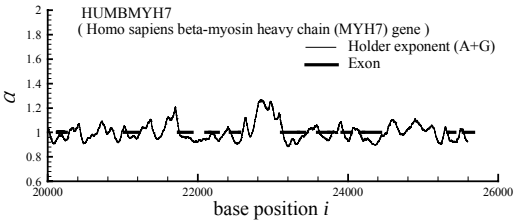
關鍵字：DNA 序列，局部碎形分布尺度比例指數，Hölder 指數，蛋白質編碼區域預測











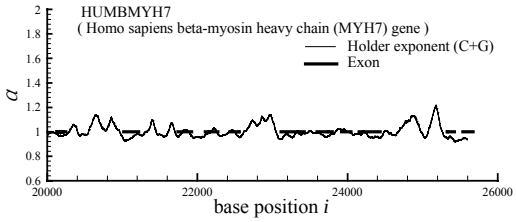


Fig.1 Comparison of exons locations and distribution of α value (human myosin heavy chain gene) calculated by using different rules.

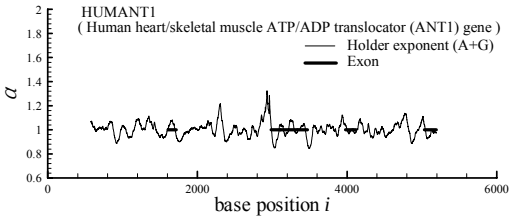


Fig.2 Comparison of exons locations and distribution of α value (Human translocator gene) calculated with A+G rule..

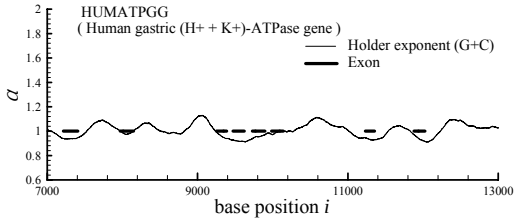


Fig.3 Comparison of exons locations and distribution of α value (Human gastric ATPase gene) calculated with C+G rule.

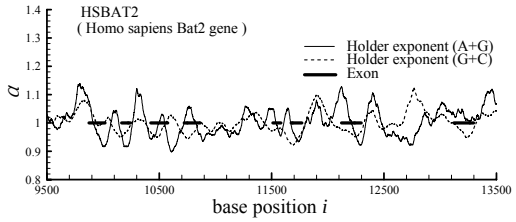


Fig.4 Comparison of exons locations and distributions of α value (human Bat2 gene) calculated by A+G and C+G rules.

Table 1. Matching ratio and it uncertainties

single base rule	
% α <1 in exon / Type	matching ratio
A	$68.5^{+3.6}_{-3.5}$
T	$9.3^{+2.9}_{-3.3}$
C	$40.4^{+6.0}_{-5.7}$
G	$82.1^{+2.4}_{-4.0}$
purine-pyrimidine and hydrogen bond rule	
% α <1 in exon / Type	matching ratio
A+G	$86.8^{+3.2}_{-4.1}$
C+G	$79.4^{+4.3}_{-4.8}$

Table 2. Results of local scaling analysis (A+G) on myosin heavy chain genes of different species.

Family Organism	Length analyzed(bp) (total length)	# of exon segments (in analyzed region)	% exon length	% α <1 in exon	# of exon segments whose % α <1 are greater than 70
yeast	4889 (6108)	1 (1)	100	50.0+5.8 -5.7	0
C. elegans #3	9285 (11604)	7 (7)	50.9	53.0+6.7 -7.8	4
Brugia malayi	9415 (11766)	13 (11)	47.6	64.8+5.6 -6.1	6
fruit fly	18132 (22663)	30 (27)	35.4	67.7+5.6 -6.2	12
rat	20606 (25755)	41 (38)	23.4	72.8+2.6 -4.3	28
chicken	24890 (31111)	38 (34)	18.7	75.8+4.5 -5.2	21
human	22752 (28438)	40 (39)	21.1	86.8+3.2 -4.1	33

Table3. Results of local scaling analysis (A+G) on different genes of different species.

species	C. elegans	Birds	Rodents	Human
# of sequences	70	34	43	176
Total length	438192	291257	258799	1904144
Total length analyzed(bp)	350688	233069	207118	1523637
Total exon segment	509	264	261	1309
Total exon segment in analyzed region	477	247	244	1142
average %α<1 in exon of species	60.4%	68.1%	66.1%	70.2%
# of exon segments whose % α<1 are greater than 70	184	137	115	644