Multifractal and Local Scaling Analyses of Nucleotide Sequences in DNA

Tzuyin Wu[1], Zhi-Yuan Su[1], Shu-Yin Wang[2]

[1]Department of Mechanical Engineering
National Taiwan University, Taipei 106, Taiwan

[2]Department of Animal Science
Chinese Culture University, Taipei 111, Taiwan

NSC Project No.: NSC90-2213-E-002-082

Keywords: DNA sequence, local scaling analysis, Hölder exponent, multifractal, spectrum of singularities, MHC gene

ABSTRACT

The fractal nature of DNA sequences is explored by employing the singularity spectral analysis (multifractal analysis). The multifractality character of the nucleotide sequences along the DNA strand is revealed. Also, a local scaling analysis is devised and exploited to study the myosin heavy chain (MHC) gene family of different species. The result shows a tendency of clustering of base distribution in MHC 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.

INTRODUCTION

DNA is a long double helical chain composed of a large number of nucleotides, each carrying one of the four bases (either purines or pyrimidines) 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. There are, for instance, roughly a total three billion base pairs in the complete genomic sequences of a human, implying a tremendous amount of hereditary information possibly carried by DNA.

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. That is, the appearance of a particular base in a DNA sequence depends virtually to some extent on the bases at a large distance ahead of it. 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 similar to the so called $1/f^{\beta}$ —noise found in many naturally occurring fluctuations. These fluctuations are actually a time-scale analogy to many scale-invariant geometric configurations, such as mountain profiles, coastlines and fern leaves which possess the property of self-similarity (small portions resemble the structure in whole when magnified) and are now categorized as ‘fractals’ [7]. Although there has been a debate about the existence of correlations between bases in DNA sequences [8-13], increasing evidences show that long-range correlations are intimately tied with non-coding regions (intergenetic sequences or introns) of the DNA [14-19]. In coding regions (genes or exons), the base sequences are less correlated and appear to be more random-like. Origin of this long-range correlation and its biological implications are still not well understood at this moment. Some [20,21] suggest that it is simply a result of repetition of short sequences in the non-coding region of DNA, while others [22] believe it might be related to the dynamical mechanisms (mutation, transposition, replication, insertion, substitution, alternative splicing, etc.) developed in the evolutionary process.

From a geometric point of view, the sequence of a particular base in the DNA strand can be viewed as a distribution of a set of points along a line. It is common that naturally evolving systems are seldom characterized by a single scaling ratio; different parts of a system may be scaling differently. That is, the clustering pattern is not uniform over the whole system. Such a system is better characterized as a 'multifractal' [7, 23]. Distributions of stars in a galaxy and epicenters of earthquakes in a seismically active region are just two examples. The notion is in the same way applied here to the study of DNA sequences. 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. Specifically, the type II myosin heavy chain (MHC) genes belonging to seven different species are analyzed in this study. Local scaling properties of coding and non-coding segments of this MHC gene family are also investigated by examining the Hölder exponent—a crowding index that quantifies the local clustering of base distributions. 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.

MULTIFRACTAL FORMALISM

Basically, the multifractal formalism 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) in the i^{th} – box, 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. If we count the number of boxes $N(\alpha)$ where the probability measure P_i has singularity strength between α and $\alpha + d\alpha$, then $f(\alpha)$ can be loosely defined as the fractal dimension of the set of boxes with singularity strength α by [23]

$$N(\alpha) \propto l^{-f(\alpha)}. \tag{2}$$

The formalism thus describes a multifractal measure in terms of interwoven sets of different singularity strengths α, each characterized by its own fractal dimension $f(\alpha)$.

Another useful multifractal formalism is the so-called generalized dimension defined as [24,25]

$$D_q = \lim_{l \to 0} \frac{1}{q-1} \frac{\log \sum_i P_i^q(l)}{\log l} \tag{3}$$

where the probability P_i is raised to the power of q. Thus different values of q emphasize the distribution with different degrees of clustering vicinities. In a point distribution set, the D_q with the limit $q \to +\infty$ is associated with the fractal dimension of most densely occupied regions in the set, while D_q with $q \to -\infty$ is associated with the fractal dimension of least populated regions. This formalism then quantifies non-uniformity of a distribution based on the statistical moments of its probability measure.

As the generalized dimension D_q is easier to compute, conventionally, the multifractal spectrum $f(\alpha)$ is usually evaluated from D_q via a Legendre transformation [23]:

$$\begin{aligned} \tau(q) &= (q-1)D_q \\ f &= q\alpha - \tau \\ \alpha &= d\tau/dq \end{aligned} \tag{4}$$

However, as pointed out in [26,27], validity of the Legendre transformation relies on the smoothness of functions $f(\alpha)$ and D_q. In the attempt to obtaining D_q from scaling the probability measures P_i^q with box sizes l, naturally evolving and experimentally observed data often produce a log-log plot featured by oscillations and scatters rather than showing a perfect linear behavior, especially when q is large. This then results in a D_q curve with large uncertainties. Applying Legendre transformation to such a curve may generate false result and make the error estimation in the $f-\alpha$ formalism a difficult task. To circumvent this pitfall, a direct determination of $f(\alpha)$ was proposed by [26,27]. The method involves first constructing a one-parameter family of normalized measures $\mu(q)$ at each box i from probabilities $P_i(l)$:

$$\mu_i(q,l) = \frac{\left[P_i(l)\right]^q}{\sum_j \left[P_j(l)\right]^q}, \tag{5}$$

then $f(\alpha)$ is just the Hausdorff dimension of the measure-theoretic support of $\mu(q)$, which is given by

$$f(q) = \lim_{l\to 0} \frac{\sum_i \mu_i(q,l)\log \mu_i(q,l)}{\log l}. \tag{6}$$

The value of the singularity strength α, averaged with respect to $\mu(q)$, can be computed by

$$\alpha(q) = \lim_{l\to 0} \frac{\sum_i \mu_i(q,l)\log P_i(l)}{\log l}. \tag{7}$$

Equations (6) and (7) provide an alternative definition of the singularity spectrum which can be used to obtain $f(\alpha)$ directly from real-world data without appealing to the Legendre transformation. The method is used in the subsequent calculations of this study.

APPLICATION TO MHC GENE

We begin our analysis by forming a subsequence containing purines (or pyrimidines) only. 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 purines sequence is treated as a distribution of a set of points in a one-dimensional line. We then study the spatial pattern of this set of points using the multifractal formalism (6) and (7). There are also other possible rules of forming subsequences, for examples, we can form subsequence of each different base A, T, C, G separately (single base rule); subsequence containing A and T (or C and G) only (hydrogen bond rule); etc. In general, we find that the original purine-pyrimidine rule provides the most substantial results, probably due to the difference in the molecular structures between purines and pyrimidines, and a chemically complementary role of purine-pyrimidine.

Figure 1 plots the singularity spectrum ($f-\alpha$ curve) of human cardiac β -myosin heavy chain gene (GenBank accession # M57965) of total length (# of base pairs) 28438. The familiar, inverted, downward-opening parabola shape of curve is seen. The cross bars on the symbols represent uncertainties in the values of α and f arising from the linear fitting procedure in (6) and (7). The wide opening $(\alpha \sim 0.83-1.4)$ of the parabola indicates that purine bases are not uniformly distributed along the human MHC gene; rather, they tend to form clusters of different sizes. To confirm this heterogeneity in base distribution, we scramble the positions of these bases by a random scheme, and plot the resulting $f(\alpha)$ spectrum in the same figure. A much smaller opening curve is shown, indicating that the base sequence after scrambling has a more uniform distribution. The remaining slight opening $(\alpha \sim 0.94-1.08)$ of the curve is perhaps due to the so-called 'strand bias' (there are slightly more purines than pyrimidines in MHC) normally observed in genomes.

Detail spatial organization of the nucleotides sequence can be further analyzed by inspecting the distribution of the singularity strength α. The Hölder exponent defined in Eq.(1) 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. In Fig.2, for example, 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.

Figure 3 shows a typical log-log plot of purine base populations $P_i(l)$ vs. sizes of boxes centered at base position $i = 4259$ of human MHC gene. The smallest size of the box has a width of just a few base pairs (bp), while the largest box can extend to a length of a few thousand bp. To avoid edge effect, the largest size of the box is limited to 1/5 of the total bp in the chain, in this case, about 5600 bps. The local Hölder exponent α is then obtained from linear fitting the data points within this range. The error in α is estimated from the standard deviation of fitted data from the linear slope.

Figure 4 presents the variation of α along the entire strand of human MHC gene. Irregular fluctuation of the curve is apparent, suggesting once again non-uniformity in base distributions. Note that in this case we did not analyze α from the very beginning of the sequence. Instead, the analysis was initiated at about 1/10 position of the sequence. The reason is to eliminate the edge effect that will enter into analysis as the largest box size extending beyond the ends of the sequence. Similarly, the last 1/10 portion of the sequence was not analyzed for the same reason.

In Fig.5, we delete all introns and stitch together the remaining exon segments of the original MHC to form a shorter sequence containing protein-coding region only. The result shows a relatively less fluctuating α curve. The corresponding $f(\alpha)$ spectrum is given in Fig.6. Comparing with the spectrum in Fig.1, a much narrower $f(\alpha)$ is observed. Scrambling the sequence produces little difference in $f(\alpha)$, implying that the protein-coding sequence has a more uniform and random-like base distribution than the original intron-rich sequence. This observation is consistent with previous findings based on random-walk model that long-range correlation is associated with intron parts of DNA sequence.

FURTHER RESULTS AND DISCUSSION

When we overlay the known positions of exon segments (extracted from the GenBank) on the α curve calculated from the human MHC gene, it is surprising to find the striking feature that most exons appear to be at the locations where the local Hölder exponent α is less than one (see Fig.7). Some of the segments may even fit perfectly into the valley of $\alpha < 1$. If we define a 'matching ratio' as the percentage of exon that fall in the region with $\alpha < 1$, then in this case, the matching ratio is $(86.8^{+3.2}_{-4.1})\%$. The ± errors are estimated from the uncertainties in α values (see Fig.3), which then lead to an up- or downward shifting of the α curve and hence changing the matching ratio.

Table 1 summarizes the results of our local scaling analysis on the MHC gene family chosen from seven different species, ranging from yeast to homo sapiens (human). The third column of the table lists the total length (measured in bp) of the MHC gene of each organism and the length that was actually analyzed. The fourth column shows the number of exon segments in the genes and in the fifth column, the percentage of length exons occupy. It is noticed from all these documentary data that MHC gene of higher species has longer total length and more fragmented coding regions that take up less portion of the whole sequence. The last column lists the matching ratio in an ascending order. Apparently, there is a tendency of increasing correlation between the exon locations and regions in the sequence where $\alpha < 1$ with phylogenetic order. In order to assess whether such a trend could be resulted from a possible bias in α value (e.g., there are probably more places with $\alpha < 1$ in the genetic sequence of higher species), we list in the sixth column of Table 1 the percentage of length in the sequence with $\alpha < 1$. All species have roughly the same percentage value; the maximum difference is about 4% only. Yet the matching ratio is 50% in yeast and 87% in human. Obviously, the result is not fortuitous.

The biological explanation of this phenomenon is not well established at this moment. However, our present results do lend support to the findings of [28,29] who studied the cluster-size distributions in coding and non-coding DNA sequences. Notice that in Fig.7, 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. This is in consistence with the claim made in [28,29] that the power-law behavior of the base sequence is associated with the tendency of large pyrimidine and purine cluster formation in the non-coding regions. The possible evolutionary mechanisms for such formation may need further study. Nevertheless the present scaling analysis, when implemented with other statistical techniques, does have the potential to become one of the effective tools for rapid location of possible coding sites in genomic sequences.

Figure 8 compares the singularity spectra obtained from the seven species. The gradual opening of the $f(\alpha)$ curve suggests an increasing complexity in the structures of DNA sequences. Again, the degree of complexity is shown to follow the evolutionary order.

CONCLUSION

A multifractal formalism has been employed to investigate the fractal nature of DNA sequences. Phylogenetic study of spatial organizations of base distributions in MHC gene family has also been performed using a local scaling analysis technique. Both the singularity spectra $f(\alpha)$ and local Hölder exponent distribution curves ($\alpha -$curve) display a tendency of increasing non-uniformity and clustering of bases in the structures of DNA chains with evolution. Furthermore, it is observed that coding segments of genes fit well with the $\alpha < 1$ sites of chains in higher-order species, suggesting the formation of pyrimidine clusters in the evolution of intron sequences. Present findings may point to an important direction for better understanding of the functional roles of evolutionary mechanisms in terms of structural properties in coding and non-coding parts of 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. Li, W. and Kaneko, K., Long-Range Correlation and Partial 1/*f* Spectrum in a Noncoding DNA Sequence, Europhys. Lett., Vol.17, pp.655-660.
9. Prabhu, V. V. and Claverie, J. M., Correlations in Intronless DNA, Nature, Vol.359, p.782, 1992.
10. Larhammar, D. and Chatzidimitrious -Dreismann, C. A., Biological Origins of Long-Range Correlations and Compositional Variations in DNA, Nucleic Acids Res., Vol.21, pp.5167-5170, 1993.
11. Chatzidimitrious-Dreismann, C. A., Streffer, R. M. F., Larhammar, D., Variations in Base Pair Composition and Associated Long-Range Correlations in DNA Sequences–Computer Simulation Results, Biochimica et Biophysica Acta, Vol.1217, pp.181-187, 1994.
12. Nee, S., Uncorrelated DNA Walks, Nature, Vol.357, p.450, 1992.
13. Karlin, S. and Brendel, V., Patchiness and Correlations in DNA Sequences, Science, Vol.259, pp.677-680, 1993.
14. Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Stanley, H. E., Stanley, M. H. R. and Simons, M, Fractal Landscapes and Molecular Evolution: Modeling the Myosin Heavy Chain Gene Family, Biophys. J., Vol.65, pp.2673-2679, 1993.
15. Ossadnik, S. M., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Peng, C.-K., Simons, M. and Stanley, H. E., Correlation Approach to Identify Coding Regions in DNA Sequences, Biophys. J., Vol.67, pp.64-70, 1994.
16. Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. and Goldberger, A. L., Mosaic Organization of DNA Nucleotides, Phys. Rev. E, Vol.49, pp.1685-1689, 1994.
17. Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Goldberger, Z. D., Havlin, S., Mantegna, R. N., Ossadnik, S. M., Peng, C.-K. and Simons, M., Statistical Mechanics in Biology: How Ubiquitous are Long-Range Correlations?,

Physica A, Vol.205, pp.214-253, 1994.

18. Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsa, M. E., Peng, C.-K., Simons, M. and Stanley, H. E., Long-Range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis, Phy. Rev. E, Vol.51, pp.5084-5091, 1995.
19. Li, W., The Study of Correlation Structures of DNA Sequences: A Critical Review, Comput. Chem., Vol.21, pp.257-271, 1997.
20. Li, W., Generating Nontrivial Long-Range Correlations and 1/*f* Spectra by Replication and Mutation, Int. J. Bifur. Chaos in Appl. Sci. Eng., Vol.2, pp.137-154, 1992.
21. Buldyrev, S. V., Dokholyan, N. V., Havlin, S., Stanley, H. E., Stanley, R. H. R., Expansion of Tandem Repeats and Oligomer Clustering in Coding and Noncoding DNA Sequences, Physica A, Vol. 273, pp.19-32, 1999.
22. Provata, A. and Almirantis, Y., Statistical Dynamics of Clustering in the Genome Structure, J. Stat. Phys., Vol.106, pp.23-56, 2002.
23. 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.
24. Grassberger, P., Generalized Dimensions of Strange Attractors, Phys. Lett. A, Vol.97, pp.227-230, 1983.
25. Hentschel, H. and Procaccia, I., The Infinite Number of Generalized Dimensions of Fractals and Strange Attractors, Physica D, Vol.8, pp.435-444, 1983.
26. Chhabra, A. V. and Jensen, R. V., Direct Determination of the $f(\alpha)$ Singularity Spectrum, Phys. Rev. Lett., Vol.62, No.12, pp.1327-1330, 1989.
27. Chhabra, A. B., Meneveau, C., Jensen, R. V. and Sreenivasan, K. R., Direct Determination of the $f(\alpha)$ Singularity Spectrum and Its Application to Fully Developed Turbulence, Phys. Rev. A, Vol.40, No.9, pp.5284-5294, 1989.
28. Provata, A. and Almirantis, Y., Scaling Properties of Coding and Non-coding DNA Sequences, Physica A, Vol.247, pp.482-496, 1997.
29. Provata, A., Random Aggregation Models for the Formation and Evolution of Coding and Non-coding DNA, Physica A, Vol.264, pp.570-580, 1999.

FIGURES AND TABLES

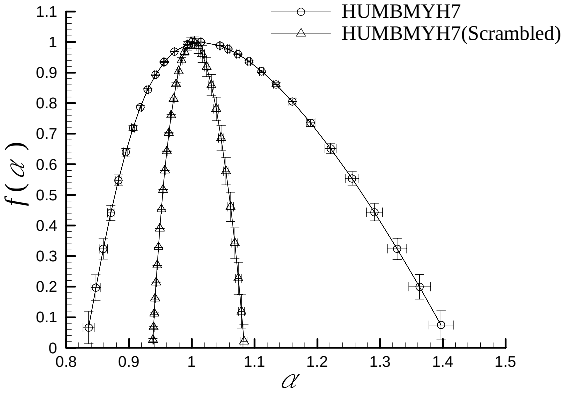


Fig.1 Singularity spectrum of human cardiac β -myosin heavy chain (MHC) gene.

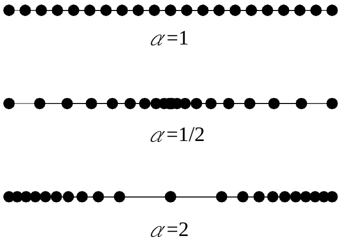


Fig.2 Fractal distributions with different α values.

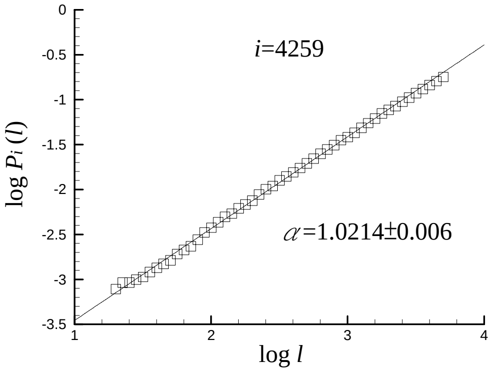


Fig.3 Scaling of $P_i(l)$ with box size *l*.

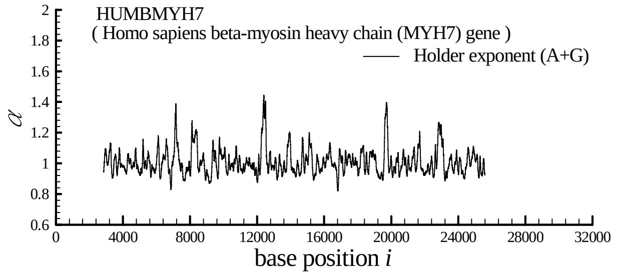


Fig.4　Distribution of Hölder exponent α for human MHC gene.

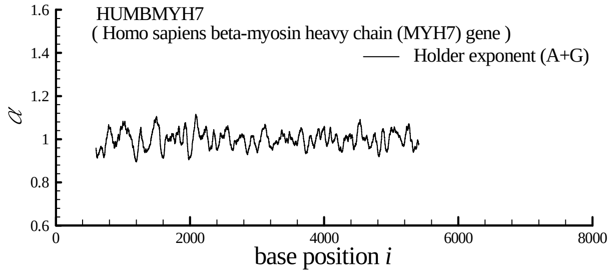


Fig.5　Distribution of Hölder exponent α for human MHC gene with introns removed.

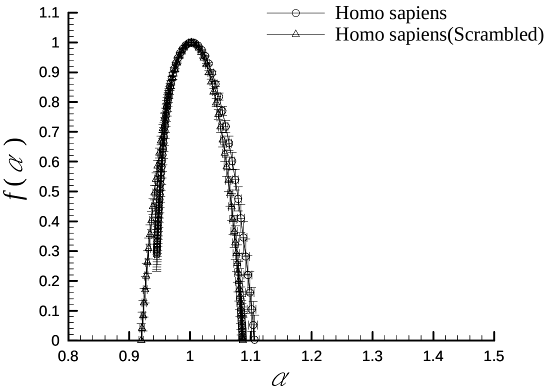
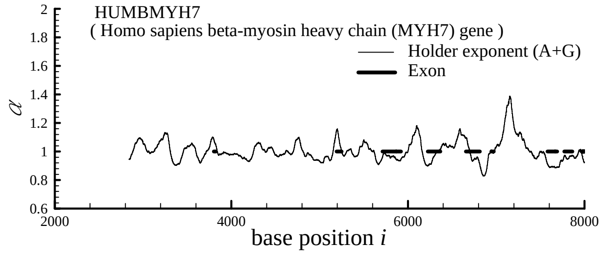


Fig.6　Singularity spectrum of human MHC gene with introns removed.



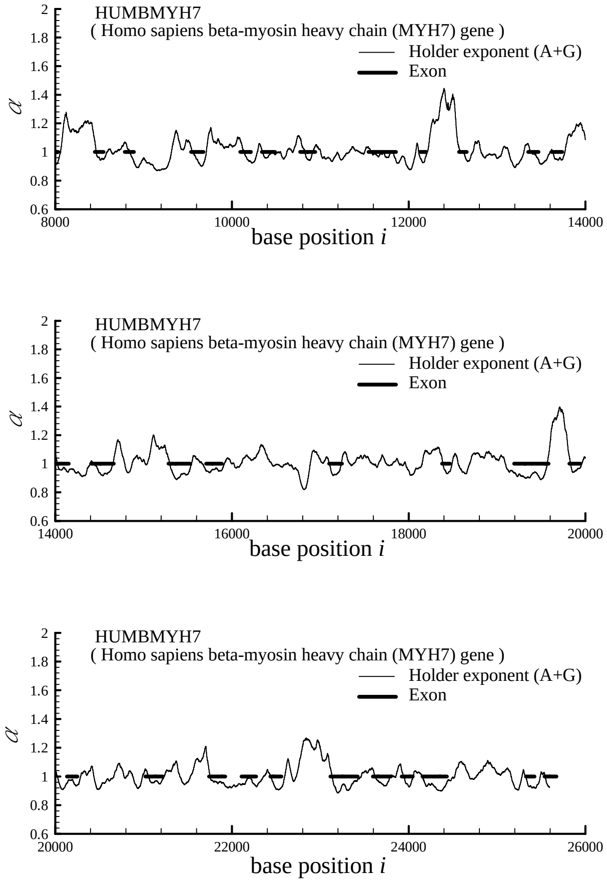


Fig.7　Comparison of exons locations and distribution of α value (human MHC gene).

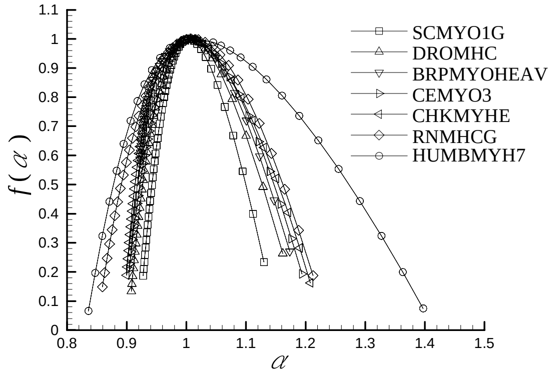


Fig.8　Singularity spectra for MHC genes of different species.

Table 1. Results of local scaling analysis on MHC genes of different species.

Family Organism	GenBank Accession # (locus)	Length analyzed(bp) (total length)	# of exon segments	% exon length	%α<1 in analysis	%α<1 in exon
Saccharomyces cerevisiae (yeast)	X53947 (SCMYO1G)	4889 (6108)	1	100	50.0	50.0+5.8 -5.7
Caenorhabditis elegans#3 (worm)	X08067 (CEMYO3)	9285 (11604)	7	50.9	51.6	53.0+6.7 -7.8
Brugia malayi (worm)	M74000 (BRPMYOHEA)	9415 (11766)	13	47.6	52.3	64.8+5.6 -6.1
Drosophila melanogaster (fruit fly)	M61229 (DROMHC)	18132 (22663)	30	35.4	52.2	67.7+5.6 -6.2
Rattus norvegicus (rat)	X04267 (RNMHCG)	20606 (25755)	41	23.4	50.1	72.8+2.6 -4.3
Gallus gallus (chicken)	J02714 (CHKMYHE)	24890 (31111)	38	18.7	50.9	75.8+4.5 -5.2
Homo sapiens (human)	M57965 (HUMBMYH7)	22752 (28438)	40	21.1	54.3	86.8+3.2 -4.1