嘉南藥理科技大學 96 年度教師研究計畫成果報告

以局部碎形分布比例指數分析法分析 GenBank 基因資料庫之 DNA 序列

計畫類別：個人型計畫
計畫編號：CN9637
執行期間：96 年 1 月 1 日至 96 年 12 月 31 日

計畫主持人：蘇致遠
計畫參與人員：
成果報告類型：精簡報告
處理方式：本計劃可公開查詢
執行單位：資訊管理系
中　華　民　國　97　年　3　月　3　日

Catalog

Abstract

Base sequences of deoxyribonucleic acid (DNA) in an organism carry all the instructions regarding its growth and development. On the surface, such sequences seem irregular; yet in reality, they are symbolic sequences with an organized structure. This study investigates the characteristics of base arrangement and distribution in DNA sequences from the fractal theory viewpoint. In addition to multifractal features demonstrated by the DNA sequence, this study also compares the multifractal spectra derived from a particular family of gene among several different species. The results reveal that a considerable correlation exists between base distribution and evolutionary order. Furthermore, local scaling exponent (Hölder exponent) differences between coding segments (exon) and non-coding segments (intron) are also examined. It is suggested that such differences in the local distribution of bases can be applied to find coding segments within the DNA sequence that is to be translated into protein. This local scaling analysis is feasible and has the potential to become an effective tool for rapid location of possible coding sites in DNA sequences. The authors hope that future studies using more complicated bioinformatics methods for analyzing DNA sequences can benefit from this study.

Keywords: DNA sequence, local scaling analysis, Hölder exponent, multifractal.

1. Introduction

Complex DNA base sequence studies began twenty years ago. In the 1980s, Gates started to map bases in a DNA sequence onto a two-dimensional space [1]. In the 1990s, Peng *et al.* [2] published a paper in the *Nature* journal. It was the first time that DNA base sequences transformed into a one-dimensional irregular walk, named "DNA walk." One-dimensional DNA walk can be defined as: if pyrimidine is located in the *i*-th position along the DNA chain, then the walker steps up ($u(i) = +1$); if purine is found, then the walker steps down ($u(i) = -1$). Calculate the sum of $u(i)$ after *l* steps and a curve *y*(*l*) for DNA walk can be obtained. DNA walk not only provides a simple and intuitional representation of each DNA sequence but also assists in analyzing correlations among DNA base sequences through calculating root mean square fluctuation *F*(*l*) of average displacement. *F*(*l*) can be used to distinguish three types of behaviors: (1) if nucleotide sequence is random, then $F(l) \sim l^{1/2}$; (2) if local correlation can be found within a certain characteristic range (such as Markov chain) with asymptotic behavior to a purely random sequence, then, eventually $F(l) \sim l^{1/2}$ for large *l*; (3) if characteristic length does not exist, then the relation between fluctuation *F*(*l*) and *l* is power law, $F(l) \sim l^{\delta}$, with $\delta \neq 1/2$, indicating that DNA walk features self-similarity, or fractal property [3]. Peng *et al.* [2] analyzed nucleotide sequences of genes and discovered a long-range correlation among genes abundant with intron, counting up to thousands of bases. On the contrary, for cDNA and genes with scarce intron, $\delta \approx 1/2$; in other words, long-range correlation is not observed. A major obstacle to overcome in long-range correlation analysis is that the DNA sequence's mosaic structure causes the four bases A (adenine), G (guanine), T (thymine), C (cytosine) strand biases and "trend" presence in a DNA walk [4–6]. Peng [2] proposed a Min-Max Method to overcome this problem, which unfortunately required determination on the number of local maximum and minimum in a DNA walk curve, suffering from practical application difficulties. Therefore, Peng *et al.* [7] later proposed the Bridge Method and

Detrended Fluctuation Analysis, or DFA. In 1993, Buldyrev *et al.* [8] used DFA to investigate genes in the myosin heavy chain family to examine the relation of base sequence fractal complexity to evolution. The use of single gene family can avoid potential bias caused by different evolutionary pressures and various base compositions among irrelevant gene families. Their analytical results reveal that, for species from eukaryotes to invertebrates to vertebrates, the correlation coefficient features monotonic increase. After that, Ossadnik *et al.* [9] proposed the Coding Sequence Finder Algorithm (CSF), based on DFA. First, a certain "window size" is selected. Then the DNA sequence is divided into several regions, which are respectively measured for individual DFA exponents. The results reveal that the local value of the DFA exponent typically displays minima where genes are suspected. Reviews about DFA's applications can be referred to in the bibliography [10].

DNA walk is very intuitional, but for the nucleotide sequences, it involves introducing artificial correlations into DNA walks of less than four dimensions. To avoid introducing artificial correlation, Voss [11] proposed a simple method that separates a DNA base sequence into four sub-sequences of A, G, T, and C, each of which stands for base location in the original DNA sequence. For example, sub-sequence A is a numeric sequence obtained by replacing A with 1 and replacing the other three bases with 0 in the original DNA nucleotide sequence, resulting in a time sequence $x_A(n)$ that oscillates with position coordinate *n*. In this manner, time sequences for other bases can be defined as $x_G(n)$, $x_T(n)$, and $x_C(n)$. Through Fourier transform, time sequences of these bases can be illustrated into a power spectra $S_k(f)$, showing that basically they share similar tendency. In cases of larger *f*, $S_k(f)$ is mostly white noise. Under low frequency, $S_k(f)$ demonstrates a power law $S_k(f) \sim 1/f^{\beta}$. In many cases, $1/f^{\beta}$ noise is assumed independent of high-frequency white noise, so that high-frequency white noise can be deducted from $S_k(f)$, making the long-range correlation within a low-frequency area more obvious. The Voss approach is

special in that the original DNA base sequence's correlation function can be regarded as the sum of correlation functions of the four sub-sequences, $x_A(n)$, $x_G(n)$, $x_T(n)$, and $x_C(n)$. This conclusion also applies to spectral density. Thus, by figuring out the spectral densities of all sub-sequences, the spectral density of original DNA sequence can be obtained without assuming any correlation among the four bases.

Besides aforementioned approaches, other DNA sequence analyses include general entropy function calculation [12], the Zipf linguistics method [13–14], Hurst exponent analysis [15–16], Hao's geometric representations [17–19], two-dimensional DNA walk [20], nonlinear prediction method [21], cluster-size distribution [22–24] and so on. However, most of research methods focus on the overall properties of DNA sequences. The power spectra, for example, depict the average distribution of energy under each oscillation frequency contained in the sequence. Therefore, local energy variations in the oscillation frequency of the sequence cannot be obtained. An average component energy distribution for oscillations presented in natural phenomena is very rare, and this is true for DNA sequences. From a geometric point of view, the sequence of a particular base in a DNA strand can be viewed as a distribution of a set of points along a line. 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" system [3, 25]. This study applies that notion in the same way to the study of DNA sequences. Since the production of a polypeptide chain (protein) only depends on the linear order of bases along the DNA strand, spatial distribution patterns of bases are scrutinized using multifractal formalism.

2. Local Scaling and Multifractal Analyses

2.1 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 $l \to 0$, an exponent (singularity strength, or Hölder exponent) α can be defined by

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

In general, α is not uniformly distributed and can therefore serve as the crowding index of a local cluster. If the number of boxes $N(\alpha)$ where the probability measure P_i has a singularity strength between α and $\alpha + d\alpha$ is counted, then $f(\alpha)$ can be loosely defined as the fractal dimension of the set of boxes with singularity strength α by [25]

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

This formalism thus describes a multifractal measure in terms of interwoven sets of different singularity strengths α, where each set is characterized by its own fractal dimension $f(\alpha)$. Another useful multifractal formalism is the so-called generalized dimension, defined as [26–27]

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

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

moments of its probability measure.

As the generalized dimension D_q is computed, the multifractal spectrum $f(\alpha)$ is usually evaluated from D_q *via* a Legendre transformation [25]:

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

However, as other studies indicate [28–29], the validity of the Legendre transformation relies on the smoothness of functions $f(\alpha)$ and D_q. In the attempt to obtain D_q by scaling the probability measures P_i^q with box sizes l, naturally evolving and experimentally observed data often produce a log–log plot featuring oscillations and scattering rather than perfect linear behavior, especially when the value of q is large. This then produces a D_q curve with large uncertainties. Applying the Legendre transformation to such a curve may generate false results and make error estimation in the $f-\alpha$. Previous studies propose a direct determination of $f(\alpha)$ to circumvent this pitfall. This method first involves 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 simply the Hausdorff dimension of the measure-theoretic support of $\mu(q)$, which is given by

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

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

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

Equations (6) and (7) provide an alternative definition of the multifractal spectrum, which can be used to obtain $f(\alpha)$ directly from real-world data without using the Legendre transformation. Subsequent calculation in make use of this method.

2.2 Calculations of Hölder exponent α for DNA sequences

Featuring DNA's structure by the variation of local scaling exponent curve (or Hölder exponent α) according to the location of base in DNA chain is the idea of this study. Since exon and intron in the DNA sequences differ notably in Hölder exponent distribution, it can be applied to find coding DNA sequences. The definition of α is as follows:

$$\alpha = \lim_{l \to 0} \frac{\log P_i(l)}{\log l} \tag{8}$$

in which $P_i(l) = N_i(l)/N$ refers to the proportion of designated bases (depending on the rule applied) fall into sub-covers centered at the *i-th* base position with radius of *l*/2 against the total number of bases. α =1 refers to an evenly distributed structure of base pairs, while α <1 and α >1 stand for "a densely occupied region surrounded by sparse vicinity" and "a less populated region surrounded by dense vicinity" structures respectively.

2.3 Calculations of Multifractal Spectrum for DNA sequences

In addition to the Hölder exponent, our analysis on DNA sequence also included multifractal spectrum analysis. Calculation of multifractal spectrum $f(\alpha)$ comprises several steps. First, the DNA sequence is considered point distribution by different rules. These points are then covered by boxes of size *l*. If the proportion of designated bases (depending on the rule applied) that fall into the *i-th* box is $P_i(l)$, then $\mu_i(q,l)$ can be calculate through equation (5). From equation (6), (7), the equations for $f(q)$ and $\alpha(q)$, values of $\sum_i \mu_i(q,l) \log \mu_i(q,l)$ and $\sum_i \mu_i(q,l) \log P_i(l)$

when covered with boxes of size l can be calculated first. Further, the value of l is changed to find out each corresponding value of $\sum_i \mu_i(q,l)\log\mu_i(q,l)$ and $\sum_i \mu_i(q,l)\log P_i(l)$. The results are respectively illustrated in the diagrams for $\sum_i \mu_i(q,l)\log\mu_i(q,l) - \log l$ and $\sum_i \mu_i(q,l)\log P_i(l) - \log l$. Areas of scaling region in the diagram are found and then fitted by the least square method to calculate the slope, which is the value of $f(q)$ and $\alpha(q)$. In accordance with various q, the values of $f(q)$ and $\alpha(q)$ are respectively calculated and then plotted into the coordinate system. The curve is the $f(\alpha)$ spectra.

3. GenBank Analysis and Discussion

3.1 Local Scaling Analyses

Detailed spatial organization of nucleotide sequences can be analyzed by inspecting the distribution of the Hölder exponent α. The Hölder exponent compares the invariant scaling nature of the population density of 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 bases along the DNA strand.

Various rules for forming point distribution are possible; for example, a point distribution of each different base A, T, C, G can be formed separately (the single base rule); a point distribution can contain A and G only (the purine–pyrimidine rule), and a point distribution can contain G and C only (the hydrogen bond rule), etc. [30–32]. The purine–pyrimidine rule is related to the strand chemical bias. Further, since A can hydrogen bond specifically only with T and G can bond specifically only with C, the hydrogen bond rule is related to the strand separation energy balance.

Adopting the purine-pyrimidine rule means that, reading from the beginning of a DNA sequence and down along the strand, each base position encountered is filled by either a black point

for a purine (A or G) or a white point for a pyrimidine (C or T). The resulting purine bases are treated as a distribution of a set of points in a one-dimensional line. With each point (black or white), various range l applied, count the black point number $N(l)$ in the range considered. Dividing $N(l)$ by the total number of black points N reveals the proportion value under different l, $P(l)$. The results are then illustrated onto $\log l - \log P(l)$, where the slope obtained from least square fitting the data points can be regarded as the Hölder exponent of this base location. The error in α is estimated from the standard deviation of fitted data from the linear slope. A similar approach can be applied to each base location to produce a Hölder exponent curve. When the base selected is too close to either end and the range l is too large, the sub-cover will exceed the DNA sequence, causing an edge effect. Hence, this study does not analyze α from the very beginning of the sequence, nor near the end of the sequence. Instead, it analyzes DNA sequences within the 1/10 to 9/10 sections of the sequence.

The myosin heavy chain gene family represents one of the few gene families whose complete sequences are well documented in the GenBank for a phylogenetically diverse group of organisms. This documentation provides a good opportunity to look into the fractal property changes of their components' spatial organization with evolution. Thus, the type II myosin heavy chain gene family was selected as the subject of this research and was first investigated by examining the Hölder exponent. In general, the purine–pyrimidine rule provides the most robust results, probably due to it reflects chemical structure similarities and preserves the most common point mutations, from purine to purine and from pyrimidine to pyrimidine, in the original genes [33]. Hence, listing the results of purine–pyrimidine rule first, as in other rules, is also feasible.

Figure 1 shows a typical log-log plot of purine base populations $P_i(l)$ vs. sizes of boxes and fitting lines centered at the base position $i = 12458$ of the human cardiac–myosin heavy chain gene. The smallest box has a width of just a few base pairs, while the largest box can extend to a

length of a few thousand base pairs. To avoid an edge effect, the largest box size is limited to 1/5 of the total base pairs in the chain, in this case, about 5600 bps. As Figure 1 indicates, most points fall on the fitting line very well. The slope of the line is 1.2987, which means the Hölder exponent α is 1.2987.

Figure 2 shows the variation of α along the entire strand of the human myosin heavy chain gene as calculated by the purine–pyrimidine rule for DNA sequence in the human myosin heavy chain gene. An irregular fluctuation of the curve is apparent, suggesting non-uniformity in base distributions. The α value calculated by the purine–pyrimidine rule fell between 0.8 and 1.5, similar to the range in most other rules. The diagram shows that, to avoid the aforementioned edge effect, the curve is not calculated from the first to the last base. Overlaying the known positions of exon segments (extracted from the GenBank) on the α curve calculated from the human myosin heavy chain gene reveals the surprising feature that most exons appear to be at the regions where the Hölder exponent α is less than one (see Figure 2, border lines in Figure 2 represent exons while thinner lines represent the Hölder exponent oscillation curve). If a "matching ratio" is the percentage of exons that falls in the region with $\alpha < 1$, then computation reveals that the matching ratio reached $(86.8\ ^{+3.2}_{-4.1})\%$. The $\pm$ errors are estimated from uncertainties in the α values. Uncertainties are inevitable for least square fitting similar to that in Figure 1, which then lead to an upward or downward shifting of the α curve and hence change the matching ratio.

The biological explanation of this phenomenon is currently not well established. However, the present results lend support to the findings of [34–36], which analyze the cluster-size distributions in coding and non-coding DNA sequences. Notice that in Figure 2, large peaks in the α curve are normally found between exon segments, indicating the existence of large clusters (either pyrimidine or purine) in the non-coding regions. This is consistent with the claims made in [34–36] 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. Moreover, Raghavan et al. [37] also observed that polypurines preferentially occur in genome coding regions, whereas such a bias does not occur in non-coding regions. Most exon segments observed would therefore be located in sections where the Hölder exponent curve of the purine distribution is less than one.

The results of myosin heavy chain gene family including Baker's Yeast, *Caenorhabditis elegans*, *Brugia Malayi*, fruit fly, chicken, rat, and human are summarized in Table 1. The fourth column of Table 1 shows that the number of exon segments in the myosin heavy chain gene family increases with the evolutionary order. The fifth column in Table 1 indicates that, except for the fruit fly with a total exon length of 8024 base pairs, most species have a length of around 6000 base pairs. The total exon length does not vary drastically with evolution, yet the total DNA sequence length increases. Therefore, the proportion of exon in a DNA sequence decreases from 100% to about 21%. Further, the seventh column in Table 1 shows that the matching ratio tends to increase (from 50% in yeast to 87% in human genes), indicating that exon segments of higher species are more likely to coincide with sections where A and G bases cluster. To assess whether such a trend could result from a possible α value bias (e.g., there are probably more places with $\alpha<1$ in the genetic sequence of higher species), the sixth column of Table 1 lists the percentage of length in the sequence with $\alpha<1$. All species have roughly the same percentage value; the maximum difference is only about 4%. Yet the matching ratio is 50% in yeast and 87% in human. Obviously, this result is not fortuitous. The eighth column in Table 1 indicates that, in human myosin heavy chain genes, a total of 39 exon segments fall within the range analyzed (number in parentheses). This total contains 33 exon segments with a 70% length that fall within the sections where $\alpha<1$. The proportion is about 84.6%, which also increases with the evolutionary order.

In addition to the myosin heavy chain gene family, this study analyzes another 335 DNA sequences, including 74 *Caenorhabditis elegans* genes, 36 bird genes, 45 rat genes, and 180 human genes. Including the myosin heavy chain gene, the total number of DNA sequences analyzed is 342.

Table 2 summarizes the analysis results. The seventh row in Table 2 indicates that the average matching rates for *Caenorhabditis elegans*, birds, rats, and human are 59.5%, 68.2%, 66.2% and 70.3%, respectively, as calculated from the purine base distribution (purine-pyrimidine rule). The eighth row in Table 2 shows that the proportion of exon segments with 70% length that fall within sections of the curve where α <1 to total exon segments in the analyzed sections are 37.1%, 55.6%, 47.5% and 56.8%, respectively. Figure 3 shows calculation results for the human gene DNA sequences, according to purine distribution.

Sometimes consideration of the Hölder exponent curve calculated by purine distribution only cannot produce the best results. Therefore, it is necessary to search other common characteristics of protein-coding segments to increase the accuracy. Some studies suggest that coding sequences in genes tend to contain more G and C bases than non-coding sequences, and that is particularly obvious in the DNA sequences of warm-blooded vertebrates [38–41]. This means that the DNA sequence coding segments could also fall within sections where the Hölder exponent of the G and C base distribution is less than one. For this reason, this study also calculates the Hölder exponent curves for G and C base distribution for all DNA sequences. Figure 4 shows some of these results. The ninth to eleventh columns of Table 1 summarize in detail the data for the myosin heavy chain gene family, and the ninth to tenth rows of Table 2 summarize all 339 DNA sequences. The exon segments fit well with the Hölder exponent curve of G and C base distribution where the Hölder exponent is less than one.

Because different species or different genes differ dramatically in base arrangement, it is sometimes not enough to consider the Hölder exponent curve based on only one rule. Integrating different rules may be a good idea [42]. Therefore, two rules mentioned previously can perhaps be combined to obtain more accurate results. Figure 5 compares exon segments, Hölder exponent curves for AG base distribution, and GC base distribution of several human DNA sequences. This figure indicates that some predictions on the exon segments of the DNA sequence using sections

where the Hölder exponent curve of the purine distribution is less than one are not very accurate. Yet these exon segments can fit well with sections where the G and C base distribution's Hölder exponent curve is less than one. The fifth column in Table 2 shows that the 181 human genes have a total of 1171 exon segments. Of this, total 665 segments (56.8%) with 70% length fall within the sections where the purine distribution's Hölder exponent curve is less than one and 794 segments (67.8%) with 70% length fall within the sections where the G and C base distribution's Hölder exponent curve is less than one. However, integrating the Hölder exponent curves for AG bases and GC bases shows that 1022 exon segments (87.3%) with 70% length fall within sections where the GC base or AG base distribution's Hölder exponent curve is less than one. Human myosin heavy chain genes have 36 exon segments with 70% length that fall within sections where the GC base or AG base distribution's Hölder exponent curve is less than one (see the twelfth column in Table 1). This is more than the original 33 segments, which comes from considering AG base distribution only (see the eighth column in Table 1).

As mention above, the coding sequences tend to appear in regions of the DNA sequence with abundant AG or GC bases. In other words, the coding sequences tend to exist in DNA sequence regions with a scarcity of the T base. Therefore, exon segments might be found within sections where the T base distribution's Hölder exponent curve is greater than 1. Figure 6 compares the T base distribution's Hölder exponent curve and exon segments in several human gene DNA sequences. This figure shows that the exon segments of some human genes fit well with the $\alpha > 1$ sites of the T base's Hölder exponent curve.

Many methods of bioinformatics about predicting genes or genetic sequences that can be coded into proteins have recently been proposed. Despite the diversity of these approaches, none of the above methods is perfect. In order to achieve a better prediction, it is sometimes necessary to integrate results from various programs. The GRAIL algorithm [43], for example, collects various methods of analysis and determines regions for coding DNA sequences as its "sensors" before

applying a neural network to predict DNA coding segments: one sensor, among many others, is a fractal dimension. The approach to find segments to be coded into proteins using the Hölder exponent curve proposed in this study has the advantages of simplicity, less calculation and do not need specify any window size. Perhaps in the future, this method can be integrated into GRAIL's sensor, as the fractal dimension, to help other methods of bioinformatics.

3.2 Multifractal Spectrum Analyses

Previous studies use two multifractal methods to study the structure of DNA sequences, including generalized dimensions [44–49] and the multifractal spectrum [50–51]. This study focuses on the multifractal spectrum behavior of DNA sequences only. Most researchers first transfer DNA sequences in DNA walk before calculating the multifractal spectrum, but this study does adopt this approach because it involves artificial correlation. Instead, the DNA sequence is viewed directly as the set of the four letters of the alphabet (A, G, T and C). Further, different rules were applied to calculate the probability that a certain kind of base falls into the box, and multifractal spectra are then calculated using formulas mentioned before.

3.2.1 Application to Myosin Heavy Chain Gene

Figure 7 illustrates the multifractal spectra ($f(\alpha)$ spectra) of the human cardiac β–myosin heavy chain gene, calculated by different rules. In this diagram, A, T, G, and C stand for various single base rules, respectively, while G+C represents the hydrogen bond rule and A+G represents the purine-pyrimidine rule. All the $f(\alpha)$ curves have an inverted and downward-opening parabolic shape; that is, the DNA sequence features a multifractal structure, but the opening of the $f(\alpha)$ curve varies with different rules. For example, the α value of the purine distribution (A+G) is between approximately 0.8 and 1.4.

To confirm this heterogeneity in base distribution, the positions of these bases are scrambled by a random scheme (base pair numbers, A, G, T, and C, remain the same as those in the original

sequence). Figure 8 plots the resulting $f(\alpha)$ spectra of the purine distribution and the result of original DNA sequences in the same figure. The crosses on the symbols in the diagram stand for uncertainties in the values of α and $f(\alpha)$ arising from the least square fitting procedure in Equation (6) and (7). A wider opening of the parabola in the diagram indicates that purine bases of original DNA sequences before scrambling are not uniformly distributed along the human myosin heavy chain gene; rather, they tend to form clusters of different sizes. After scrambling, a much smaller opening curve occurs, indicating that the base sequence has a more uniform distribution. This implies that the DNA sequence base arrangement in the human myosin heavy chain gene is far more complicated than any random sequence and carries considerable messages. 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 the myosin heavy chain gene) normally observed in genomes.

All introns are deleted, and the remaining exon segments of the human's myosin heavy chain gene are stitched together to form a shorter sequence containing protein-coding regions only. Figure 9 gives the corresponding $f(\alpha)$ spectra of purine distribution. Compared with the spectra of original DNA sequence in Figure 8, a much narrower $f(\alpha)$ is observed. Scrambling this shorter sequence produces little difference in $f(\alpha)$ (see Figure 9), 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 the random-walk model, which state that long-range correlation is associated with intron parts of DNA sequence.

Next, this study calculates the $f(\alpha)$ spectra of seven species' DNA sequences in the myosin heavy chain gene family, ranging from yeast to human (*homo sapiens*). Figure 10 shows the result of purine distribution in DNA sequence. Comparing the $f(\alpha)$ spectra obtained from the seven species in Figure 10, the gradual opening of the $f(\alpha)$ curve indicates that the purine distribution

of DNA sequences of higher species is more uneven, irregular, and has more local clusters. These variations suggest an increasing complexity in the structures of DNA sequences. Again, the degree of complexity follows the evolutionary order from fungus, invertebrates, vertebrates, to human. This result agrees with Table 1, which reveals that higher species have more fragmented coding regions in their DNA sequences.

The multifractal spectra of DNA sequences obtained by deleting all introns and stitching together the remaining exon segments for the corresponding seven species are shown in Figure 11. Comparing Figure 11 and Figure 10, the opening of $f(\alpha)$ curves obtained from stitching together the remaining exon segments are much smaller than original DNA sequences for most species, except yeast. This means that protein-coding sequences are much more evenly distributed than the original sequence. In Figure 11, the difference in the openings of all $f(\alpha)$ curves obtained from the seven species is insignificant; suggesting the complexity of protein-coding sequences in DNA is almost the same among different species.

3.2.2 Application to Mitochondria Genome

The primary function of mitochondria is to convert oxygen and nutrients into adenosine triphosphate (ATP), so they can be considered the power generators of cells. Because many species have this organelle, the DNA sequence of the mitochondrion can also be used to study evolution using the multifractal spectrum. Figure 12 shows the multifractal spectrum for the purine distribution of the human mitochondrion DNA sequence. The multifractal spectrum obtained from the same sequence but with bases positions scrambled by using a random scheme is also provided for comparison, from which it can be inferred that the disturbance does not have a considerable influence on the $f(\alpha)$ curve opening. This phenomenon may be related to the fact that human mitochondrion DNA is extremely compact and contains no introns. A comparison of multifractal spectra for the mitochondrion DNA sequences of different species, shown in Figure 13, shows that

the openings of these $f(\alpha)$ curves are not wide, ranging from only 0.9 to 1.1. Considering the insignificant difference in opening size and error margins, it is clear that the order of opening size cannot be determined; i.e., degrees of complexity in the structure of these DNA sequences are about the same. In fact, except for yeast, almost all the mitochondrion DNA sequences are of the same length and, quite different from the myosin heavy chain gene family, the exon segments of higher species with inheritance messages do not scatter more widely along the DNA sequence. This may account for the insignificant opening size difference in the $f(\alpha)$ curves. Mitochondrion DNA is derived from bacteria that were engulfed by early precursors of eukaryotic cells. In mammals and for most, but not all, organisms, mitochondrion DNA is inherited from the mother. Unlike nuclear DNA in which the genes are rearranged each generation due to the process called recombination, there is usually no change in mitochondrion DNA from parent to offspring by this mechanism. Hence, the mitochondria genomes in animal cells show extensive homology in organization. This may account for the insignificant difference in opening size of the $f(\alpha)$ curves for different species.

4. Conclusion

This study employed a multifractal formalism to investigate the fractal nature of DNA sequences. Phylogenetic study of the spatial organization of nucleotide sequence collected and organized from GenBank was performed using a local scaling and multifractal spectrum analysis technique. As for the multifractal spectrum $f(\alpha)$ of the myosin heavy chain gene family, the familiar, inverted, downward-opening parabola curve shape is present. The finding demonstrates the arrangement of bases in DNA sequence does exhibit a multifractal feature. The gradual opening of the $f(\alpha)$ curves associated with different species suggests an increasing complexity in the structures of DNA sequences with evolutionary order. The widest opening $f(\alpha)$ of the parabola indicates that purine bases are not uniformly distributed along the human myosin heavy chain gene; rather, they tend to form clusters of different sizes. Therefore, it may be possible to determine the evolutionary level of an unknown genetic DNA sequence by considering the multifractal spectrum opening. Unlike myosin heavy chain gene family, the openings of the $f(\alpha)$ curves for mitochondria DNA show insignificant difference among different species. This indicates that the coding segments along the mitochondria DNA sequences of higher species do not disperse widely; and different species share similar structural complexity in the mitochondria DNA sequences.

The local scaling analysis of the myosin heavy chain gene family suggests that higher species have more fragmented exon segments that take up a smaller portion of the whole sequence, and more exon segments fall within sections where the purine distribution's Hölder exponent α is less than one; this is related to the phenomenon that base (purine or pyrimidine) distribution in the DNA sequence of higher species tends to be uneven, irregular, and dense in local regions. This result also agrees with the observed increase in the width of the $f(\alpha)$ curve openings for higher-level species. Furthermore, analyzing 339 DNA sequences obtained from *Caenorhabditis elegans*, birds,

rats, and humans shows that there is also an apparent tendency of increasing correlation between the exon locations in the sequence and sections where the purine distribution's Hölder exponent α is less than one with phylogenetic order. On the other hand, there are cases where coding sequences in genes tend to contain more G and C bases than non-coding sequences, and exon segments may also fit well with the G and C base distribution's Hölder exponent curve where the Hölder exponent is less than one. Therefore, considering the Hölder exponent curve for either AG or GC bases distribution only may not achieve the best prediction for coding segments. Simultaneously considering Hölder exponent curves calculated by adopting different rules, such as the purine–pyrimidine rule and hydrogen bond rule, may have a complementary effect and can lead to a more accurate prediction. The phenomenon that exon segments of higher species fall within sections where the purine distribution's Hölder exponent curve is less than one may be due to clustering pyrimidine in DNA sequence intron in the process of evolution. While possible reasons for exon segments of higher species fall within sections where the Hölder exponent curve of the GC base distribution is less than one (i.e. GC rich) include adaptation to elevated temperatures, active gene transcription, mutation bias, or biased gene conversion (BGC) [41]. The present analysis substantiates the feasibility of finding exon segments in DNA sequences using the local scaling exponent. Because the method is simple, with less calculation work and dispenses with specifying window sizes. The authors hope that other more complicated bioinformatics methods for analyzing DNA sequences can benefit from this study in the future.

References

[1] Gates, M. A. J. Theor. Biol., 1986; 119:319-328.

[2] Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., Stanley, H. E. Nature, 1992; 356:168-170.

[3] Mandelbrot, B. B. The fractal geometry of nature. New York: Freeman, 1983.

[4] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F. Science, 1985; 228:953-958.

[5] Churchill, G. A. Bull. Math. Biol., 1989; 51:79-94.

[6] Fickett, J. W., Torney, D. C., Wolf, D. R. Genomics, 1992; 13:1056-1064.

[7] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., Goldberger, A. L. Phys. Rev., E, 1994; 49:1685-1689.

[8] Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Stanley, H. E., Stanley, M. H. R., Simons, M. Biophys. J., 1993; 65: 2673-2679.

[9] Ossadnik, S. M., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Peng, C.-K., Simons, M., Stanley, H. E. Biophys. J., 1994; 67:64-70.

[10] Havlin, S., Buldyrev, S. V., Goldberger, A. L., Mantegna, R. N., Ossadnik, S. M., Peng, C.-K., Simons, M., Stanley, H. E. Chaos, Solitons & Fractals, 1995; 6:171-201.

[11] Voss, R. F. Phys. Rev. Lett., 1992; 68:3805-3808.

[12] Bickel, D. R. Chaos, Solitons & Fractals, 2002; 13:491-497.

[13] Zhang, L., Sun T. Chaos, Solitons & Fractals, 2005; 23:1077-1085.

[14] Kim, J. K., Yang, S. I., Kwon, Y. H., Lee, E. I. Chaos, Solitons & Fractals, 2005; 23:1795-1807.

[15] Yu, Z.-G., Wang, B., Chaos, Solitons & Fractals, 2001; 12:519-526.

[16] Yu, Z.-G., Anh, V. Chaos, Solitons & Fractals, 2001; 12:1827-1834.

[17] Hao, B.-L., Physica, A, 2000, 282:225-246.

[18] Hao, B.-L., Lee, H. C., Zhang, S.-Y. Chaos, Solitons & Fractals, 2000, 11:825-836.

[19] Yu, Z.-G., Hao, B.-L., Xie, H.-M., Chen, G.-Y. Chaos, Solitons & Fractals, 2000, 11:2215-2222.

[20] Zhang, L., Jiang, Z. Chaos, Solitons & Fractals, 2004, 22:947-955.

[21] Xiao, Y., Huang, Y. Chaos, Solitons & Fractals, 2004, 20:497-502.

[22] Sun, T., Zhang, L., Chen, J., Jiang, Z. Chaos, Solitons & Fractals, 2004, 20:1075-1084.

[23] Cheng, J., Zhang, L. Chaos, Solitons & Fractals, 2005, 25:339-346.

[24] Zhang, L., Chen, J. Chaos, Solitons & Fractals, 2005, 24:115-123.

[25] Halsey, T. C., Jensen, M. H., Kadanoff, L. P., Procaccia, I., Shraiman, B. I. Phys. Rev. A, 1986; 33:1141-1151.

[26] Grassberger, P. Phys. Lett. A, 1983; 97:227-230.

[27] Hentschel, H., Procaccia, I. Physica D, 1983; 8:435-444.

[28] Chhabra, A. V., Jensen, R. V. Phys. Rev. Lett., 1989; 62:1327-1330.

[29] Chhabra, A. B., Meneveau, C., Jensen, R. V., Sreenivasan, K. R. Phys. Rev. A, 1989; 40:5284-5294.

[30] Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsa, M. E., Peng, C.-K., Simons, M., Stanley, H. E. Phy. Rev. E, 1995; 51: 5084-5091.

[31] Viswanathan, G. M., Buldyrev, S. V., Havlin, S., Stanley, H. E. Biophys. J., 1997; 72: 866-875.

[32] Viswanathan, G. M., Buldyrev, S. V., Havlin, S., Stanley, H. E. Physica A, 1998; 249:581-586.

[33] Provata, A., Almirantis, Y. J. Stat. Phys., 2002; 106:23-56.

[34] Provata, A., Almirantis, Y. Physica A, 1997; 247:482-496.

[35] Almirantis, Y., Provata, A. J. Stat. Phys., 1999; 97:233-262.

[36] Provata, A. Physica A, 1999; 264:570-580.

[37] Raghavan, S., Hariharan, R., Brahmachari, S. K. Gene, 2000; 242: 275-283.

[38] Burset, M., Guigo, R. Genomics, 1996; 34:353-367.

[39] Bernardi, G., Bernardi, G. J. Mol. Evol., 1990; 31:282-293.

[40] Vinogradov, A. E. Nucleic Acids Research, 2003; 31:1838-1844.

[41] Vinogradov, A. E. Nucleic Acids Res., 2003; 31:5212-5220.

[42] Kotlar, D., Lavner, Y. Genome Research, 2003; 13: 1930-1937.

[43] Uberbacher, E. C., Mural, R. J. Proc. Natl. Acad. Sci. U. S. A., 1991; 88: 11261-11265.

[44] Glazier, J. A., Raghavachari, S., Berthelsen, C. L., Skolnick, M. H. Phys. Rev., E, 1995; 51:2665-2668.

[45] Anh, V., Lau, K.-S., Yu, Z.-G. J. Phys. A, 2001; 34:7127-7139.

[46] Anh, V., Lau, K.-S., Yu, Z.-G. Phys. Rev., E, 2002; 66:031910-1-031910-9.

[47] Bershadskii, A. Phys. Lett. A, 2001; 284:136-140.

[48] Yu, Z.-G., Anh, V., Lau, K.-S. Physica A, 2001; 301:351-361.

[49] Yu, Z.-G., Anh, V., Lau, K.-S. Phys. Rev., E, 2001; 64:031903-1-031903-9.

[50] Gutierrez, J. M., Rodriguez, M. A., Abramson, G. Physica A, 2001; 300:271-284.

[51] Oiwa, N. N., Goldman, C. Phys. Rev. Lett., 2000; 85:2396-2399.

List of Figures

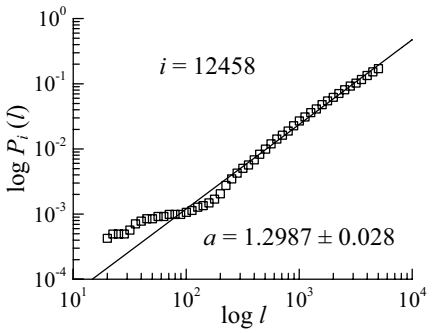


Figure 1

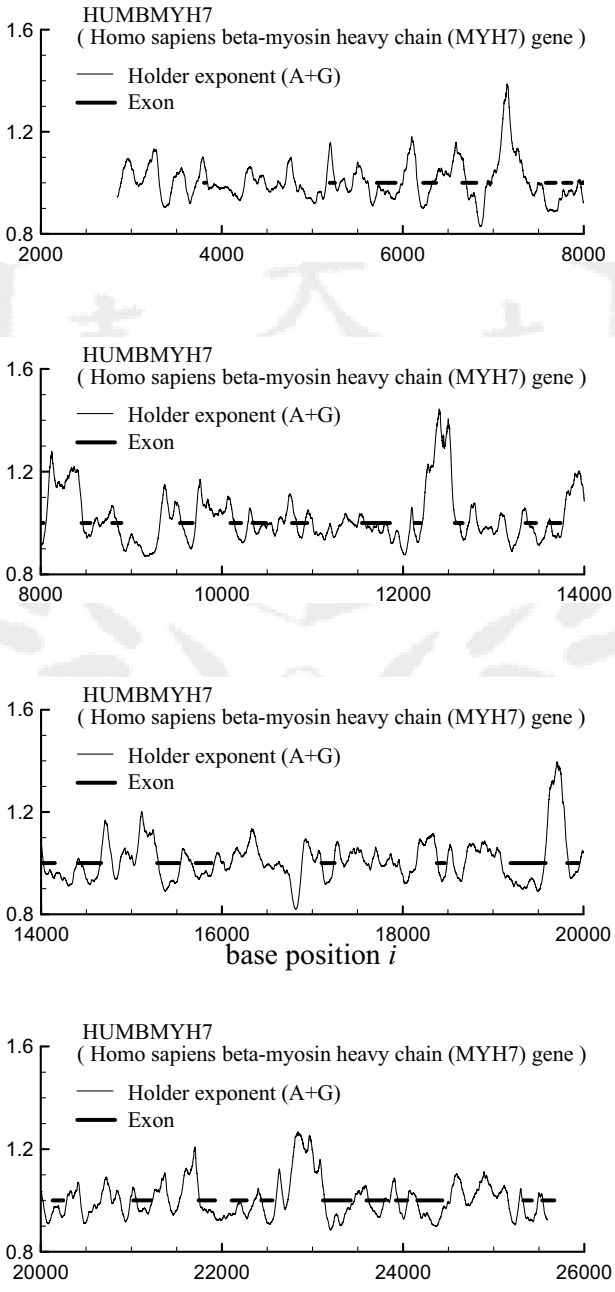
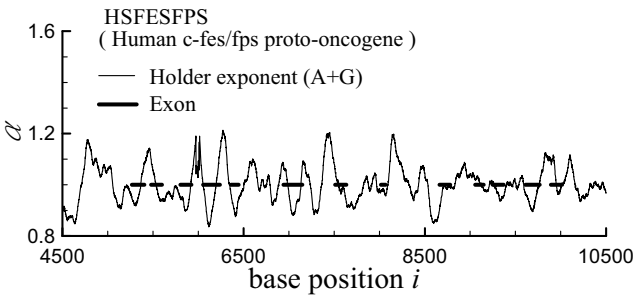
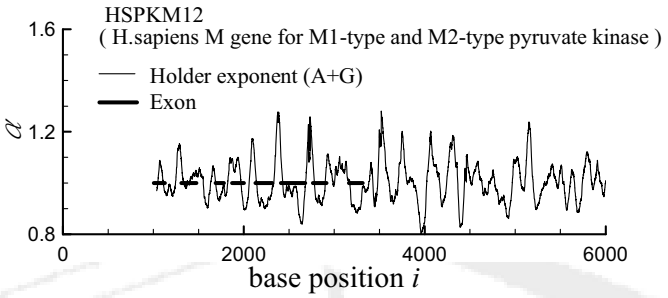


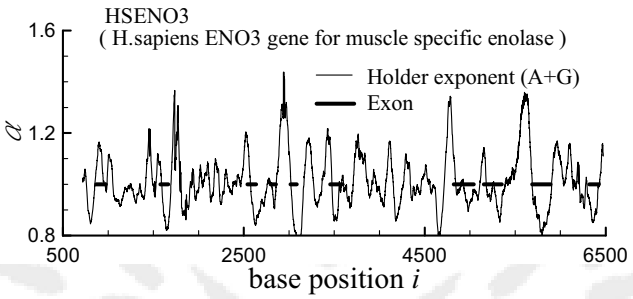
Figure 2


HSFESFPS
(Human c-fes/fps proto-oncogene)
Holder exponent (A+G)
Exon
α
1.6
1.2
0.8
4500
6500
8500
10500
base position *i*

(a)

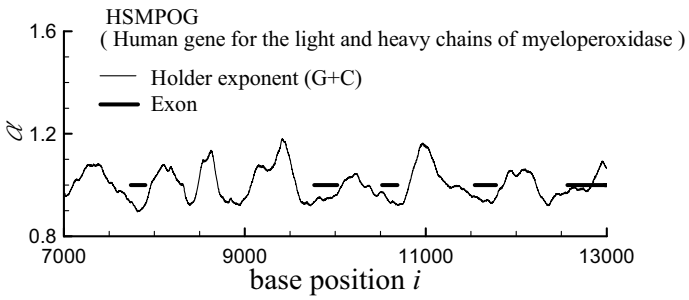

HSPKM12
(H.sapiens M gene for M1-type and M2-type pyruvate kinase)
Holder exponent (A+G)
Exon
α
1.6
1.2
0.8
0
2000
4000
6000
base position *i*

(b)

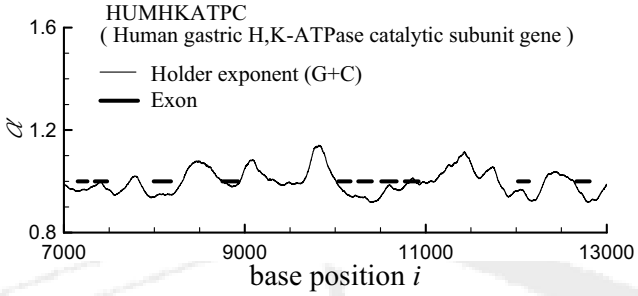

HSENO3
(H.sapiens ENO3 gene for muscle specific enolase)
Holder exponent (A+G)
Exon
α
1.6
1.2
0.8
500
2500
4500
6500
base position *i*

(c)

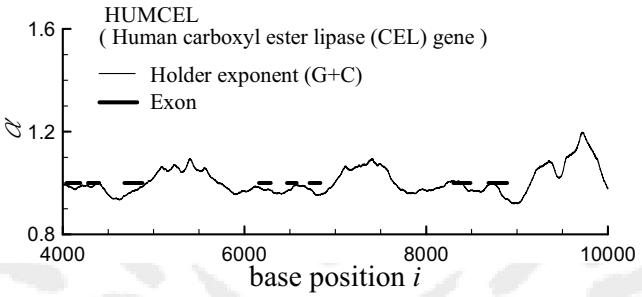
Figure 3


HSMPOG
(Human gene for the light and heavy chains of myeloperoxidase)
Holder exponent (G+C)
Exon
α
1.6
1.2
0.8
7000
9000
11000
13000
base position i

(a)

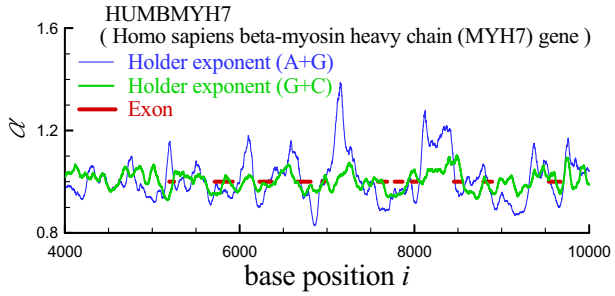

HUMHKATPC
(Human gastric H,K-ATPase catalytic subunit gene)
Holder exponent (G+C)
Exon
α
1.6
1.2
0.8
7000
9000
11000
13000
base position i

(b)

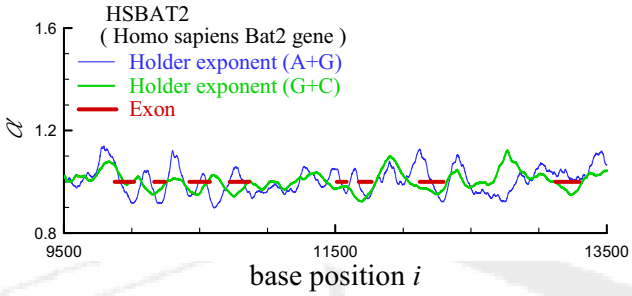

HUMCEL
(Human carboxyl ester lipase (CEL) gene)
Holder exponent (G+C)
Exon
α
1.6
1.2
0.8
4000
6000
8000
10000
base position i

(c)

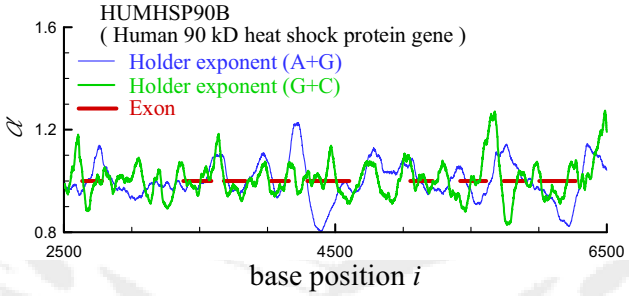
Figure 4



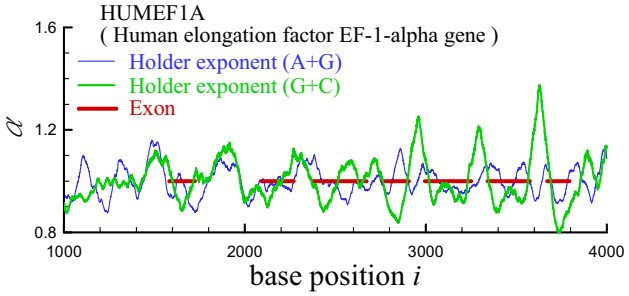
(a)



(b)

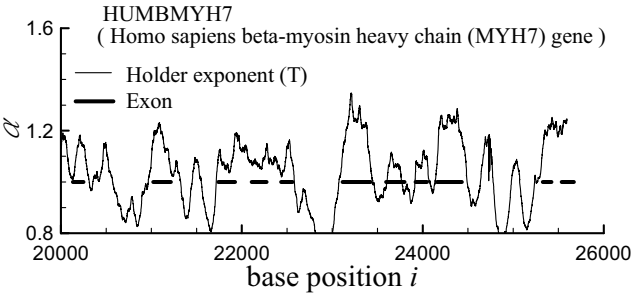


(c)

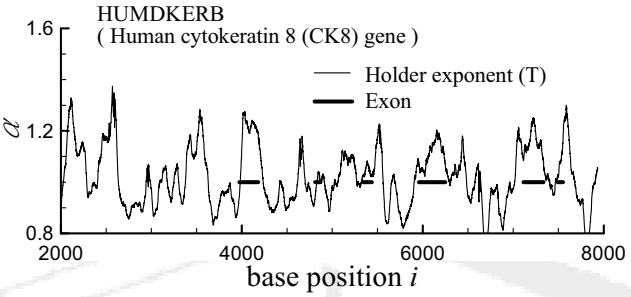


(d)

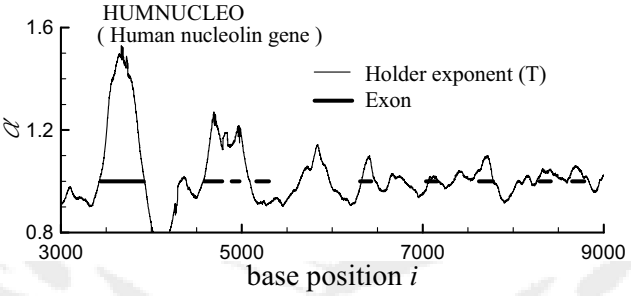
Figure 5


HUMBMYH7
(Homo sapiens beta-myosin heavy chain (MYH7) gene)
Holder exponent (T)
Exon
α
1.6
1.2
0.8
20000
22000
24000
26000
base position i

(a)


HUMDKERB
(Human cytokeratin 8 (CK8) gene)
Holder exponent (T)
Exon
α
1.6
1.2
0.8
2000
4000
6000
8000
base position i

(b)


HUMNUCLEO
(Human nucleolin gene)
Holder exponent (T)
Exon
α
1.6
1.2
0.8
3000
5000
7000
9000
base position i

(c)

Figure 6

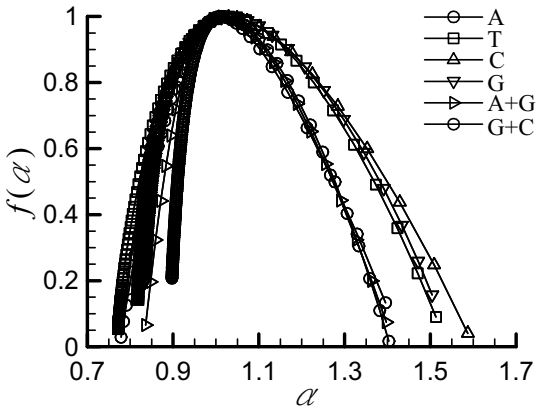


Figure 7

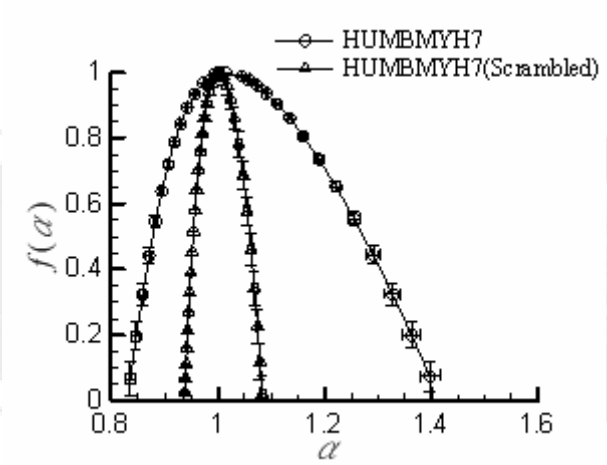


Figure 8

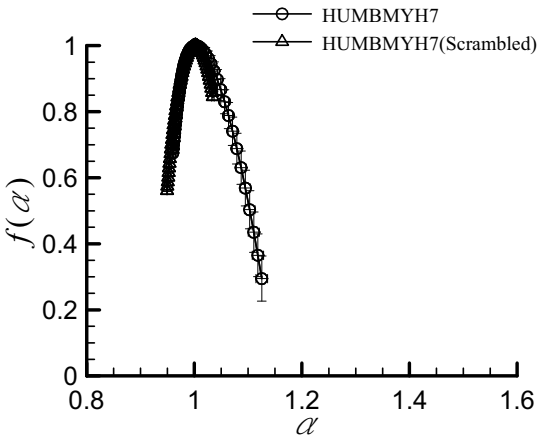


Figure 9

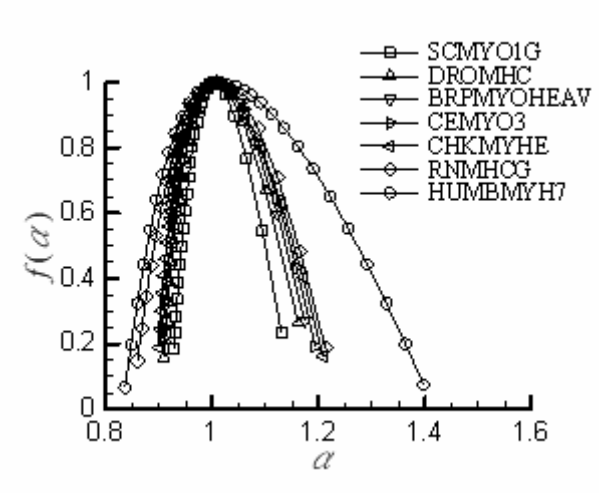


Figure 10

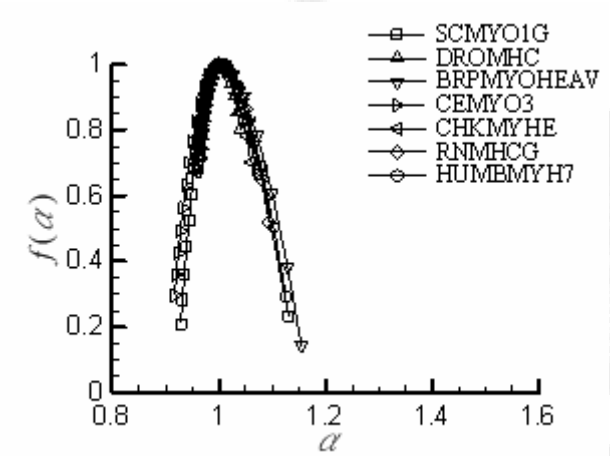


Figure 11

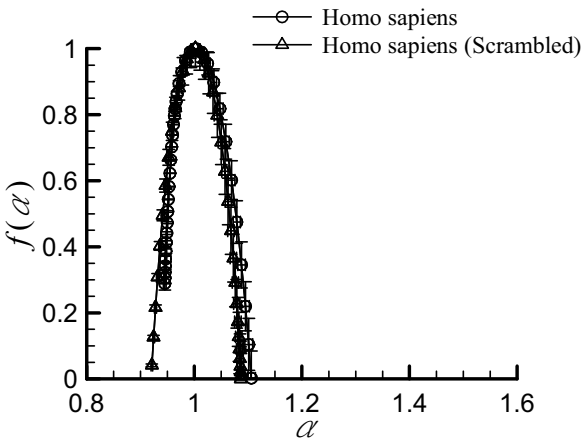


Figure 12

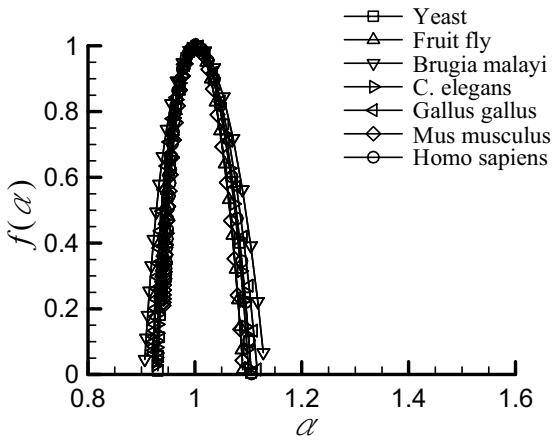

Yeast
Fruit fly
Brugia malayi
C. elegans
Gallus gallus
Mus musculus
Homo sapiens
f(α)
α
0
0.2
0.4
0.6
0.8
1
0.8
1
1.2
1.4
1.6

Figure 13

Table 1 Summary of local scaling analysis on myosin heavy chain genes family.

Family Organism	GenBank Accession # (locus)	Length Analyzed (bp) (total length)	# of exon segments	Total exon length (bp) (% exon length)	%α < 1 in analysis (A+G)	%α < 1 in exon (A+G)	# of exon segments whose %α < 1 are greater than 70 (# of exon segments in analyzed region) (A+G)	%α < 1 in analysis (G+C)	%α < 1 in exon (G+C)	# of exon segments whose %α < 1 are greater than 70 (# of exon segments in analyzed region) (G+C)	# of exon segments whose %α < 1 are greater than 70 (# of exon segments in analyzed region) (A+G or G+C)
Saccharomyces cerevisiae (yeast)	X53947 (SCMYO1G)	4889 (6108)	1	6108 (100)	50.0	50.0+5.8 -5.7	0 (1)	49.1	49.1+6.4 -4.8	0 (1)	0 (1)
Caenorhabditis elegans #3 (worm)	X08067 (CEMYO3)	9285 (11604)	7	5911 (50.9)	51.6	53.0+6.7 -7.8	4 (7)	49.1	54.6+6.8 -6.6	5 (7)	5 (7)
Brugia malayi (worm)	M74000 (BRPMYOHEA)	9415 (11766)	13	5598 (47.6)	52.3	64.8+5.6 -6.1	6 (11)	51.7	69.3+5.1 -5.4	6 (11)	7 (11)
Drosophila melanogaster (fruit fly)	M61229 (DROMHC)	18132 (22663)	30	8024 (35.4)	52.2	67.7+5.6 -6.2	12 (27)	52.2	85.8+2.6 -3.7	21 (27)	23 (27)
Gallus gallus (chicken)	J02714 (CHKMYHE)	24890 (31111)	38	5823 (18.7)	50.9	75.8+4.5 -5.2	28 (38)	49.2	90.8+2.1 -2.9	31 (38)	36 (38)
Rattus norvegicus (rat)	X04267 (RNMHCG)	20606 (25755)	41	6015 (23.4)	50.1	72.8+2.6 -4.3	21 (34)	54.4	80.8+3.8 -5.1	24 (34)	31 (34)
Homo sapiens (human)	M57965 (HUMBMYH7)	22752 (28438)	40	6008 (21.1)	54.3	86.8+3.2 -4.1	33 (39)	55.2	79.4+4.3 -4.9	25 (39)	36 (39)

Table 2 Summary of local scaling analysis on 339 DNA sequences of different species collected from GenBank

Species	*Caenorhabditis elegans*	Bird	Rat	Human
Number of sequences (including myosin heavy chain gene family)	75	37	46	181
Total length of sequences	465459	306376	267478	1946379
Total length analyzed	372514	245171	214068	1557435
Total number of exon segments	533	275	278	1340
Exon segments in the analyzed sections	501	257	261	1171
Average matching rate (A+G)	59.5%	68.2%	66.2%	70.3%
Number and percentage of segments with the matching rate over 70% (A+G)	186(37.1%)	143(55.6%)	124(47.5%)	665(56.8%)
Average matching rate (G+C)	71.8%	72.0%	68.7%	75.9%
Number and percentage of segments with the matching rate over 70% (G+C)	323(64.5%)	157(61.1%)	142(54.4%)	794(67.8%)
Number and percentage of segments with the matching rate over 70% (A+G or G+C)	361(72.1%)	204(79.4%)	204(78.2%)	1022(87.3%)

Appendix

Available online at www.sciencedirect.com
ScienceDirect
Chaos, Solitons and Fractals xxx (2007) xxx–xxx
CHAOS SOLITONS & FRACTALS
www.elsevier.com/locate/chaos

Local scaling and multifractal spectrum analyses of DNA sequences – GenBank data analysis

Zhi-Yuan Su [a,*], Tzuyin Wu [b], Shu-Yin Wang [c]

[a] Department of Information Management, Chia Nan University of Pharmacy and Science, Tainan 717, Taiwan, ROC
[b] Department of Mechanical Engineering, National Taiwan University, Taipei 106, Taiwan, ROC
[c] Department of Animal Science, Chinese Culture University, Taipei 111, Taiwan, ROC

Accepted 17 September 2007

Abstract

Base sequences of deoxyribonucleic acid (DNA) in an organism carry all the instructions regarding its growth and development. On the surface, such sequences seem irregular; yet in reality, they are symbolic sequences with an organized structure. This study investigates the characteristics of base arrangement and distribution in DNA sequences from the fractal theory viewpoint. In addition to multifractal features demonstrated by the DNA sequence, this study also compares the multifractal spectra derived from a particular family of gene among several different species. The results reveal that a considerable correlation exists between base distribution and evolutionary order. Furthermore, local scaling exponent (Hölder exponent) differences between coding segments (exon) and non-coding segments (intron) are also examined. It is suggested that such differences in the local distribution of bases can be applied to find coding segments within the DNA sequence that is to be translated into protein. This local scaling analysis is feasible and has the potential to become an effective tool for rapid location of possible coding sites in DNA sequences. The authors hope that future studies using more complicated bioinformatics methods for analyzing DNA sequences can benefit from this study.

1. Introduction

Complex DNA base sequence studies began twenty years ago. In the 1980s, Gates started to map bases in a DNA sequence onto a two-dimensional space [1]. In the 1990s, Peng et al. [2] published a paper in the *Nature* journal. It was the first time that DNA base sequences transformed into a one-dimensional irregular walk, named "DNA walk." One-dimensional DNA walk can be defined as: if pyrimidine is located in the ith position along the DNA chain, then the walker steps up ($u(i) = +1$); if purine is found, then the walker steps down ($u(i) = -1$). Calculate the sum of $u(i)$ after l steps and a curve $y(l)$ for DNA walk can be obtained. DNA walk not only provides a simple and intuitional representation of each DNA sequence but also assists in analyzing correlations among DNA base sequences through calculating root mean square fluctuation $F(l)$ of average displacement. $F(l)$ can be used to distinguish three types of

* Corresponding author. Tel.: +886 6 2664911x5305; fax: +886 6 3660607.
E-mail address: zysu@mail.chna.edu.tw (Z.-Y. Su).

0960-0779/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chaos.2007.09.078

behaviors: (1) if nucleotide sequence is random, then $F(l) \sim l^{1/2}$; (2) if local correlation can be found within a certain characteristic range (such as Markov chain) with asymptotic behavior to a purely random sequence, then, eventually $F(l) \sim l^{1/2}$ for large l; (3) if characteristic length does not exist, then the relation between fluctuation $F(l)$ and l is power-law, $F(l) \sim l^{\delta}$, with $\delta \neq 1/2$, indicating that DNA walk features self-similarity, or fractal property [3]. Peng et al. [2] analysed nucleotide sequences of genes and discovered a long-range correlation among genes abundant with intron, counting up to thousands of bases. On the contrary, for cDNA and genes with scarce intron, $\delta \approx 1/2$; in other words, long-range correlation is not observed. A major obstacle to overcome in long-range correlation analysis is that the DNA sequence's mosaic structure causes the four bases A (adenine), G (guanine), T (thymine), C (cytosine) strand biases and "trend" presence in a DNA walk [4–6]. Peng [2] proposed a min–max method to overcome this problem, which unfortunately required determination on the number of local maximum and minimum in a DNA walk curve, suffering from practical application difficulties. Therefore, Peng et al. [7] later proposed the bridge method and detrended fluctuation analysis, or DFA. In 1993, Buldyrev et al. [8] used DFA to investigate genes in the myosin heavy chain family to examine the relation of base sequence fractal complexity to evolution. The use of single gene family can avoid potential bias caused by different evolutionary pressures and various base compositions among irrelevant gene families. Their analytical results reveal that, for species from eukaryotes to invertebrates to vertebrates, the correlation coefficient features monotonic increase. After that, Ossadnik et al. [9] proposed the coding sequence finder algorithm (CSF), based on DFA. First, a certain "window size" is selected. Then the DNA sequence is divided into several regions, which are respectively measured for individual DFA exponents. The results reveal that the local value of the DFA exponent typically displays minima where genes are suspected. Reviews about DFA's applications can be referred to in the bibliography [10].

DNA walk is very intuitional, but for the nucleotide sequences, it involves introducing artificial correlations into DNA walks of less than four dimensions. To avoid introducing artificial correlation, Voss [11] proposed a simple method that separates a DNA base sequence into four sub-sequences of A, G, T, and C, each of which stands for base location in the original DNA sequence. For example, sub-sequence A is a numeric sequence obtained by replacing A with 1 and replacing the other three bases with 0 in the original DNA nucleotide sequence, resulting in a time sequence $x_A(n)$ that oscillates with position coordinate n. In this manner, time sequences for other bases can be defined as $x_G(n)$, $x_T(n)$, and $x_C(n)$. Through Fourier transform, time sequences of these bases can be illustrated into a power spectra $S_k(f)$, showing that basically they share similar tendency. In cases of larger f, $S_k(f)$ is mostly white noise. Under low-frequency, $S_k(f)$ demonstrates a power-law $S_k(f) \sim 1/f^{\beta}$. In many cases, $1/f^{\beta}$ noise is assumed independent of high-frequency white noise, so that high-frequency white noise can be deducted from $S_k(f)$, making the long-range correlation within a low-frequency area more obvious. The Voss approach is special in that the original DNA base sequence's correlation function can be regarded as the sum of correlation functions of the four sub-sequences, $x_A(n)$, $x_G(n)$, $x_T(n)$, and $x_C(n)$. This conclusion also applies to spectral density. Thus, by figuring out the spectral densities of all sub-sequences, the spectral density of original DNA sequence can be obtained without assuming any correlation among the four bases.

Besides aforementioned approaches, other DNA sequence analyses include general entropy function calculation [12], the Zipf linguistics method [13,14], Hurst exponent analysis [15,16], Hao's geometric representations [17–19], two-dimensional DNA walk [20], nonlinear prediction method [21], cluster-size distribution [22–24] and so on. However, most of research methods focus on the overall properties of DNA sequences. The power spectra, for example, depict the average distribution of energy under each oscillation frequency contained in the sequence. Therefore, local energy variations in the oscillation frequency of the sequence cannot be obtained. An average component energy distribution for oscillations presented in natural phenomena is very rare, and this is true for DNA sequences. From a geometric point of view, the sequence of a particular base in a DNA strand can be viewed as a distribution of a set of points along a line. 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" system [3,25]. This study applies that notion in the same way to the study of DNA sequences. Since the production of a polypeptide chain (protein) only depends on the linear order of bases along the DNA strand, spatial distribution patterns of bases are scrutinized using multifractal formalism.

2. Local scaling and multifractal analyses

2.1. 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 ith box. Then, in the limit $l \to 0$, an exponent (singularity strength, or Hölder exponent) α can be defined by

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

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

In general, α is not uniformly distributed and can therefore serve as the crowding index of a local cluster. If the number of boxes $N(\alpha)$ where the probability measure P_i has a singularity strength between α and $\alpha + d\alpha$ is counted, then $f(\alpha)$ can be loosely defined as the fractal dimension of the set of boxes with singularity strength α by [25]

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

This formalism thus describes a multifractal measure in terms of interwoven sets of different singularity strengths α, where each set is characterized by its own fractal dimension $f(\alpha)$. Another useful multifractal formalism is the so-called generalized dimension, defined as [26,27]

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

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

As the generalized dimension D_q is computed, the multifractal spectrum $f(\alpha)$ is usually evaluated from D_q via a Legendre transformation [25]:

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

However, as other studies indicate [28,29], the validity of the Legendre transformation relies on the smoothness of functions $f(\alpha)$ and D_q. In the attempt to obtain D_q by scaling the probability measures P_i^q with box sizes l, naturally evolving and experimentally observed data often produce a log–log plot featuring oscillations and scattering rather than perfect linear behavior, especially when the value of q is large. This then produces a D_q curve with large uncertainties. Applying the Legendre transformation to such a curve may generate false results and make error estimation in the $f - \alpha$. Previous studies propose a direct determination of $f(\alpha)$ to circumvent this pitfall. This method first involves constructing a one-parameter family of normalized measures $\mu(q)$ at each box i from probabilities $P_i(l)$

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

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

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

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

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

Eqs. (6) and (7) provide an alternative definition of the multifractal spectrum, which can be used to obtain $f(\alpha)$ directly from real-world data without using the Legendre transformation. Subsequent calculation in make use of this method.

2.2. Calculations of Hölder exponent α for DNA sequences

Featuring DNA's structure by the variation of local scaling exponent curve (or Hölder exponent α) according to the location of base in DNA chain is the idea of this study. Since exon and intron in the DNA sequences differ notably in Hölder exponent distribution, it can be applied to find coding DNA sequences. The definition of α is as follows:

$$\alpha = \lim_{l \to 0} \frac{\log P_i(l)}{\log l} \tag{8}$$

in which $P_i(l) = N_i(l)/N$ refers to the proportion of designated bases (depending on the rule applied) fall into sub-covers centered at the ith base position with radius of $l/2$ against the total number of bases. $\alpha = 1$ refers to an evenly distributed

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

structure of base pairs, while $\alpha < 1$ and $\alpha > 1$ stand for "a densely occupied region surrounded by sparse vicinity" and "a less populated region surrounded by dense vicinity" structures respectively.

2.3. Calculations of multifractal spectrum for DNA sequences

In addition to the Hölder exponent, our analysis on DNA sequence also included multifractal spectrum analysis. Calculation of multifractal spectrum $f(\alpha)$ comprises several steps. First, the DNA sequence is considered point distribution by different rules. These points are then covered by boxes of size l. If the proportion of designated bases (depending on the rule applied) that fall into the ith box is $P_i(l)$, then $\mu_i(q,l)$ can be calculate through Eq. (5). From Eqs. (6),(7), the equations for $f(q)$ and $\alpha(q)$, values of $\sum_i \mu_i(q,l)\log\mu_i(q,l)$ and $\sum_i \mu_i(q,l)\log P_i(l)$ when covered with boxes of size l can be calculated first. Further, the value of l is changed to find out each corresponding value of $\sum_i \mu_i(q,l)\log\mu_i(q,l)$ and $\sum_i \mu_i(q,l)\log P_i(l)$. The results are respectively illustrated in the diagrams for $\sum_i \mu_i(q,l)\log\mu_i(q,l) - \log l$ and $\sum_i \mu_i(q,l)\log P_i(l) - \log l$. Areas of scaling region in the diagram are found and then fitted by the least square method to calculate the slope, which is the value of $f(q)$ and $\alpha(q)$. In accordance with various q, the values of $f(q)$ and $\alpha(q)$ are respectively calculated and then plotted into the coordinate system. The curve is the $f(\alpha)$ spectra.

3. GenBank analysis and discussion

3.1. Local scaling analyses

Detailed spatial organization of nucleotide sequences can be analysed by inspecting the distribution of the Hölder exponent α. The Hölder exponent compares the invariant scaling nature of the population density of 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 bases along the DNA strand.

Various rules for forming point distribution are possible; for example, a point distribution of each different base A, T, C, G can be formed separately (the single base rule); a point distribution can contain A and G only (the purine–pyrimidine rule), and a point distribution can contain G and C only (the hydrogen bond rule), etc. [30–32]. The purine–pyrimidine rule is related to the strand chemical bias. Further, since A can hydrogen bond specifically only with T and G can bond specifically only with C, the hydrogen bond rule is related to the strand separation energy balance.

Adopting the purine–pyrimidine rule means that, reading from the beginning of a DNA sequence and down along the strand, each base position encountered is filled by either a black point for a purine (A or G) or a white point for a pyrimidine (C or T). The resulting purine bases are treated as a distribution of a set of points in a one-dimensional line. With each point (black or white), various range l applied, count the black point number $N(l)$ in the range considered. Dividing $N(l)$ by the total number of black points N reveals the proportion value under different l, $P(l)$. The results are then illustrated onto $\log l - \log P(l)$, where the slope obtained from least square fitting the data points can be regarded as the Hölder exponent of this base location. The error in α is estimated from the standard deviation of fitted data from the linear slope. A similar approach can be applied to each base location to produce a Hölder exponent curve. When the base selected is too close to either end and the range l is too large, the sub-cover will exceed the DNA sequence, causing an edge effect. Hence, this study does not analyse α from the very beginning of the sequence, nor near the end of the sequence. Instead, it analyses DNA sequences within the 1/10–9/10 sections of the sequence.

The myosin heavy chain gene family represents one of the few gene families whose complete sequences are well documented in the GenBank for a phylogenetically diverse group of organisms. This documentation provides a good opportunity to look into the fractal property changes of their components spatial organization with evolution. Thus, the type II myosin heavy chain gene family was selected as the subject of this research and was first investigated by examining the Hölder exponent. In general, the purine–pyrimidine rule provides the most robust results, probably due to it reflects chemical structure similarities and preserves the most common point mutations, from purine to purine and from pyrimidine to pyrimidine, in the original genes [33]. Hence, listing the results of purine–pyrimidine rule first, as in other rules, is also feasible.

Fig. 1 shows a typical log–log plot of purine base populations $P_i(l)$ vs. sizes of boxes and fitting lines centered at the base position $i = 12{,}458$ of the human cardiac-myosin heavy chain gene. The smallest box has a width of just a few base pairs, while the largest box can extend to a length of a few thousand base pairs. To avoid an edge effect, the largest box size is limited to 1/5 of the total base pairs in the chain, in this case, about 5600 bps. As Fig. 1 indicates, most points fall on the fitting line very well. The slope of the line is 1.2987, which means the Hölder exponent α is 1.2987.

Fig. 2 shows the variation of α along the entire strand of the human myosin heavy chain gene as calculated by the purine–pyrimidine rule for DNA sequence in the human myosin heavy chain gene. An irregular fluctuation of the curve

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

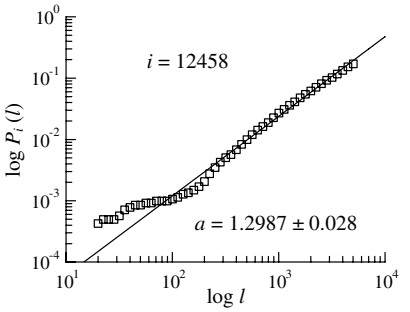


Fig. 1. Scaling of $P_i(l)$ with box size l.

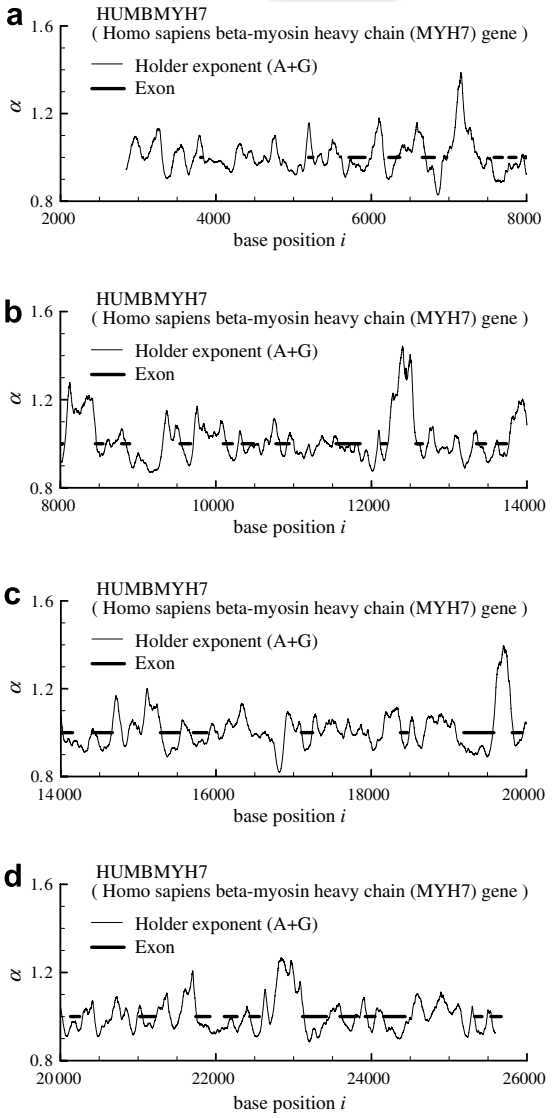


Fig. 2. Comparison in exon segments and Hölder exponent curve of purine distribution in human cardiac β-myosin heavy chain gene.

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

is apparent, suggesting non-uniformity in base distributions. The α value calculated by the purine–pyrimidine rule fell between 0.8 and 1.5, similar to the range in most other rules. The diagram shows that, to avoid the aforementioned edge effect, the curve is not calculated from the first to the last base. Overlaying the known positions of exon segments (extracted from the GenBank) on the α curve calculated from the human myosin heavy chain gene reveals the surprising feature that most exons appear to be at the regions where the Hölder exponent α is less than one (see Fig. 2, border lines in Fig. 2 represent exons while thinner lines represent the Hölder exponent oscillation curve). If a "matching ratio" is the percentage of exons that falls in the region with $\alpha < 1$, then computation reveals that the matching ratio reached $(86.8^{+3.2}_{-4.1})\%$. The $\pm$ errors are estimated from uncertainties in the α values. Uncertainties are inevitable for least square fitting similar to that in Fig. 1, which then lead to an upward or downward shifting of the α curve and hence change the matching ratio.

The biological explanation of this phenomenon is currently not well established. However, the present results lend support to the findings of [34–36], which analyse the cluster-size distributions in coding and non-coding DNA sequences. Notice that in Fig. 2, large peaks in the α curve are normally found between exon segments, indicating the existence of large clusters (either pyrimidine or purine) in the non-coding regions. This is consistent with the claims made in [34–36] 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. Moreover, Raghavan et al. [37] also observed that polypurines preferentially occur in genome coding regions, whereas such a bias does not occur in non-coding regions. Most exon segments observed would therefore be located in sections where the Hölder exponent curve of the purine distribution is less than one.

The results of myosin heavy chain gene family including Baker's Yeast, *Caenorhabditis elegans*, *Brugia malayi*, fruit fly, chicken, rat, and human are summarized in Table 1. The fourth column of Table 1 shows that the number of exon segments in the myosin heavy chain gene family increases with the evolutionary order. The fifth column in Table 1 indicates that, except for the fruit fly with a total exon length of 8024 base pairs, most species have a length of around 6000 base pairs. The total exon length does not vary drastically with evolution, yet the total DNA sequence length increases. Therefore, the proportion of exon in a DNA sequence decreases from 100% to about 21%. Further, the seventh column in Table 1 shows that the matching ratio tends to increase (from 50% in yeast to 87% in human genes), indicating that exon segments of higher species are more likely to coincide with sections where A and G bases cluster. To assess whether such a trend could result from a possible α value bias (e.g., there are probably more places with $\alpha < 1$ in the genetic sequence of higher species), the sixth column of Table 1 lists the percentage of length in the sequence with $\alpha < 1$. All species have roughly the same percentage value; the maximum difference is only about 4%. Yet the matching ratio is 50% in yeast and 87% in human. Obviously, this result is not fortuitous. The eighth column in Table 1 indicates that, in human myosin heavy chain genes, a total of 39 exon segments fall within the range analysed (number in parentheses). This total contains 33 exon segments with a 70% length that fall within the sections where $\alpha < 1$. The proportion is about 84.6%, which also increases with the evolutionary order.

In addition to the myosin heavy chain gene family, this study analyses another 335 DNA sequences, including 74 *C. elegans* genes, 36 bird genes, 45 rat genes, and 180 human genes. Including the myosin heavy chain gene, the total number of DNA sequences analysed is 342. Table 2 summarizes the analysis results. The seventh row in Table 2 indicates that the average matching rates for *C. elegans*, birds, rats, and human are 59.5%, 68.2%, 66.2% and 70.3%, respectively, as calculated from the purine base distribution (purine–pyrimidine rule). The eighth row in Table 2 shows that the proportion of exon segments with 70% length that fall within sections of the curve where $\alpha < 1$ to total exon segments in the analysed sections are 37.1%, 55.6%, 47.5% and 56.8%, respectively. Fig. 3 shows calculation results for the human gene DNA sequences, according to purine distribution.

Sometimes consideration of the Hölder exponent curve calculated by purine distribution only cannot produce the best results. Therefore, it is necessary to search other common characteristics of protein-coding segments to increase the accuracy. Some studies suggest that coding sequences in genes tend to contain more G and C bases than non-coding sequences, and that is particularly obvious in the DNA sequences of warm-blooded vertebrates [38–41]. This means that the DNA sequence coding segments could also fall within sections where the Hölder exponent of the G and C base distribution is less than one. For this reason, this study also calculates the Hölder exponent curves for G and C base distribution for all DNA sequences. Fig. 4 shows some of these results. The ninth to eleventh columns of Table 1 summarize in detail the data for the myosin heavy chain gene family, and the ninth to tenth rows of Table 2 summarize all 339 DNA sequences. The exon segments fit well with the Hölder exponent curve of G and C base distribution where the Hölder exponent is less than one.

Because different species or different genes differ dramatically in base arrangement, it is sometimes not enough to consider the Hölder exponent curve based on only one rule. Integrating different rules may be a good idea [42]. Therefore, two rules mentioned previously can perhaps be combined to obtain more accurate results. Fig. 5 compares exon segments, Hölder exponent curves for AG base distribution, and GC base distribution of several human DNA

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

Table 1
Summary of local scaling analysis on myosin heavy chain genes family

Family organism	GenBank accession # (locus)	Length analysed (bp) (total length)	# of exon segments	Total exon length (bp) (% exon length)	$\%\alpha < 1$ in analysis (A + G)	$\%\alpha < 1$ in exon (A + G)	# of exon segments whose $\%\alpha < 1$ are greater than 70 (# of exon segments in analysed region) (A + G)	$\%\alpha < 1$ in analysis (G + C)	$\%\alpha < 1$ in exon (G + C)	# of exon segments whose $\%\alpha < 1$ are greater than 70 (# of exon segments in analysed region) (G + C)	# of exon segments whose $\%\alpha < 1$ are greater than 70 (# of exon segments in analysed region) (A + G or G + C)
Saccharomyces cerevisiae (yeast)	X53947 (SCMYO1G)	4889 (6108)	1	6108 (100)	50.0	50.0+5.8 −5.7	0 (1)	49.1	49.1+6.4 −4.8	0 (1)	0 (1)
Caenorhabditis elegans #3 (worm)	X08067 (CEMYO3)	9285 (11,604)	7	5911 (50.9)	51.6	53.0+6.7 −7.8	4 (7)	49.1	54.6+6.8 −6.6	5 (7)	5 (7)
Brugia malayi (worm)	M74000 (BRPMYOHEA)	9415 (11,766)	13	5598 (47.6)	52.3	64.8+5.6 −6.1	6 (11)	51.7	69.3+5.1 −5.4	6 (11)	7 (11)
Drosophila melanogaster (fruit fly)	M61229 (DROMHC)	18,132 (22,663)	30	8024 (35.4)	52.2	67.7+5.6 −6.2	12 (27)	52.2	85.8+2.6 −3.7	21 (27)	23 (27)
Gallus gallus (chicken)	J02714 (CHKMYHE)	24,890 (31,111)	38	5823 (18.7)	50.9	75.8+4.5 −5.2	28 (38)	49.2	90.8+2.1 −2.9	31 (38)	36 (38)
Rattus norvegicus (rat)	X04267 (RNMHCG)	20,606 (25,755)	41	6015 (23.4)	50.1	72.8+2.6 −4.3	21 (34)	54.4	80.8+3.8 −5.1	24 (34)	31 (34)
Homo sapiens (human)	M57965 (HUMBMYH7)	22,752 (28,438)	40	6008 (21.1)	54.3	86.8+3.2 −4.1	33 (39)	55.2	79.4+4.3 −4.9	25 (39)	36 (39)

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

Table 2
Summary of local scaling analysis on 339 DNA sequences of different species collected from GenBank

Species	*Caenorhabditis elegans*	Bird	Rat	Human
Number of sequences (including myosin heavy chain gene family)	75	37	46	181
Total length of sequences	465,459	306,376	267,478	1,946,379
Total length analysed	372,514	245,171	214,068	1,557,435
Total number of exon segments	533	275	278	1340
Exon segments in the analysed sections	501	257	261	1171
Average matching rate (A + G)	59.5%	68.2%	66.2%	70.3%
Number and percentage of segments with the matching rate over 70% (A + G)	186(37.1%)	143(55.6%)	124(47.5%)	665(56.8%)
Average matching rate (G + C)	71.8%	72.0%	68.7%	75.9%
Number and percentage of segments with the matching rate over 70% (G + C)	323(64.5%)	157(61.1%)	142(54.4%)	794(67.8%)
Number and percentage of segments with the matching rate over 70% (A + G or G + C)	361(72.1%)	204(79.4%)	204(78.2%)	1022(87.3%)

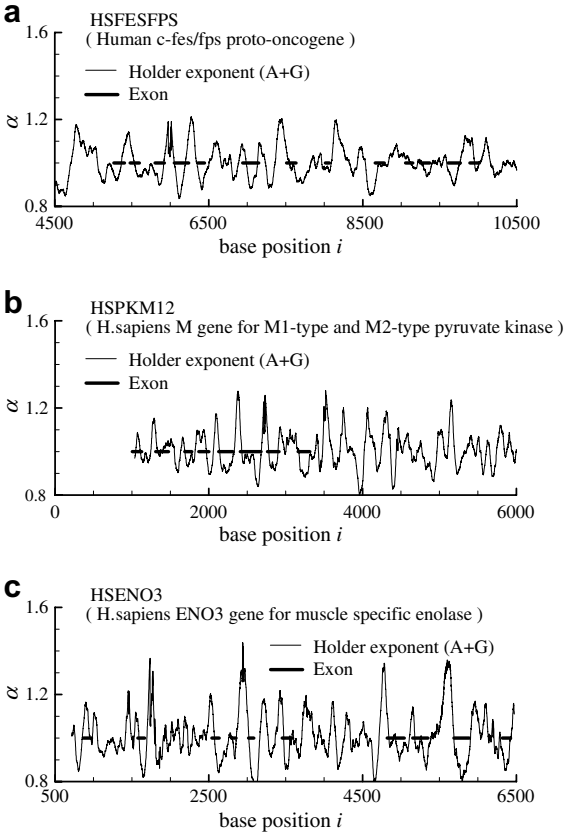


Fig. 3. Comparison between actual sites of exon segments and Hölder exponent curve of purine distribution for (a) human proto-oncogene gene, (b) human pyruvate kinase gene, and (c) human ENO3 gene.

sequences. This figure indicates that some predictions on the exon segments of the DNA sequence using sections where the Hölder exponent curve of the purine distribution is less than one are not very accurate. Yet these exon segments can fit well with sections where the G and C base distribution's Hölder exponent curve is less than one. The fifth column in Table 2 shows that the 181 human genes have a total of 1171 exon segments. Of this, total 665 segments (56.8%) with

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

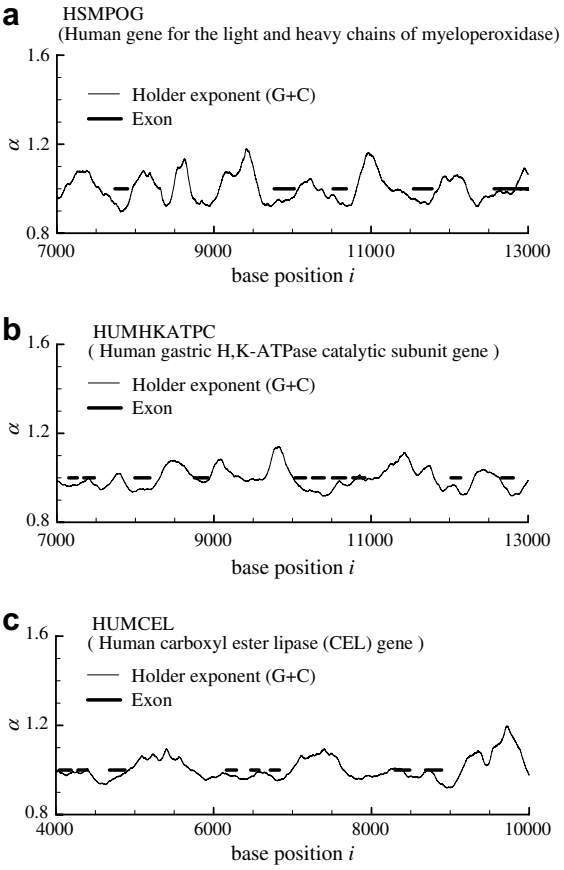


Fig. 4. Comparison between actual sites of exon segments and Hölder exponent curve of GC bases distribution for (a) human gene for light and heavy chains of myeloperoxidase, (b) human gastric H,K-ATPase catalytic subunit gene, and (c) human CEL gene.

70% length fall within the sections where the purine distribution's Hölder exponent curve is less than one and 794 segments (67.8%) with 70% length fall within the sections where the G and C base distribution's Hölder exponent curve is less than one. However, integrating the Hölder exponent curves for AG bases and GC bases shows that 1022 exon segments (87.3%) with 70% length fall within sections where the GC base or AG base distribution's Hölder exponent curve is less than one. Human myosin heavy chain genes have 36 exon segments with 70% length that fall within sections where the GC base or AG base distribution's Hölder exponent curve is less than one (see the twelfth column in Table 1). This is more than the original 33 segments, which comes from considering AG base distribution only (see the eighth column in Table 1).

As mention above, the coding sequences tend to appear in regions of the DNA sequence with abundant AG or GC bases. In other words, the coding sequences tend to exist in DNA sequence regions with a scarcity of the T base. Therefore, exon segments might be found within sections where the T base distribution's Hölder exponent curve is greater than 1. Fig. 6 compares the T base distribution's Hölder exponent curve and exon segments in several human gene DNA sequences. This figure shows that the exon segments of some human genes fit well with the $\alpha > 1$ sites of the T base's Hölder exponent curve.

Many methods of bioinformatics about predicting genes or genetic sequences that can be coded into proteins have recently been proposed. Despite the diversity of these approaches, none of the above methods is perfect. In order to achieve a better prediction, it is sometimes necessary to integrate results from various programs. The GRAIL algorithm [43], for example, collects various methods of analysis and determines regions for coding DNA sequences as its "sensors" before applying a neural network to predict DNA coding segments: one sensor, among many others, is a fractal dimension. The approach to find segments to be coded into proteins using the Hölder exponent curve proposed in this study has the advantages of simplicity, less calculation and do not need specify any window size. Perhaps in the future, this method can be integrated into GRAIL's sensor, as the fractal dimension, to help other methods of bioinformatics.

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

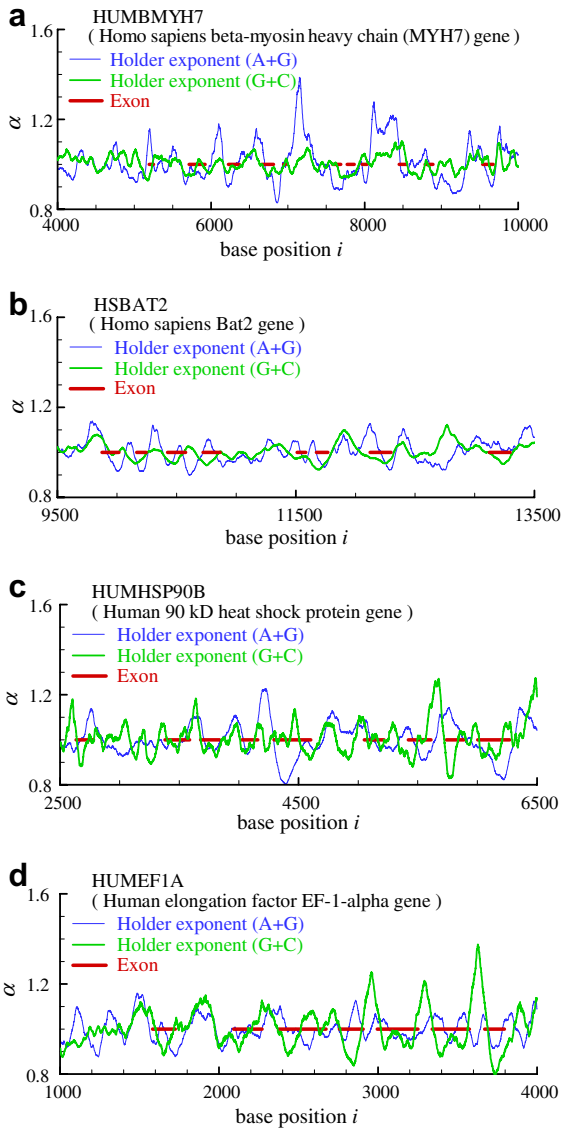


Fig. 5. Comparison between actual sites of exon segments and Hölder exponent curve of AG bases (thin line) and GC bases distribution (thick line) for (a) human myosin heavy chain gene, (b) human Bat2 gene, (c) human heat shock protein gene, and (d) human elongation factor gene.

3.2. Multifractal spectrum analyses

Previous studies use two multifractal methods to study the structure of DNA sequences, including generalized dimensions [44–49] and the multifractal spectrum [50,51]. This study focuses on the multifractal spectrum behavior of DNA sequences only. Most researchers first transfer DNA sequences in DNA walk before calculating the multifractal spectrum, but this study does adopt this approach because it involves artificial correlation. Instead, the DNA sequence is viewed directly as the set of the four letters of the alphabet (A, G, T and C). Further, different rules were applied to calculate the probability that a certain kind of base falls into the box, and multifractal spectra are then calculated using formulas mentioned before.

3.2.1. Application to myosin heavy chain gene

Fig. 7 illustrates the multifractal spectra ($f(\alpha)$ spectra) of the human cardiac β-myosin heavy chain gene, calculated by different rules. In this diagram, A, T, G, and C stand for various single base rules, respectively, while G + C

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

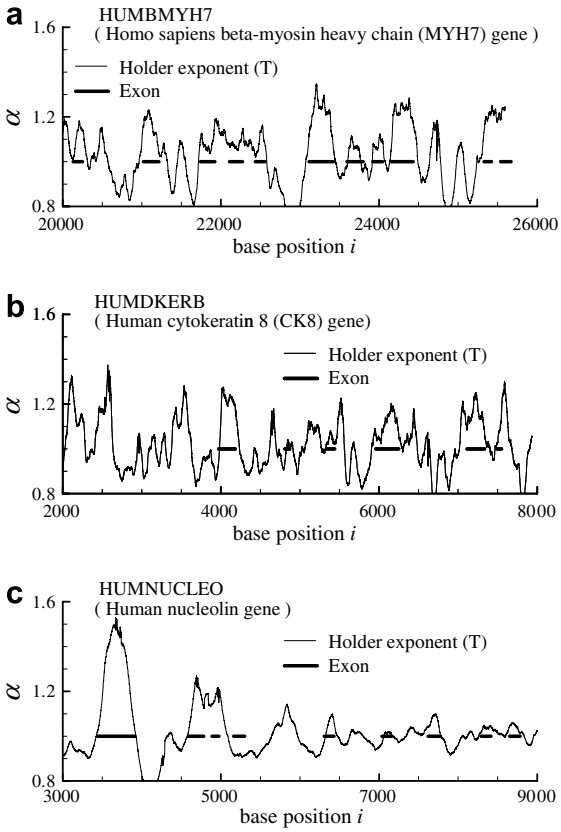


Fig. 6. Comparison between actual sites of exon segments and Hölder exponent curve of T distribution for (a) human myosin heavy chain gene, (b) human cytokeratin 8 gene, and (c) human nucleolin gene.

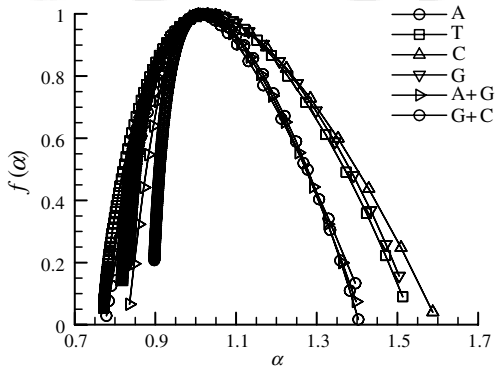


Fig. 7. Multifractal spectra of human cardiac β-myosin heavy chain genes under different rules.

represents the hydrogen bond rule and A + G represents the purine–pyrimidine rule. All the $f(\alpha)$ curves have an inverted and downward-opening parabolic shape; that is, the DNA sequence features a multifractal structure, but the opening of the $f(\alpha)$ curve varies with different rules. For example, the α value of the purine distribution (A + G) is between approximately 0.8 and 1.4.

To confirm this heterogeneity in base distribution, the positions of these bases are scrambled by a random scheme (base pair numbers, A, G, T, and C, remain the same as those in the original sequence). Fig. 8 plots the resulting $f(\alpha)$ spectra of the purine distribution and the result of original DNA sequences in the same figure. The crosses on the

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

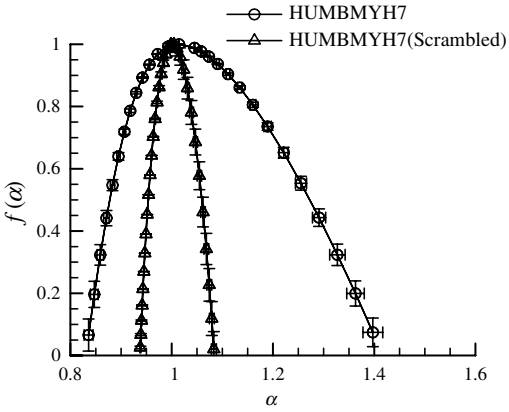


Fig. 8. Multifractal spectra for the purine distributions of human cardiac β-myosin heavy chain gene nucleotide sequence before and after shuffling.

symbols in the diagram stand for uncertainties in the values of α and $f(\alpha)$ arising from the least square fitting procedure in Eqs. (6) and (7). A wider opening of the parabola in the diagram indicates that purine bases of original DNA sequences before scrambling are not uniformly distributed along the human myosin heavy chain gene; rather, they tend to form clusters of different sizes. After scrambling, a much smaller opening curve occurs, indicating that the base sequence has a more uniform distribution. This implies that the DNA sequence base arrangement in the human myosin heavy chain gene is far more complicated than any random sequence and carries considerable messages. 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 the myosin heavy chain gene) normally observed in genomes.

All introns are deleted, and the remaining exon segments of the human's myosin heavy chain gene are stitched together to form a shorter sequence containing protein-coding regions only. Fig. 9 gives the corresponding $f(\alpha)$ spectra of purine distribution. Compared with the spectra of original DNA sequence in Fig. 8, a much narrower $f(\alpha)$ is observed. Scrambling this shorter sequence produces little difference in $f(\alpha)$ (see Fig. 9), 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 the random-walk model, which state that long-range correlation is associated with intron parts of DNA sequence.

Next, this study calculates the $f(\alpha)$ spectra of seven species DNA sequences in the myosin heavy chain gene family, ranging from yeast to human (*Homo sapiens*). Fig. 10 shows the result of purine distribution in DNA sequence. Comparing the $f(\alpha)$ spectra obtained from the seven species in Fig. 10, the gradual opening of the $f(\alpha)$ curve indicates that the purine distribution of DNA sequences of higher species is more uneven, irregular, and has more local clusters. These variations suggest an increasing complexity in the structures of DNA sequences. Again, the degree of complexity follows the evolutionary order from fungus, invertebrates, vertebrates, to human. This result agrees with Table 1, which reveals that higher species have more fragmented coding regions in their DNA sequences.

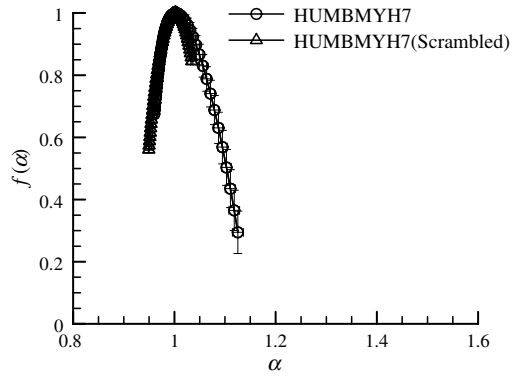


Fig. 9. Multifractal spectra of human cardiac β-myosin heavy chain gene with introns removed, before and after shuffling.

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

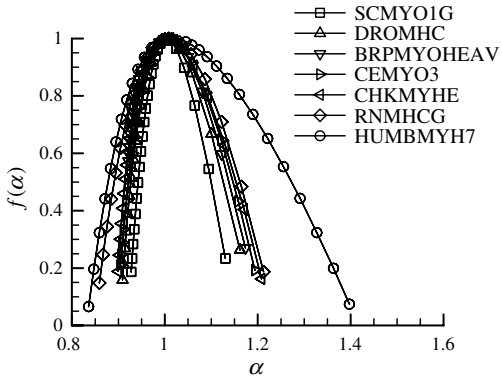


Fig. 10. Comparison of multifractal spectra of myosin heavy chain genes among different species.

The multifractal spectra of DNA sequences obtained by deleting all introns and stitching together the remaining exon segments for the corresponding seven species are shown in Fig. 11. Comparing Figs. 11 and 10, the opening of $f(\alpha)$ curves obtained from stitching together the remaining exon segments are much smaller than original DNA sequences for most species, except yeast. This means that protein-coding sequences are much more evenly distributed than the original sequence. In Fig. 11, the difference in the openings of all $f(\alpha)$ curves obtained from the seven species is insignificant; suggesting the complexity of protein-coding sequences in DNA is almost the same among different species.

3.2.2. Application to mitochondria genome

The primary function of mitochondria is to convert oxygen and nutrients into adenosine triphosphate (ATP), so they can be considered the power generators of cells. Because many species have this organelle, the DNA sequence of the mitochondrion can also be used to study evolution using the multifractal spectrum. Fig. 12 shows the multifractal spectrum for the purine distribution of the human mitochondrion DNA sequence. The multifractal spectrum obtained from the same sequence but with bases positions scrambled by using a random scheme is also provided for comparison, from which it can be inferred that the disturbance does not have a considerable influence on the $f(\alpha)$ curve opening. This phenomenon may be related to the fact that human mitochondrion DNA is extremely compact and contains no introns. A comparison of multifractal spectra for the mitochondrion DNA sequences of different species, shown in Fig. 13, shows that the openings of these $f(\alpha)$ curves are not wide, ranging from only 0.9–1.1. Considering the insignificant difference in opening size and error margins, it is clear that the order of opening size cannot be determined; i.e., degrees of complexity in the structure of these DNA sequences are about the same. In fact, except for yeast, almost all the mitochondrion DNA sequences are of the same length and, quite different from the myosin heavy chain gene family, the exon segments of higher species with inheritance messages do not scatter more widely along the DNA sequence. This may account for the insignificant opening size difference in the $f(\alpha)$ curves. Mitochondrion DNA is derived from

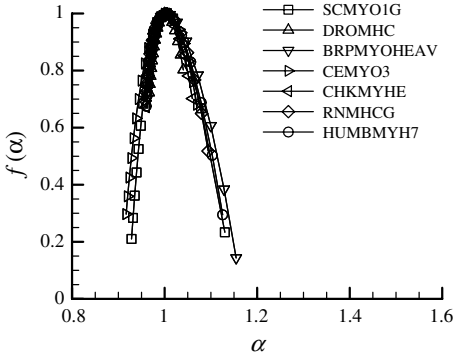


Fig. 11. Comparison of multifractal spectra of myosin heavy chain genes of different species with introns removed, before and after shuffling.

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

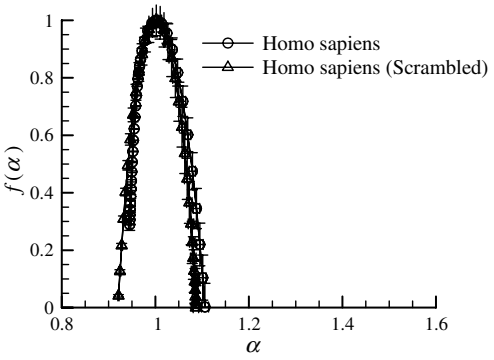


Fig. 12. Multifractal spectra for the purine distributions of human mitochondrion DNA sequence before and after shuffling.

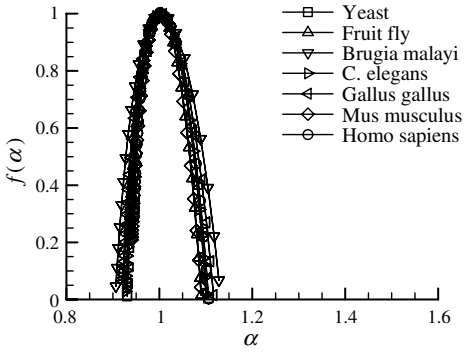


Fig. 13. Comparison of multifractal spectra of mitochondrion DNA sequences among different species.

bacteria that were engulfed by early precursors of eukaryotic cells. In mammals and for most, but not all, organisms, mitochondrion DNA is inherited from the mother. Unlike nuclear DNA in which the genes are rearranged each generation due to the process called recombination, there is usually no change in mitochondrion DNA from parent to offspring by this mechanism. Hence, the mitochondria genomes in animal cells show extensive homology in organization. This may account for the insignificant difference in opening size of the $f(\alpha)$ curves for different species.

4. Conclusion

This study employed a multifractal formalism to investigate the fractal nature of DNA sequences. Phylogenetic study of the spatial organization of nucleotide sequence collected and organized from GenBank was performed using a local scaling and multifractal spectrum analysis technique. As for the multifractal spectrum $f(\alpha)$ of the myosin heavy chain gene family, the familiar, inverted, downward-opening parabola curve shape is present. The finding demonstrates the arrangement of bases in DNA sequence does exhibit a multifractal feature. The gradual opening of the $f(\alpha)$ curves associated with different species suggests an increasing complexity in the structures of DNA sequences with evolutionary order. The widest opening $f(\alpha)$ of the parabola indicates that purine bases are not uniformly distributed along the human myosin heavy chain gene; rather, they tend to form clusters of different sizes. Therefore, it may be possible to determine the evolutionary level of an unknown genetic DNA sequence by considering the multifractal spectrum opening. Unlike myosin heavy chain gene family, the openings of the $f(\alpha)$ curves for mitochondria DNA show insignificant difference among different species. This indicates that the coding segments along the mitochondria DNA sequences of higher species do not disperse widely; and different species share similar structural complexity in the mitochondria DNA sequences.

The local scaling analysis of the myosin heavy chain gene family suggests that higher species have more fragmented exon segments that take up a smaller portion of the whole sequence, and more exon segments fall within sections where

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

the purine distribution's Hölder exponent α is less than one; this is related to the phenomenon that base (purine or pyrimidine) distribution in the DNA sequence of higher species tends to be uneven, irregular, and dense in local regions. This result also agrees with the observed increase in the width of the $f(\alpha)$ curve openings for higher-level species. Furthermore, analyzing 339 DNA sequences obtained from *C. elegans*, birds, rats, and humans show that there is also an apparent tendency of increasing correlation between the exon locations in the sequence and sections where the purine distribution's Hölder exponent α is less than one with phylogenetic order. On the other hand, there are cases where coding sequences in genes tend to contain more G and C bases than non-coding sequences, and exon segments may also fit well with the G and C base distribution's Hölder exponent curve where the Hölder exponent is less than one. Therefore, considering the Hölder exponent curve for either AG or GC bases distribution only may not achieve the best prediction for coding segments. Simultaneously considering Hölder exponent curves calculated by adopting different rules, such as the purine–pyrimidine rule and hydrogen bond rule, may have a complementary effect and can lead to a more accurate prediction. The phenomenon that exon segments of higher species fall within sections where the purine distribution's Hölder exponent curve is less than one may be due to clustering pyrimidine in DNA sequence intron in the process of evolution. While possible reasons for exon segments of higher species fall within sections where the Hölder exponent curve of the GC base distribution is less than one (i.e., GC rich) include adaptation to elevated temperatures, active gene transcription, mutation bias, or biased gene conversion (BGC) [41]. The present analysis substantiates the feasibility of finding exon segments in DNA sequences using the local scaling exponent. Because the method is simple, with less calculation work and dispenses with specifying window sizes. The authors hope that other more complicated bioinformatics methods for analyzing DNA sequences can benefit from this study in the future.

Acknowledgements

This study is sponsored by Chia Nan University of Pharmacy and Science, Taiwan, ROC under the project number: CN 9637.

References

[1] Gates MA. J Theor Biol 1986;119:319–28.
[2] Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, et al. Nature 1992;356:168–70.
[3] Mandelbrot BB. The fractal geometry of nature. New York: Freeman; 1983.
[4] Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, et al. Science 1985;228:953–8.
[5] Churchill GA. Bull Math Biol 1989;51:79–94.
[6] Fickett JW, Torney DC, Wolf DR. Genomics 1992;13:1056–64.
[7] Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Phys Rev E 1994;49:1685–9.
[8] Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Stanley HE, Stanley MHR, et al. Biophys J 1993;65:2673–9.
[9] Ossadnik SM, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Peng C-K, et al. Biophys J 1994;67:64–70.
[10] Havlin S, Buldyrev SV, Goldberger AL, Mantegna RN, Ossadnik SM, Peng C-K, et al. Chaos, Solitons & Fractals 1995;6:171–201.
[11] Voss RF. Phys Rev Lett 1992;68:3805–8.
[12] Bickel DR. Chaos, Solitons & Fractals 2002;13:491–7.
[13] Zhang L, Sun T. Chaos, Solitons & Fractals 2005;23:1077–85.
[14] Kim JK, Yang SI, Kwon YH, Lee EI. Chaos, Solitons & Fractals 2005;23:1795–807.
[15] Yu Z-G, Wang B. Chaos, Solitons & Fractals 2001;12:519–26.
[16] Yu Z-G, Anh V. Chaos, Solitons & Fractals 2001;12:1827–34.
[17] Hao B-L. Physica A 2000;282:225–46.
[18] Hao B-L, Lee HC, Zhang S-Y. Chaos, Solitons & Fractals 2000;11:825–36.
[19] Yu Z-G, Hao B-L, Xie H-M, Chen G-Y. Chaos, Solitons & Fractals 2000;11:2215–22.
[20] Zhang L, Jiang Z. Chaos, Solitons & Fractals 2004;22:947–55.
[21] Xiao Y, Huang Y. Chaos, Solitons & Fractals 2004;20:497–502.
[22] Sun T, Zhang L, Chen J, Jiang Z. Chaos, Solitons & Fractals 2004;20:1075–84.
[23] Cheng J, Zhang L. Chaos, Solitons & Fractals 2005;25:339–46.
[24] Zhang L, Chen J. Chaos, Solitons & Fractals 2005;24:115–23.
[25] Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI. Phys Rev A 1986;33:1141–51.
[26] Grassberger P. Phys Lett A 1983;97:227–30.
[27] Hentschel H, Procaccia I. Physica D 1983;8:435–44.
[28] Chhabra AV, Jensen RV. Phys Rev Lett 1989;62:1327–30.
[29] Chhabra AB, Meneveau C, Jensen RV, Sreenivasan KR. Phys Rev A 1989;40:5284–94.

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078

[30] Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng C-K, et al. Phy Rev E 1995;51:5084–91.
[31] Viswanathan GM, Buldyrev SV, Havlin S, Stanley HE. Biophys J 1997;72:866–75.
[32] Viswanathan GM, Buldyrev SV, Havlin S, Stanley HE. Physica A 1998;249:581–6.
[33] Provata A, Almirantis Y. J Stat Phys 2002;106:23–56.
[34] Provata A, Almirantis Y. Physica A 1997;247:482–96.
[35] Almirantis Y, Provata A. J Stat Phys 1999;97:233–62.
[36] Provata A. Physica A 1999;264:570–80.
[37] Raghavan S, Hariharan R, Brahmachari SK. Gene 2000;242:275–83.
[38] Burset M, Guigo R. Genomics 1996;34:353–67.
[39] Bernardi G, Bernardi G. J Mol Evol 1990;31:282–93.
[40] Vinogradov AE. Nucleic Acids Res 2003;31:1838–44.
[41] Vinogradov AE. Nucleic Acids Res 2003;31:5212–20.
[42] Kotlar D, Lavner Y. Genome Res 2003;13:1930–7.
[43] Uberbacher EC, Mural RJ. Proc Natl Acad Sci USA 1991;88:11261–5.
[44] Glazier JA, Raghavachari S, Berthelsen CL, Skolnick MH. Phys Rev E 1995;51:2665–8.
[45] Anh V, Lau K-S, Yu Z-G. J Phys A 2001;34:7127–39.
[46] Anh V, Lau K-S, Yu Z-G. Phys Rev E 2002;66:031910-1–0-9.
[47] Bershadskii A. Phys Lett A 2001;284:136–40.
[48] Yu Z-G, Anh V, Lau K-S. Physica A 2001;301:351–61.
[49] Yu Z-G, Anh V, Lau K-S. Phys. Rev., E 2001;64:031903-1–3-9.
[50] Gutierrez JM, Rodriguez MA, Abramson G. Physica A 2001;300:271–84.
[51] Oiwa NN, Goldman C. Phys Rev Lett 2000;85:2396–9.

Please cite this article in press as: Su Z-Y et al., Local scaling and multifractal spectrum analyses of DNA ..., Chaos, Solitons & Fractals (2007), doi:10.1016/j.chaos.2007.09.078