# Wavelet-based multifractal analysis of DNA sequences by using chaos-game representation*

Han Jia-Jing(韩佳静) and Fu Wei-Juan (符维娟)†

*Surface Physics Laboratory (National Key Laboratory) and Physics Department, Fudan University, Shanghai* 200433, *China*

(Received 24 March 2009; revised manuscript received 9 June 2009)

Chaos game representation (CGR) is proposed as a scale-independent representation for DNA sequences and provides information about the statistical distribution of oligonucleotides in a DNA sequence. CGR images of DNA sequences represent some kinds of fractal patterns, but the common multifractal analysis based on the box counting method cannot deal with CGR images perfectly. Here, the wavelet transform modulus maxima (WTMM) method is applied to the multifractal analysis of CGR images. The results show that the scale-invariance range of CGR edge images can be extended to three orders of magnitude, and complete singularity spectra can be calculated. Spectrum parameters such as the singularity spectrum span are extracted to describe the statistical character of DNA sequences. Compared with the singularity spectrum span, exon sequences with a minimal spectrum span have the most uniform fractal structure. Also, the singularity spectrum parameters are related to oligonucleotide length, sequence component and species, thereby providing a method of studying the length polymorphism of repeat oligonucleotides.

**Keywords:** chaos game representation (CGR), multifractal, wavelet transform modulus maxima(WTMM), singularity spectrum

**PACC:** 0250, 0555, 8710

## 1. Introduction

In 1990 Jeffrey proposed a method of using chaos game representation (CGR) to visualize DNA primary sequence structure.[1] CGR images of DNA sequences show various geometric patterns, such as parallel lines, squares, rectangles and triangles. The major advantage of the CGR method is that it can deal with a DNA sequence of any length including entire genomes. Patrick *et al.* provided a link between CGR images and genomic signatures,[2] and so the CGR method became a powerful tool of genomic signatures. The key observation behind this concept is that the variation between CGR images along a genome is smaller than the variation among genomes. In the further research, a $k$-th-order frequency matrix (FCGR) is extracted from a CGR image,[3] and each element in the matrix corresponds to the frequency of an oligonucleotide with length $k$. Thus the frequencies of oligonucleotides found in a sequence are displayed in a square CGR image. Wang *et al.* challenged the idea, indicating that CGR was merely a representation of nucleotide, dinucleotide, and trinucleotide frequencies.[4] It was pointed out in their research that short oligonucleotide frequencies provided major organizational information of a DNA sequence, while longer oligonucleotide frequencies bring more precise information about the details of the genome organization. Milan designed a procedure that converts the CGR images into a 'spectrum-like' format serving as the CGR representation of short DNA sequences (about 100 bases instead of 100 k bases that were considered by Jeffrey and others).[5] In recent years, the CGR method has been extended to the research of secondary structure of the genome,[6] protein sequences (amino acid sequences) and protein structures.[7–9]

Multifractal analysis,[10] initially proposed to treat turbulence data, is a useful way to characterize the spatial inhomogeneity of both theoretical and experimental fractal patterns. In the last decade it has been applied successfully to many different fields including time series analysis and financial modelling.[11] Gutiérrez *et al.* proposed an alternative 1D iterated function system representation of DNA sequences instead of Jeffrey's 2D representation, and calculated a multifractal spectrum to analyse the correlation length of a given coding or non-coding sequence.[12] Yu *et al.*[13,14] provided a simple yet powerful 1D visualization method of DNA sequences and discussed the multifractal property of the measure representation.

*Project supported by the Science and Technology Commission of Shanghai Municipality (Grant No. 05DZ19747) and the National Basic Research Program of China (Grant No. 2006CB504509).

†Corresponding author. E-mail: softmat@fudan.edu.cn

© 2010 Chinese Physical Society and IOP Publishing Ltd

Then the difference between a DNA sequence and a random sequence was expanded to distinguish DNA sequences themselves in more detail and the classification of bacteria was carried out. Furthermore, Yu *et al.* performed multifractal and correlation analyses of the measures on the CGR representation of protein sequences.[15]

However, multifractal research on fractal patterns of CGR images has been carried out less frequently. Fu *et al.* used the box counting method to calculate the multifractal spectrum of CGR images and recognize gene sequences by the spectrum parameters.[16,17] The CGR images are experimental multifractal images with noise as a result of molecular evolution driven by mutation and random drift. As a simple denoise method, the relative probability set used in Refs. [16] and [17] can but give a limited scaling range of a CGR image. In recent research on practical multifractal images, such as clouds and surface profiles of physics process, the wavelet transform modulus maxima (WTMM) method was used to obtain complete multifractal spectra.[18−20] In the present paper, we apply this 2D wavelet-based multifractal method to CGR images of DNA sequences and obtain the character of statistical distribution of oligonucleotides in DNA sequences. Firstly, the edge image is extracted from a CGR image with the WTMM method, and its fractal scaling range is discussed. Secondly, the multifractal spectra of three different component sequences (exon, intron and blank) are calculated and compared with each other. Finally, the dependence of spectrum parameters on oligonucleotide length, sequence component and species is used to investigate the length polymorphism of repeat oligonucleotides.

# 2. System and methods

## 2.1. Multifractal signal—CGR images of DNA sequences

The CGR method is an algorithm which produces pictures revealing patterns in DNA sequences by an iterative mapping technique.[1] The CGR geometric shape generated by DNA sequences is planar and is confined by the four possible nucleotides as vertices of a binary square. Suppose the length of a sequence is $L$, and assign the binary CGR vertices to the four nucleotides as $A = (0, 0)$, $C = (0, 1)$, $G = (1, 1)$, $T = (1, 0)$; the CGR positions $\mathrm{CGR}_i$ of each nucleotide $\mathrm{g}_i$ can be calculated by moving a pointer to half of the distance between the previous position and the current binary representation as follows:

$$\mathrm{CGR}_i = \mathrm{CGR}_{i-1} + 0.5(\mathrm{g}_i - \mathrm{CGR}_{i-1}), \qquad (1)$$

where $i = 1, \ldots, L$ and $\mathrm{CGR}_0 = (0.5, 0.5)$.

As a consequence, the trajectory of identical sequences, independent of their location in different genomes, will converge in the CGR space. Since there are four kinds of nucleotides, a $k$-length oligonucleotide has $4^k$ kinds of different compositions. The CGR image is equally divided into $2^k \times 2^k$ lattices in each of which the spots represent one kind of $k$-length oligonucleotide composition, and then $4^k$ lattices correspond to such $4^k$ kinds of oligonucleotide compositions. Also it was pointed out in Ref. [2] that CGR images presented fractal patterns, and different species exhibited different fractal patterns.

## 2.2. Theory of two-dimensional wavelet transform modulus maxima (WTMM)

Mallat and Zhong[21] and his collaborators carefully studied WTMM theory and its application. Here is a simple introduction to two-dimensional WTMM theory.

Let $\theta(x, y)$ be a smooth and twice differentiable function, whose integral is equal to 1 and converges to 0 at infinity. The two-dimensional dyadic wavelet transform (DWT) of a function $f(x,y) \in L^2(R^2)$ is a set of functions $(W^1_{2^j} f(x, y), W^2_{2^j} f(x, y))$, which are, respectively, the partial derivatives along the horizontal and the vertical directions of the convolution of $f(x, y)$ that is smoothed by function $\theta(x, y)$ dilated along a dyadic sequence $(2^j)_{j\in Z}$. The DWT at a scale of $2^j$ is given by

$$\begin{aligned}\begin{pmatrix} W^1_{2^j} f(x,y) \\ W^2_{2^j} f(x,y) \end{pmatrix} &= f * \begin{pmatrix} \psi^1_{2^j} \\ \psi^2_{2^j} \end{pmatrix} \\ &= 2^j \begin{pmatrix} \dfrac{\partial}{\partial x}(f * \theta_{2^j})(x,y) \\ \dfrac{\partial}{\partial y}(f * \theta_{2^j})(x,y) \end{pmatrix}, \end{aligned} \qquad (2)$$

where $\psi^1(x, y)$ and $\psi^2(x, y)$ are the wavelets under study and $*$ denotes the convolution operation. Equation (2) indicates that the above set of functions can be viewed as the two components of the gradient vector of $f(x, y)$ smoothed by $\theta(x, y)$ at scale $2^j$, i.e.

$$\begin{pmatrix} W^1_{2^j} f(x,y) \\ W^2_{2^j} f(x,y) \end{pmatrix} = 2^j \cdot \nabla(f * \theta_{2^j})(x,y). \qquad (3)$$

The wavelet transform modulus maxima are then evaluated by using the modulus $M_2^j f(x, y)$ and argument $A_2^j f(x, y)$ of the gradient vector $W_2^j f(x, y)$. Comparing the modulus $M_2^j f(x, y)$ of point $(x, y)$ with another two neighbouring modules in the direction specified by$A_2^j f(x, y)$, the wavelet transform modulus maxima are defined at the positions where the modulus is locally maximum. The multiscale edge points can be obtained by calculating the modulus maxima points of $f(x, y)$ at different scales. Furthermore, the modulus maxima lines on a scale could be acquired by connecting the edge points along the gradient direction on this scale.

## 2.3. Multifractal analysis based on WTMM theory

The multifractal analysis is employed to describe the spatial proportion distribution of geometry and physical parameters,[10] as well as the indescribable distributive characters of physics-quantum in a self-similarity structure system. In multifractal analysis, a singularity exponent $\alpha$ and density function $f(\alpha)$ may be used to carry out the quantitative evaluation of the degree of reaction probability distribution inhomogeneity.[11] In the box counting method of multifractal analysis,[14] a square image with length $l$ can be divided into $2^k \times 2^k$ grids, and then the length of a box $\varepsilon = l/2^k$. The probability of pixel number in the box $(i, j)$ is defined by $P_{ij}(\varepsilon) = h_{ij}/\sum h_{ij}$, where $h_{ij}$ is the pixel number in the box $(i, j)$. The partition function $Z_q(\varepsilon)$ is defined and expressed as a power law of scale $\varepsilon$ with an exponent $\tau(q)$ as follows:

$$Z_q(\varepsilon) = \sum P_{ij}(\varepsilon)^q = \varepsilon^{\tau(q)}, \tag{4}$$

where $q$ is a real parameter, which indicates the moment order of the measure. The exponent $\tau(q)$ can be obtained from the slope of the $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curve. Therefore, information at different scales and moments can be obtained by the partition function $Z(q,\varepsilon)$.

The multifractal is described as

$$P_{ij}(\varepsilon) \sim \varepsilon^{\alpha}, \tag{5}$$

$$N \sim \varepsilon^{f(\alpha)}, \tag{6}$$

where the exponent $\alpha$ is the singularity of the probability subset which can be used as a token of the local regularity of a fractal image, $N$ is the number of boxes of size $\varepsilon$ with the same probability, and $f(\alpha)$ is the fractal dimension of the subset. The singularity spectrum $f(\alpha)$ is then determined by Legendre transformation[13] of the $\tau(q)$ curve according to the following expressions:

$$f(q) = \min_q(\alpha q - \tau(q)), \tag{7}$$

$$\alpha(q) = \frac{\mathrm{d}\tau(q)}{\mathrm{d}q}. \tag{8}$$

The hybrid wavelet and multifractal method applied to signal and image was discussed in detail in Ref. [19]. The compactly supported character of the wavelet is the basic reason for the application of the wavelet to the study of fractals. The wavelet is the analogy of the standard box counting method, corresponding to a cover with a box at scale $2^j$. Then the wavelet transform coefficient $W_2^j f(x, y)$ of function $f(x, y)$ equals one of the parameters in fractal theory, such as probability, mass, etc. Here, the modulus maxima $M_2^j f(x, y)$ of wavelet transform coefficient $W_2^j f(x, y)$ is treated as a kind of fractal parameter, and so a scale-adaptive partition function $\mathrm{Z}(q, 2^j)$ is defined as

$$Z(q, 2^j) = \sum_{L\in L(s)} \Big(\sup_{(x,y,s')\in L, s'\le s} M_\psi f(x, y)\Big)^q \propto s^{\tau(q)}, \quad q \in R, \tag{9}$$

where $s$ is the scale $2^j$, $\psi$ is the wavelet under study and $L(s)$ is the modulus maxima lines at scale $s$. The modulus maximum $M_s f(x, y)$ of function $f(x, y)$ at scale $s$ satisfies

$$M_\psi[f](L_{x,y}(s)) \sim s^{\alpha(x,y)}. \tag{10}$$

Therefore, the singularity spectrum $f(\alpha)$ based WTMM theory can be obtained according to expressions (7) and (8). In real calculation, the value of $q$ cannot be infinite. One can take the maximum $|q|$ when $\alpha$ and $f(\alpha)$ approach to the saturated values with the increase of $q$. The maximum distribution probability $P_{ij}(\varepsilon)$ increases with the decrease of $\alpha_{\min}$; and the minimum distribution probability $P_{ij}(\varepsilon)$ decreases with the increase of $\alpha_{\max}$. Therefore $f(\alpha)$ in the case of small values of $\alpha$ describes the property of a large distribution probability subset, and $f(\alpha)$ in the case of large values of $\alpha$ describes the property of a small distribution probability subset, also $\Delta f = f(\alpha_{\min}) - f(\alpha_{\max})$ corresponds to the difference between the maximal and the minimal probability subset numbers. The multifractal spectrum $f(\alpha)$ with the shape of a right-hook shows that the small

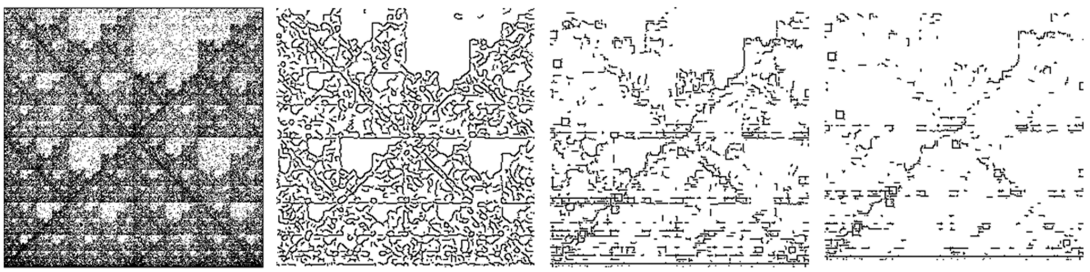
probability subsets take the leading part, $f(\alpha)$ with the shape of a left-hook shows that the large probability subsets take the leading part, and $f(\alpha)$ with the shape of a symmetrical bell jar shows that the numbers of maximal and minimal probability subsets are the same.

# 3. Results and discussion

## 3.1. Extraction of multiscale edge images from CGR images

In the application of the WTMM method to the edge extraction of CGR images, the smoothing function $\theta(x, y)$ is chosen to be a Gaussian function. Since the modulus maximum site is the edge point, the tendency of modulus maxima can reflect the variation of edge points. The singularities of CGR images can be extracted accurately by the edge images that denote the distribution of edge intensity.

Firstly, two DNA sequences, each with 100 k base length, are selected as research sequences, respectively from the human and *Caenorhabditis elegans* genome. Their CGR images were plotted in Figs. 1(a) and 2(a) according to expression (1), and different fractal patterns can be observed. Secondly, the multiscale edge image is extracted from a CGR image divided into $2^8 \times 2^8$ grids, and the edge image describes the detailed distribution information of 8-length oligonucleotides in the sequence. The edge images at scales $s = 2^1$, $2^3$ and $2^5$ of two research sequences are shown in Figs. 1(b)–1(d) and Figs. 2(b)–2(d) respectively, and the image details are richest at scales $s = 2^1$, but filtered mostly at scale $2^5$.



(a) (b) (c) (d)

**Fig. 1.** Human CGR image (a) and edge images at scales $s = 2$, $2^3$, and $2^5$, (b), (c) and (d), respectively.



(a) (b) (c) (d)

**Fig. 2.** *Caenorhabditis elegans* CGR image (a) and edge images at scales $s = 2$, $2^3$, and $2^5$ (b), (c) and (d), respectively.

Considering the multiscale edge analysis theoretically, the smaller the scale size is, the better the locations of WTMM and the singularity point will be. However, when the scale is small, pseudo extreme points will emerge because the wavelet coefficients are sensitive to the noise. In contrast, when the scale becomes large, the extreme sites will be relatively stable, but the drift of WTMM positions cause some detail points to smooth. To solve this problem, we choose the smallest scale $s = 2^1$, and make use of an algorithm

that can self-adaptively calculate the double thresholds to get rid of the pseudo extreme points. Actually Figs. 1 and 2 show that the de-pseudo edge images at scale $2^1$ can reflect the morphology of CGR images reasonably. The quantity of edge points on an edge image can be described by its fractal dimension (FD); the more edge points there are, the higher the FD will be. The calculation shows that Fig. 1(b) has a higher FD value (1.6863) than Fig. 2(b) (1.6629), and it indicates that Fig. 1(b) is more abundant in details and rougher in morphology than Fig. 2(b).

## 3.2. Range of scale-invariance

The linearity of the $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curve is very important for the multifractal, and its linear range is called the scaling range. A regular multifractal has strict linearity in $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curves at all values of moment $q$, and this linearity can approach zero infinitely when $\ln\varepsilon \to -\infty$. For a random multifractal (for example clouds, thin film growth and CGR images of DNA sequences), however, scaling usually takes place in a limited range, and the scale-invariance range is related to the calculation method.

In the multifractal analysis of CGR images based on the box counting method,[16,17], the partition function $Z_q(\varepsilon)$ is calculated from expression (4) and the $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curves are shown in Fig. 3(a). The moment order $q$ ranges from –30 to 30 with interval 5, and the box size $\varepsilon$ is taken from $1/2^{10}$ to $1/2^3$. It can be seen that the slope of the $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curve has a pronounced deviation from linearity, especially the range of scale-invariance is divided into two segments when $q < 0$. When the WTMM method is used to calculate the partition function $Z(q, 2^1)$ (as expression (9)), the $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curves in Fig. 3(b) show better linearity than that in Fig. 3(a). By selecting an appropriate filter length and threshold parameters of de-pseudo edge, the range of scale-invariance can be extended to three orders of magnitude. Thus, the WTMM method can extract the multifractal character from CGR images with a more extended scale-invariance range.

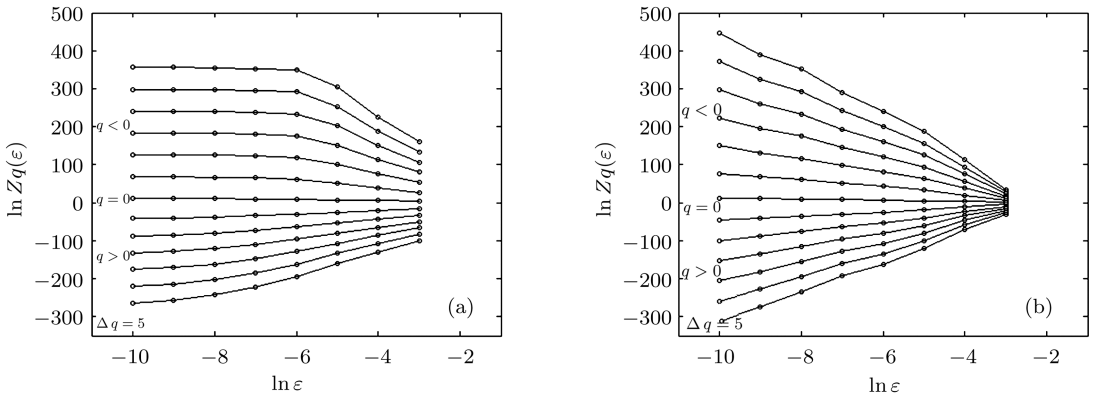



**Fig. 3.** $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curves calculated by the box counting method (a) and the WTMM method (b) separately.

## 3.3. Multifractal analysis of CGR images

Jeffrey considered that the CGR images possessed fractal structures for the properties of statistical self-similarity, and they could be studied with the fractal method.[1] The recent study indicates that more information can be obtained from a multifractal spectrum than from a fractal dimension value. In this work, a multifractal method based on the WTMM method is utilized to analyse the fractal character of CGR images, and the singularity spectrum $f(\alpha)$ can be calculated from expressions (7–9).

We extract exon, intron and blank sequences from a Caenorhabditis elegans genome sequence with length 10 M, and the three kinds of sequences are components independent of each other in genome. Figure 4 shows the singularity spectra $f(\alpha)$ of CGR images divided into $2^8 \times 2^8$ grids, and then the spectra describe the statistic distribution of 8-length oligonucleotides in the sequences under study. The singularity spectrum shape of the exon sequence is like a hook to the left, while those of intron and blank sequences

are almost symmetrical.

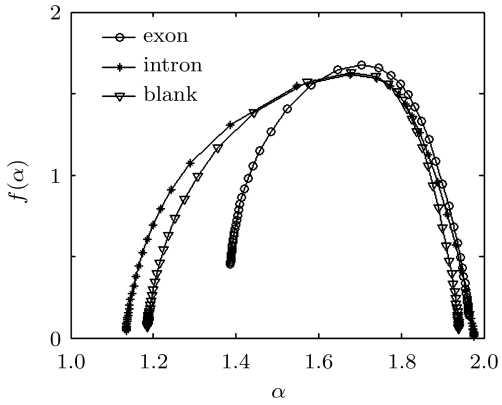



**Fig. 4.** Singularity spectrum of *Caenorhabditis elegans* ($k = 8$).

The parameters of singularity spectra in Fig. 4 are summarized in Table 1, and it can be seen that spectrum span $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$ is the most notable parameter to distinguish different sequences. Spectrum span $\Delta\alpha$ characterizes the distribution of singularity intensities, the narrower the span $\Delta\alpha$ is, the more uniform the distribution will be. Table 1 shows that the exon sequence has the narrowest spectrum span $\Delta\alpha$, and the intron sequence has the widest spectrum span $\Delta\alpha$. It indicates that the 8-length oligonucleotide distribution of the exon sequence is the most uniform, and that of the intron sequence is the most nonuniform. Additionally, the difference between spectra mostly appear in a small $\alpha$ region that describes the property of a large distribution probability subset.

**Table 1.** Parameters of the singularity spectrum in Fig. 4.

| | $\alpha_{\min}$ | $\alpha_{\max}$ | $\Delta\alpha$ | $\Delta\alpha'$ | $\alpha_0$ | $\Delta f(\alpha)$ | $f(\alpha_0)$ |
|---|---|---|---|---|---|---|---|
| exon | 1.387 | 1.964 | 0.577 | –0.058 | 1.704 | –0.315 | 1.677 |
| intron | 1.134 | 1.976 | 0.842 | –0.245 | 1.677 | –0.033 | 1.619 |
| blank | 1.186 | 1.940 | 0.754 | –0.233 | 1.679 | –0.012 | 1.630 |

$\alpha_0$: $\alpha(q = 0)$, $\Delta\alpha' = (\alpha_{\min} + \alpha_{\max}) - 2 \times \alpha_0$.

In the further study on the singularity spectra of different component sequences, 10 DNA segments with length 1 M are selected from human chromosome 21, and their exon, intron and blank sequences are extracted respectively. A comparison between the values of average spectrum span $\Delta\alpha$ shown in Table 2 indicates that the relation between different component sequences is the same as that in Table 1. Thus a general rule may be inferred that exon sequences have the most uniform distribution of oligonucleotides in DNA sequences in the three kinds of sequences mentioned above. Additionally, it can be seen that the differences in singularity spectrum span among blank sequences are more distinct than among exon sequences or intron sequences.

**Table 2.** Singularity spectrum span $\Delta\alpha$ of 10 human DNA sequences ($k = 8$).

| sequence | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | A | S | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| exon | 0.5917 | 0.5841 | 0.6485 | 0.6192 | 0.5889 | 0.6383 | 0.5023 | 0.6164 | 0.6137 | 0.5832 | 0.5986 | 0.041 | 6.8% |
| intron | 0.7759 | 0.7937 | 0.8124 | 0.7446 | 0.8367 | 0.7015 | 0.7813 | 0.7230 | 0.6908 | 0.8408 | 0.7701 | 0.054 | 7.0% |
| blank | 0.7750 | 0.7389 | 0.7377 | 0.6594 | 0.7429 | 0.7315 | 0.5933 | 0.7927 | 0.7465 | 0.6559 | 0.7174 | 0.062 | 8.6% |

A: average. S: standard deviation. R: relative deviation.

In genome structure research, the 1–10 length repeat oligonucleotides distribute widely in the genome of eukaryotes and some prokaryotes, but not stochastically. The oligonucleotides repeat in different classes between coding regions and non-coding regions, and the longer oligonucleotide frequencies describe more precisely the structure of gemome organization. The genesis and the evolution of oligonucleotide repeat polymorphisms are complex and may have several typical biological functions. Yu *et al.*[14] discussed the fractal character of human chromosome 22 for oligonucleotide length $k = 1, \ldots, 8$, and found that some fractal parameters of $k = 6, 7, 8$ were very close to each other. Then, we continue the oligonucleotide repeat polymorphism study by the relation between the singularity spectrum and oligonucleotide length

$k$. Theoretically speaking, the probability including $k$-oligonucleotide decreases as $k$ increases, which will cause the singularity spectrum to shift toward the right and leads its parameters to change monotonically. Here, an exon sequence is extracted from a *Caenorhabditis elegans* DNA sequence with length 10 M. Figure 5 shows the singularity spectra with oligonucleotide length $k$ changing from 5 to 9.

Figure 5 shows that the variation of singularity spectrum with oligonucleotide length $k$ accords with the theoretical analysis, except the singularity spectrum of $k = 6$. Also, the shift of the spectrum becomes slower and slower with $k$ increasing, especially when $k$ reaches 6. It accords with the viewpoint in Ref. [14] that the increase is stable when $k$ reaches 6. Furthermore, the multifractal calculation is extended to 18 kinds of species, DNA sequences with length 10 M are selected and their exon, intron blank sequence are extracted respectively. The singularity spectra are calculated, and the values of spectrum span $\Delta\alpha$ for four different values of oligonucleotide length $k$, i.e. $k = 6, 7, 8$ and 9 are shown in Table 3.

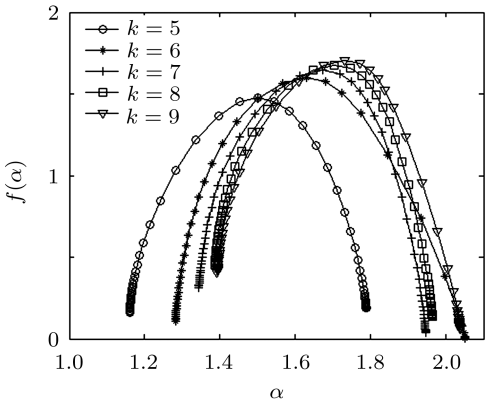



**Fig. 5.** Singularity spectra with different values of $k$.

**Table 3.** Singularity exponent span $\Delta\alpha$ versus oligonucleotide length $k$, component and species.

| | $\Delta\alpha$(exon) | | | | $\Delta\alpha$(intron) | | | | $\Delta\alpha$(blank) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $k$ | | | | $k$ | | | | $k$ | | | |
| | 6 | 7 | 8 | 9 | 6 | 7 | 8 | 9 | 6 | 7 | 8 | 9 |
| 1 | 0.671 | 0.600 | 0.636 | 0.698 | 0.734 | 0.673 | 0.767 | 0.811 | 0.676 | 0.745 | 0.733 | 0.858 |
| 2 | 0.770 | 0.602 | 0.577 | 0.645 | 0.843 | 0.818 | 0.842 | 0.940 | 0.715 | 0.777 | 0.754 | 0.888 |
| 3 | 0.584 | 0.599 | 0.613 | 0.680 | 0.573 | 0.683 | 0.819 | 0.733 | 0.688 | 0.659 | 0.656 | 0.730 |
| 4 | 0.507 | 0.568 | 0.705 | 0.735 | 0.754 | 0.617 | 0.790 | 0.928 | 0.713 | 0.812 | 0.803 | 0.795 |
| 5 | 0.608 | 0.577 | 0.554 | 0.650 | 0.663 | 0.668 | 0.924 | 0.791 | 0.541 | 0.552 | 0.710 | 0.916 |
| 6 | 0.521 | 0.533 | 0.710 | 0.659 | 0.711 | 0.670 | 0.693 | 0.815 | 0.748 | 0.752 | 0.661 | 0.817 |
| 7 | 0.506 | 0.597 | 0.592 | 0.663 | 0.596 | 0.661 | 0.818 | 0.998 | 0.486 | 0.635 | 0.759 | 0.988 |
| 8 | 0.587 | 0.573 | 0.578 | 0.692 | 0.555 | 0.603 | 0.804 | 0.840 | 0.519 | 0.621 | 0.759 | 0.850 |
| 9 | 0.482 | 0.567 | 0.552 | 0.640 | 0.637 | 0.605 | 0.777 | 0.904 | 0.552 | 0.751 | 0.877 | 0.984 |
| 10 | 0.542 | 0.603 | 0.578 | 0.685 | 0.669 | 0.581 | 0.694 | 0.833 | 0.509 | 0.608 | 0.711 | 0.868 |
| 11 | 0.454 | 0.458 | 0.537 | 0.616 | 0.504 | 0.572 | 0.718 | 0.838 | 0.646 | 0.528 | 0.753 | 0.871 |
| 12 | 0.517 | 0.575 | 0.682 | 0.747 | 0.616 | 0.561 | 0.709 | 0.799 | 0.590 | 0.546 | 0.594 | 0.738 |
| 13 | 0.486 | 0.579 | 0.636 | 0.675 | 0.545 | 0.756 | 0.660 | 0.804 | 0.516 | 0.536 | 0.637 | 0.842 |
| 14 | 0.581 | 0.581 | 0.611 | 0.553 | 0.790 | 0.697 | 0.804 | 0.887 | 0.519 | 0.576 | 0.744 | 0.885 |
| 15 | 0.580 | 0.547 | 0.628 | 0.657 | 0.572 | 0.592 | 0.593 | 0.807 | 0.564 | 0.515 | 0.639 | 0.717 |
| 16 | 0.651 | 0.602 | 0.723 | 0.706 | 0.500 | 0.622 | 0.715 | 0.844 | 0.697 | 0.600 | 0.797 | 0.879 |
| 17 | 0.568 | 0.659 | 0.676 | 0.814 | 0.484 | 0.611 | 0.772 | 0.836 | 0.500 | 0.506 | 0.642 | 0.878 |
| 18 | 0.525 | 0.504 | 0.626 | 0.680 | 0.507 | 0.681 | 0.773 | 0.843 | 0.426 | 0.644 | 0.772 | 0.857 |
| *A* | 0.563 | 0.574 | 0.623 | 0.677 | 0.625 | 0.648 | 0.760 | 0.847 | 0.589 | 0.631 | 0.722 | 0.853 |
| *S* | 0.078 | 0.044 | 0.057 | 0.055 | 0.107 | 0.066 | 0.076 | 0.063 | 0.097 | 0.098 | 0.072 | 0.075 |
| *R* | 13.8% | 7.6% | 9.2% | 8.1% | 17.2% | 10.2% | 10.0% | 7.5% | 16.5% | 15.5% | 10.0% | 8.8% |

A: average. S: standard deviation. R: relative deviation.

1. *Saccharomyces cerevisiae*, 2. *Caenorhabditis elegans*, 3. *Ciona savignyi*, 4. *Drosophila melanogaster*, 5. *mosquito*, 6. *medaka*, 7. *Gasterosteus aculeatus*, 8. *Tetraodon nigroviridis*, 9. *Danio rerio*, 10. *Gallus*, 11. *Rattus norvegicus*, 12. *Mus musculus*, 13. *opossum*, 14. *dog*, 15. *Bos taurus*, 16. *Pan troglodytes*, 17. *Macaca mulatta*, 18. *Homo sapiens*.

Table 3 shows that the values of average singularity spectrum span $\Delta\alpha$ of all species g increase distinctly with $k$, which indicates that the detailed fractal information is more abundant in the case of larger $k$ values. Also, the exon sequence has the least change of spectrum span average $\Delta\alpha$ with $k$ increasing and the least relative deviation of spectrum span $\Delta\alpha$ in species. Thus, it can be concluded that oligonucleotide repeat length polymorphism is more distinct in non-coding sequences than in coding sequences. In addition, the values of relative deviation of spectrum span $\Delta\alpha$ in different species decrease with $k$ increasing, especially when $k$ changes from 6 to 7 for exon and intron sequences and from 7 to 8 for blank sequences. In Table 2, the values of relative deviation of spectrum span $\Delta\alpha$ in different human component sequences are exon 6.8%, intron 7.0%, blank 8.6% ($k = 8$), while in Table 3, the values of relative deviation of $\Delta\alpha$ in different species reach 9.2%, 10%, 10% respectively. It agrees with the viewpoint in Ref. [3] that the difference of oligonucleotide distribution along a genome was smaller than that among genomes. Furthermore, Table 3 shows that the way the spectrum span varies with $k$ depends on what the species is. The spectrum span does not always increase monotonically with $k$ increasing, and then the inflexion emerges in different values of $k$ for different species. The oligonucleotide with length in which the inflexion emerges has the most uniform or non-uniform distribution. So the relation between singularity spectrum span and oligonucleotide length $k$ provides a useful method to study oligonucleotide repeat length polymorphism.

# 4. Conclusions

CGR is a very handy statistical approach of dealing with such a large quantity of genome data for its scale-independent characteristics. CGR images of DNA sequences show the fractal character, and multifractal analysis can provide statistical information on DNA sequences.

In the present paper, we perform a 2D WTMM-based multifractal analysis on CGR images. Firstly, we extract the edge images of CGR images; the $\ln Z_q(\varepsilon)$–$\ln\varepsilon$ curves show that the edge images hold the morphology character of CGR images, and have a wider scale-invariability range. Hence, the sequence structure feature can be analysed more accurately with the multifractal analysis of edge images.

Secondly, the multifractal spectra of exon, intron and blank sequences are calculated and compared with each other. The exon sequence has a left-hooked shape and the lowest singularity spectrum width. Since there exists a certain connection between the sequence activity and the distribution of singularity points with different singularity intensities, the exon sequence has the most uniform distribution of the oligonucleotides.

Finally, the dependence of singularity spectrum parameters on oligonucleotide length, sequence component and species is investigated, and the results show that the spectrum span $\Delta\alpha$ is a parameter sensitive to sequence component, oligonucleotide length and species. Thus it can be regarded as a handy tool for length polymorphism research of repeat oligonucleotides.

# References


[1] Jeffrey H 1990 *Nucl. Acids Res.* **18** 2163
[2] Patrick J Deschavanne, Alain Giron, Joseph Vilain, Guillaume Fagot and Bernard Fertil 1999 *Mol. Biol. Evol.* **16** 1391
[3] Almeida J, Carrico J, Maretzek A, Noble P A and Fletcher M 2001 *Bioinformatics* **17** 429
[4] Wang Y W, Hill K, Singb S and Kari L 2005 *Gene.* **346** 173
[5] Milan R 2008 *Chem. Phys. Lett.* **456** 84
[6] Jie F and Tianming W 2008 *Chem. Phys. Lett.* **454** 355
[7] Fiser A, Tusnady G E and Simon I 1994 *J. Mol. Graphics* **12** 302
[8] Yu Z G, Anh V and Lau K S 2004 *J. Theor. Biology* **226** 341
[9] Yang J Y, Peng Z L, Yu Z G, Zhang R J, Anh V and Wang D S 2009 *J. Theor. Biology* **257** 618
[10] Evertsz C J G and Mandelbrot B B 1992 *Multifractal Measures* In: Peitgen H, Juergens H and Saupe D (eds.) *Chaos and Fractals* (Berlin: Springer) p921
[11] Stanley H E and Meakin P 1998 *Nature* **335** 405
[12] Gutiérrez J, Rodríguez M and Abramson G 2001 *Physica* A **300** 271
[13] Yu Z G, Anh V and Lau K S 2001 *Physica* A **301** 351
[14] Yu Z G, Anh V and Lau K S 2001 *Phys. Rev.* E **64** 031903
[15] Yu Z G, Anh V and Lau K S 2003 *Phys. Rev.* E **68** 021913
[16] Fu W J, Wang Y Y and Lu D R 2005 *The 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **17-18** 4783
[17] Fu W J, Wang Y Y and Lu D R 2007 *J. Biomed. Engineer.* **24** 522 (in Chinese)
[18] Muzy J F, Bacry E and Arnéodo A 1991 *Phys. Rev. Lett.* **67** 3515
[19] Arnéodo A, Decoster N and Roux S G 2000 *Europ. Phys. J.* B **15** 567
[20] Camilo Rodrigues N, Kevin B and Adrienn C 2004 *Physica* A **344** 5806
[21] Mallat S and Zhong S 1992 *IEEE Trans. Pattern Anal. Mach. Intell.* **14** 710