Iterated function system and multifractal analysis of biological sequences[*]

Zu-Guo Yu[1,2][†], Vo Anh[1] and Ka-Sing Lau[3]
[1]Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia.
[2]Department of Mathematics, Xiangtan University, Hunan 411105, P. R. China.
[3]Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong

Abstract

The fractal method has been successfully used to study many problems in Physics, Mathematics, Engineering, Finance, even in Biology till now. In the past decade or so there has been a ground swell of interest in unravelling the mysteries of DNA. How to get more bioinformations from these DNA sequences is a challenging problem. The problem of classification and evolution relationship of organisms are the central problems in Bioinformatics. And it is also very hard to predict the secondary and space structure of a protein from its amino acid sequence. In this paper, some recent results related these problems obtained through multifractal analysis and iterated function system (IFS) model are introduced.

Key words: Measure representation, Multifractal analysis, IFS (RIFS) model, complete genome, length sequence, protein.

1 Introduction

The concept of "fractal" was proposed by Benoit Mandelbrot [1] in the later of 1970s. Fractal geometry provides a mathematical formalism for describing complex spatial and dynamical structures [1, 2] (e.g. the strange attractor of a chaotic dynamical system is usually a fractal). The fractal method has been successfully used to study many problems in Physics, Mathematics, Engineering, and Biology. Multifractal analysis was initially proposed to treat turbulence data. This kind of analysis is a useful way to characterise the spatial inhomogeneity of both theoretical and experimental fractal patterns [3] and play an important role in the fractal theory. In recent years it has been applied successfully in many different fields including time series analysis and financial modelling.

In the past decade or so there has been a ground swell of interest in unravelling the mysteries of DNA. The heredity information of most organisms is encoded in a universal way in long chains of nucleic acids formed by four different nucleotides, namely adenine (a), cytosine (a), guanine (g) and thymine (t). The DNA sequence identifies a given species, distinguishing it from all other species, even those with the same nucleotide composition. A large number of these DNA sequences is widely available in recent times. One of the challenges of DNA sequence analysis is to determine

[*]The research was partially supported by QUT's Postdoctoral Research Support Grant No. 9900658 and the Youth Foundation of Chinese National Natural Science Foundation (No. 10101022), the RGC Earmarked Grant CUHK 4215/99P.

[†]Email: yuzg@hotmail.com,z.yu@qut.edu.au

the patterns in these sequences. It is useful to distinguish coding from noncoding sequences. Problems related to the classification and evolution of organisms are also important. A significant contribution in these studies is to investigate the long-range correlation in DNA sequences [4-19]. Berthelsen *et al.* [20] considered the global fractal dimensions of human DNA sequences treated as pseudorandom walks.

Since the first complete genome of the free-living bacterium *Mycoplasma genitalium* was sequenced in 1995 [21], an ever-growing number of complete genomes has been deposited in public databases. The availability of complete genomes induces the possibility to establish some global properties of these sequences. A time series model was proposed by Yu *et al.* [22, 23, 24] to study the correlation property of coding segments and length sequences of complete genome.

The global and visual methods can amplify the difference between a DNA sequence and a random sequence as well as to distinguish DNA sequences themselves in more details [25]. For this purpose, after the famous chaos game representation of DNA sequences proposed by Jeffrey *et al* [26, 27], Hao *et al.* [25] proposed a visualisation method based on counting and coarse-graining the frequency of appearance of substrings with a given length. They called it the *portrait* of an organism. They found that there exist some fractal patterns in the portraits which are induced by avoiding and under-represented strings. The fractal dimension of the limit set of portraits was also discussed [28, 29]. The connection between the Hao's scheme and the chaos game representation is established through the multifractal property [30]. In [31], Yu *et al.* proposed the measure representation of complete genomes followed by the multifractal analysis. The multifractal analysis of the length sequences based on the complete genome was performed in [32].

A protein is composed of one or more chains that are covalently joined. The chain of amino acids are called *polypeptides*. Twenty different kinds of amino acids are found in proteins. The three-dimensional structure of proteins is a complex physical and mathematical problem of prime importance in molecular biology, medicine, and pharmacology [33, 34]. The central dogma motivating structure prediction is that: 'the three dimensional structure of a protein is determined by its amino acid sequence and its environment without the obligatory role of extrinsic factors' [35, 36]. Once an amino acid sequence is known, the number of possible space structures it can fold to is enormous. How to predict the high level structures (secondary and space structures) from the amino acid sequence is a challenge problem in science, in particular to the large proteins. A number of coarse-grained models have been proposed to provide insight to these very complicated issues [36]. A well known model in this class is the HP model proposed by Dill *et al.* [37]. In this model 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). In last decade the HP model has been extensively studied by several groups (e.g. [34, 38, 39]). After studying the model on lattices, Li *et al.* [38] found there are small number of structures with exceptionally high designability which a large number of protein sequences possess as their ground states. These highly designable structures are found to have protein-like secondary structures [34, 38, 40]. But the HP model may be too simple and lacks enough consideration on the heterogeneity and the complexity of the natural set of residues [41]. According to Brown [42], in the HP model, one can divide the polar class into three classes: positive polar, uncharged polar and negative polar. So 20 different kinds of amino acids can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. In this model, one gets more details than in the HP model. We call this model a *detailed HP model.* In this paper we will adopt the detailed HP model.

The fractal method has been used to study the protein backbone [43], the accessible surface of protein [43, 44, 45, 46] and protein potential energy landscapes [47]. The multifractal analysis of solvent accessibilities in proteins was done by Balafas and Dewey [48]. In [48], the model used to fit the multifractal spectrum was also discussed. But the parameters derived in their multifractal analysis cannot be used to predict the structural classification of a protein from its amino acid sequence.

The amino acid sequence of a protein is also called a *protein sequence* in this paper. Based the idea of DNA walk model and different mapping, a decoded walk model was proposed to study the correlation property of protein sequences by Pande *et al.* [49] using "Bridge analysis" and Straint and Dewey [50] using multifractal analysis. Deviations of the decoded walk from random behaviour provides evidence of memory.

Inspired by the idea of measure representation of DNA sequence [31], we also proposed a visual representation — measure representation of protein sequences based on the detailed HP model [51].

To our knowledge [52], it is much harder to simulate a measure than to fit its multifractal spectrum (because different measures may have the same multifractal spectrum). The iterated function systems (IFS) model proposed by Barnsley and Demko [53] is a powerful tool in fractal theory (many fractals such as the Cantor set can be generated by the IFS model). We found that the recurrent IFS (RIFS) model can be used to simulate the measure representation of complete genomes while the IFS model can be used to simulate the measure representation of protein sequences. In this paper, the estimated parameters in RIFS or IFS model are used to discuss the classification and evolutionary tree of living organisms and the structural classification of large proteins.

2 Measure representation of complete genomes

We call any string made of K letters from the set $\{g, c, a, t\}$ a K-string. For a given K there are in total 4^K different K-strings. In order to count the number of each kind of K-strings in a given DNA sequence 4^K counters are needed. We divide the interval $[0, 1[$ into 4^K disjoint subintervals, and use each subinterval to represent a counter. Letting $s = s_1 \cdots s_K, s_i \in \{a, c, g, t\}, i = 1, \cdots, K$, be a substring with length K, we define

$$x_{left}(s) = \sum_{i=1}^{K} \frac{x_i}{4^i}, \tag{1}$$

where

$$x_i = \begin{cases} 0, & \text{if } s_i = a, \\ 1, & \text{if } s_i = c, \\ 2, & \text{if } s_i = g, \\ 3, & \text{if } s_i = t, \end{cases} \tag{2}$$

and

$$x_{right}(s) = x_{left}(s) + \frac{1}{4^K}. \tag{3}$$

We then use the subinterval $[x_{left}(s), x_{right}(s)[$ to represent substring s. Let $N_K(s)$ be the number of times that substring s with length K appears in the complete genome. If the total number of K-strings appeared in the complete genome is denoted as $N_K(total)$, we define

$$F_K(s) = N_K(s)/N_K(total) \tag{4}$$

to be the frequency of substring s. It follows that $\sum_{\{s\}} F_K(s) = 1$. Now we can define a measure μ_K on $[0, 1[$ by $d\mu_K(x) = Y(x)dx$, where

$$Y_K(x) = 4^K F_K(s), \quad \text{when} \quad x \in [x_{left}(s), x_{right}(s)[. \tag{5}$$

It is easy to see $\int_0^1 d\mu_K(x) = 1$ and $\mu_K([x_{left}(s), x_{right}(s)[) = F_K(s)$. We call μ_K the *measure representation* of the organism corresponding to the given K.

For simplicity of notation, the index K is dropped in $F_K(s)$, etc., from now on, where its meaning is clear.

Remark: The ordering of a, c, g, t in (2) will give the natural dictionary ordering of K-strings in the one-dimensional space. A different ordering of K-strings would change the nature of the correlations. When we want to compare different organisms using the measure representation, once the ordering of a, c, g, t in (2) is given, it is fixed for all organisms.

3 Length sequences based on the complete genomes

For the importance of the numbers, sizes and ordering of genes along the chromosome, one can refer to Part 5 of Lewin [54]. Here one may ignore the composition of the four kinds of bases in coding and noncoding segments and only considers the rough structure of the complete genome or long DNA sequences. Provata and Almirantis [55] proposed a fractal Cantor pattern of DNA. They map coding segments to filled regions and noncoding segments to empty regions of random Cantor set and then calculate the fractal dimension of the random fractal set. They found that the coding/noncoding partition in DNA sequences of lower organisms is homogeneous-like, while in the higher eucariotes the partition is fractal. This result seems too rough to distinguish bacteria because the fractal dimensions of bacteria they gave out are all the same.

Viewing from the level of structure, the complete genome of an organism is made up of coding and noncoding segments. Here the length of a coding/noncoding segment means the number of its bases. Based on the lengths of coding/noncoding segments in the complete genome, one can get two kinds of integer sequences by the following ways:

i) Order all lengths of coding segments according to the order of coding segments in the complete genome. This integer sequence is named *coding length sequence.*

ii) Order all lengths of noncoding segments according to the order of noncoding segments in the complete genome. This integer sequence is named *noncoding length sequence.*

Let $T_t,\ t = 1, 2, \cdots, N$, be the length sequence of coding or noncoding segments in the complete genome of an organism. First we define

$$F_t = T_t / (\sum_{j=1}^{N} T_j) \tag{6}$$

to be the frequency of T_t. It follows that $\sum_t F_t = 1$. Now we can define a measure μ on $[0, 1[$ by $d\mu(x) = Y(x)dx$, where

$$Y(x) = N \times F_t, \quad \text{when} \quad x \in [\frac{t-1}{N}, \frac{t}{N}[. \tag{7}$$

It is easy to see that $\int_0^1 d\mu(x) = 1$ and $\mu([(t-1)/N, t/N[) = F_t$.

4 Detailed HP model and measure representation of protein sequences

Twenty different kinds of amino acids are found in proteins. In the detailed HP model they can be divided in to four classes: non-polar, negative polar, uncharged polar and positive polar. The eight residues designating the non-polar class are: ALA, ILE, LEU, MET, PHE, PRO, TRP, VAL; the two residues designating the negative polar class are: ASP, GLU; the seven residues designating the uncharged polar class are: ASN, CYS, GLN, GLY, SER, THR, TYR; and the remaining three residues: ARG, HIS, LYS designate the positive polar class.

For a given protein sequence with length L, $s = s_1 \cdots s_L$, where s_i is one of the twenty kinds of amino acids for $i = 1, \cdots, L$, we define

$$a_i = \begin{cases} 0, & \text{if } s_i \text{ is non-polar,} \\ 1, & \text{if } s_i \text{ is negative polar,} \\ 2, & \text{if } s_i \text{ is uncharged polar,} \\ 3, & \text{if } s_i \text{ is positive polar.} \end{cases} \tag{8}$$

So we can obtain a sequence $X(s) = a_1 \cdots a_L$, where a_i is a letter in the alphabet $\{0, 1, 2, 3\}$. Using the same idea as in Section 2, we can define the measure representation μ_K of K-strings of the given protein sequence.

5 Multifractal analysis

The most common numerical implementations of multifractal analysis are the so-called *fixed-size box-counting algorithms* [56]. In the one-dimensional case, for a given measure μ with support $E \subset \mathbf{R}$, we consider the *partition sum*

$$Z_\epsilon(q) = \sum_{\mu(B) \neq 0} [\mu(B)]^q, \tag{9}$$

$q \in \mathbf{R}$, where the sum runs over all different nonempty boxes B of a given side ϵ in a grid covering of the support E, that is,

$$B = [k\epsilon, (k+1)\epsilon[. \tag{10}$$

The scaling exponent $\tau(q)$ is defined by

$$\tau(q) = \lim_{\epsilon \to 0} \frac{\log Z_\epsilon(q)}{\log \epsilon} \tag{11}$$

and the generalized fractal dimensions of the measure are defined as

$$D_q = \tau(q)/(q-1), \quad \text{for } q \neq 1, \tag{12}$$

and

$$D_q = \lim_{\epsilon \to 0} \frac{Z_{1,\epsilon}}{\log \epsilon}, \quad \text{for } q = 1, \tag{13}$$

where $Z_{1,\epsilon} = \sum_{\mu(B) \neq 0} \mu(B) \log \mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of

$$\frac{1}{q-1} \log Z_\epsilon(q)$$

against $\log \epsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\epsilon}$ against $\log \epsilon$ for $q = 1$. D_1 is called the *information dimension* and D_2 the *correlation dimension.* The D_q of the positive values of q give relevance to the regions where the measure is large, i.e., to the coding or noncoding segments which are relatively long. The D_q of the negative values of q deal with the structure and the properties of the most rarefied regions of the measure, i.e. to the segments which are relatively short.

By following the thermodynamic formulation of multifractal measures, Canessa [57] derived an expression for the 'analogous' specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \tag{14}$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. In the following we calculate the 'analogous' specific heat of coding and noncoding length sequences for the first time. The types of phase transitions are helpful to discuss the classification of bacteria.

6 IFS (RIFS) model and moment method for the inverse problem

6.1 IFS (RIFS) model

In order to simulate the measure representation of the complete genome, Anh *et al.* [58] proposed the *iterated function systems* (IFS) model and the recurrent IFS model. IFS is the name given by Barnsley and Demko [53] originally to a system of contractive maps $w = \{w_1, w_2, \cdots, w_N\}$. Let E_0 be a compact set in a compact metric space, $E_{\sigma_1\sigma_2\cdots\sigma_n} = w_{\sigma_1} \circ w_{\sigma_2} \circ \cdots \circ w_{\sigma_n}(E_0)$ and

$$E_n = \cup_{\sigma_1,\cdots,\sigma_n \in \{1,2,\cdots,N\}} E_{\sigma_1\sigma_2\cdots\sigma_n}.$$

Then $E = \cap_{n=1}^{\infty} E_n$ is called the *attractor* of the IFS. The attractor is usually a fractal and the IFS is a relatively general model to generate many well-known fractal sets such as the Cantor set and the Koch curve. Given a set of probabilities $p_i > 0$, $\sum_{i=1}^{N} p_i = 1$, pick an $x_0 \in E$ and define the iteration sequence

$$x_{n+1} = w_{\sigma_n}(x_n), \qquad n = 0, 1, 2, 3, \cdots, \tag{15}$$

where the indices σ_n are chosen randomly and independently from the set $\{1, 2, \cdots, N\}$ with probabilities $P(\sigma_n = i) = p_i$. Then every orbit $\{x_n\}$ is dense in the attractor E [53, 59]. For n large enough, we can view the orbit $\{x_0, x_1, \cdots, x_n\}$ as an approximation of E. This process is called *chaos game.*

Given a system of contractive maps $w = \{w_1, w_2, \cdots, w_N\}$ on a compact metric space E^*, we associate with these maps a matrix of probabilities $\mathbf{P} = (p_{ij})$ which is row stochastic, i.e. $\sum_j p_{ij} = 1$, $i = 1, 2, \cdots, N$. Consider a random chaos game sequence generated by

$$x_{n+1} = w_{\sigma_n}(x_n), \quad n = 0, 1, 2, 3, \cdots,$$

where x_0 is any starting point. The fundamental difference between this process and the usual chaos game Eq. (15) is that the indices σ_n are not chosen independently, but rather with a probability that depends on the previous index σ_{n-1}:

$$P(\sigma_{n+1} = i) = p_{\sigma_n, i}$$

Then $(E^*, w, \mathbf{P})$ is called a *recurrent IFS* (RIFS).

Let μ be the invariant measure on the attractor E of an IFS or RIFS, χ_B the characteristic function for the Borel subset $B \subset E$, then from the ergodic theorem for IFS or RIFS [53],

$$\mu(B) = \lim_{n\to\infty}[\frac{1}{n+1}\sum_{k=0}^{n} \chi_b(x_k)].$$

In other words, μ_B is the relative visitation frequency of B during the chaos game. A histogram approximation of the invariant measure may then be obtained by counting the number of visits made to each pixel on the computer screen.

6.2 Moment method to estimate the parameters of the IFS (RIFS) model

The coefficients in the contractive maps and the probabilities in the IFS or RIFS model are the parameters to be estimated for a real measure which we want to simulate. Vrscay [59] introduced a moment method to perform this task. If μ is the invariant measure and E the attractor of IFS or RIFS in $\mathbf{R}$, the moments of μ are

$$g_i = \int_E x^i d\mu, \qquad g_0 = \int_E d\mu = 1. \tag{16}$$

If $w_i(x) = c_i x + d_i,\ i = 1, \cdots, N$, then the following well-known recursion relations hold for the IFS model:

$$[1 - \sum_{i=1}^{N} p_i c_i^n] g_n = \sum_{j=1}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) g_{n-j} (\sum_{i=1}^{N} p_i c_i^{n-j} d_i^j). \tag{17}$$

Thus, setting $g_0 = 1$, the moments $g_n,\ n \geq 1$, may be computed recursively from a knowledge of $g_0, \cdots, g_{n-1}$ [59].

For the RIFS model, we have

$$g_n = \sum_{j=1}^{N} g_n^{(j)}, \tag{18}$$

where $g_n^{(j)},\ j = 1, \cdots, N$, are given by the solution of the following system of linear equations:

$$\sum_{j=1}^{N} (p_{ji} c_i^n - \delta_{ij}) g_n^{(j)} = - \sum_{k=0}^{n-1} \left(\begin{array}{c} n \\ k \end{array} \right) [\sum_{j=1}^{N} c_i^k d_i^{n-k} p_{ji} g_k^{(j)}],\ i = 1, \cdots, N,\ n \geq 1. \tag{19}$$

For $n = 0$, we set $g_0^{(i)} = m_i$, where m_i are given by the solution of the linear equations

$$\sum_{j=1}^{N} p_{ji} m_j = m_i, \quad i = 1, 2, \cdots, N, \quad \text{and } g_0 = \sum_{i=1}^{N} m_i = 1. \tag{20}$$

If we denote by G_k the moments obtained directly from the real measure using (16), and g_k the formal expression of moments obtained from (17) for IFS model and from (18-20) for RIFS model, then through solving the optimal problem

$$\min_{c_i, d_i, p_i \text{ or } p_{ij}} \sum_{k=1}^{n} (g_k - G_k)^2, \qquad \text{for some chosen } n, \tag{21}$$

we can obtain the estimated values of the parameters in the IFS or RIFS model.

From the measure representation of a complete genome or protein sequence, we see that it is natural to choose $N = 4$ and

$$w_1(x) = x/4,\ w_2(x) = x/4 + 1/4,\ w_3(x) = x/4 + 1/2,\ w_4(x) = x/4 + 3/4$$

in the IFS or RIFS model. For a given measure representation of a complete genome or protein sequence, we obtain the estimated values of the probabilities p_1, p_2, p_3, p_4 in IFS model or the matrix of probabilities $\mathbf{P} = (p_{ij})$ by solving the optimisation problem (21). Based on the estimated values of the probabilities, we can use the chaos game to generate a histogram approximation of the invariant measure of IFS or RIFS which we can compare with the real measure representation of the complete genome or protein sequence.

7 Applications to the biological sequence analysis

Till now more than 50 complete genomes of Archaea and Eubacteria are available in public databases (for example in Genbank at web site ftp://ncbi.nlm.nih.gov/genbank/genomes/ or in KEGG at web site http://www.genome.ad.jp/kegg/java/org_list.html).

In [31], the multifractal analysis were performed on the measure representations of a large number of complete genomes. For examples, the D_q and C_q curves of some organisms are shown in Figures 1 and 2 respectively. From the measure representations and the values of the D_q spectra and related C_q curves, it was concluded that these complete genomes are not random sequences. For substrings with length $K = 8$, the D_q spectra of all organisms studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. With the decreasing value of K, the multifractality lessens. The C_q curves of all bacteria resemble a classical phase transition at a critical point. But the 'analogous' phase transitions of chromosomes of non-bacteria organisms are different. Apart from Chromosome 1 of *C. elegans*, they exhibit the shape of double-peaked specific heat function.

Yu and Anh [23] proposed a time series model for the length sequences of DNA. After calculating the correlation dimensions and Hurst exponents, it was found that one can get more information from this model than that of fractal Cantor pattern [55]. The quantification of these correlations could give an insight into the role of the ordering of genes on the chromosome. Through detrended fluctuation analysis (DFA) [60] and spectral analysis, the long-range correlation was found in these length sequences [24].

The multifractal analysis was also performed on the coding and noncoding length sequences constructed from a complete genome [32]. From the shape of the D_q and C_q curves of length sequences, it was seen that there exists a clear difference between the coding/noncoding length sequences of all organisms considered and a completely random sequence. The complexity of noncoding length sequences is higher than that of coding length sequences for bacteria. Almost all D_q curves for coding length sequences are flat, so their multifractality is small whereas almost all D_q curves for noncoding length sequences are multifractal-like. It is seen that the 'analogous' specific heats of noncoding length sequences of bacteria have a rich variety of behaviour which is much more complex than that of coding length sequences.

In [58], we simulated the measure representations of the complete genomes of many organisms using the IFS and RIFS models. We found that RIFS is a good model to simulate the measure representation of complete genome of organisms. For example, the histogram of substrings in the genome of *Buchnera sp. APS* for $K = 8$ is given in the left figure of Figure 3. Self-similarity is apparent in the measure. The histogram approximation of the generated measure of *Buchnera sp. APS* using the RIFS model is shown in the right figure of Figure 3. It is seen that the RIFS simulation traces very well the original measure representation of the complete genome.

Once the matrix of probabilities is determined, the RIFS model is obtained. Hence the matrix of probabilities obtained from the RIFS model can be used to represent the measure of the complete genome of an organism. Different organisms can be compared using their matrix of probabilities obtained from the RIFS model. If $\mathbf{P} = (p_{ij})$, $\mathbf{P'} = (p'_{ij})$, $i, j = 1, 2, 3, 4$, are the matrices of probabilities of two different organisms obtained from the RIFS model for a fixed K, we propose to define the distance between the two organisms as

$$Dist = \sqrt{\sum_{i,j=1}^{4} (p_{ij} - p'_{ij})^2}. \qquad (22)$$

The genetic distance defined between two organisms is based only on the parameters derived from the fractal model, so that we can avoid artefacts associated with sequence alignment. The similarity based on the fractal model of the complete genome is global. In [61], we have done the phylogenetic analysis of more than 50 genomes using the above definition of distance. The phylogenetic tree is shown in Figure 5. The results from our phylogenetic analysis indicate that lateral gene transfer [62] must have been very common in the early history of life and thus constitutes a major source of variations in a substantial proportion of prokaryotic genome.

It is well known that all statistical method and nonlinear scale method require enough data samples. The methods introduced in the previous sections can only be used to analyse long protein sequences (corresponding to large proteins). The amino acid sequence of 32 large proteins are selected from RCSB Protein Data Bank (PDB) (http://www.rcsb.org/pdb/index.html). These 32 proteins belong to five structure classes [63] according to their secondary structures: α, β, $\alpha+\beta$ (α,β alternate), α/β (α, β segregate) and others (no α and no β) proteins. First we convert the amino acid sequences of these proteins to their measure representations with $K = 5$ according to the method introduced in Section 3. If K is too small, there are not enough combinations of letters from the set $\{0, 1, 2, 3\}$, therefore there is no statistical sense. And if K is too big, the frequencies of most substrings are zero. So we cannot obtain any biological information from the measure representation. Considering the length of the selected proteins which ranges from 350 to 1000, we think it is suitable to choose $K = 5$.

In [51], we found the IFS model is better than the RIFS model to simulate the measure representation of protein sequences. The estimated parameters in the IFS model of 32 proteins are given in Table 1. For example, we show the histograms of measure representation and simulated measures of protein *P.69 Pertactin* (PDB ID: 1DAB) in Figure 4. From Figure 4, one can see that the difference between measure representation and IFS simulated measure is very small. Once the probabilities are determined, the IFS model is obtained. Hence the probabilities obtained from the IFS model can be used to represent the measure representation of the protein sequence. From Table 1, we find the probability p_3 (which corresponding to the uncharged polar property) can be used to distinguish the structural class of proteins from α class and β class (values of p_3 of proteins in class α are less than those of proteins in class β), and the probability p_1 (which corresponds to the non-polar property) can be used to distinguish the structural class of proteins from class $\alpha+\beta$ and class α/β (values of p_1 of proteins in class α/β are less than those of proteins in class $\alpha+\beta$). Hence we believe that the non-polar residues and uncharged residues play a more important role than other kinds of residues in the protein folding process. This information is useful for the prediction of protein structure.

References

[1]

[1] B.B. Mandelbrot, *The Fractal Geometry of Nature*, Academic, New York, 1983.

[2] J. Feder, *Fractals*, Plenum, New York, 1988.

[3] P. Grassberger and I. Procaccia, *Phys. Rev. lett.* **50** (1983) 346.

[4] W. Li and K. Kaneko, *Europhys. Lett.* **17** (1992) 655; W. Li, T. Marr, and K. Kaneko, *Physica D* **75** (1994) 392.

[5] C.K. Peng, S. Buldyrev, A.L.Goldberg, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, *Nature* **356** (1992) 168.

[6] J. Maddox, *Nature* **358** (1992) 103.

[7] S. Nee, *Nature* **357** (1992) 450.

[8] C.A. Chatzidimitriou-Dreismann and D. Larhammar, *Nature* **361** (1993) 212.

[9] V.V. Prabhu and J. M. Claverie, *Nature* **359** (1992) 782.

[10] S. Karlin and V. Brendel, *Science* **259** (1993) 677.

[11] (a) R. Voss, *Phys. Rev. Lett.* **68** (1992) 3805; (b) *Fractals* **2** (1994) 1.

[12] H.E. Stanley, S.V. Buldyrev, A.L. Goldberg, Z.D. Goldberg, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.K. Peng, and M. Simons, *Physica A* **205** (1994) 214.

[13] H.Herzel, W. Ebeling, and A.O. Schmitt, *Phys. Rev. E* **50** (1994) 5061.

[14] P. Allegrini, M. Barbi, P. Grigolini, and B.J. West, *Phys. Rev. E* **52** (1995) 5281.

[15] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M, Simons, and H. E. Stanley, *Phys. Rev. E* **51(5)** (1995) 5084-5091.

[16] A. Arneodo, E. Bacry, P.V. Graves, and J. F. Muzy, *Phys. Rev. Lett.* **74** (1995) 3293.

[17] A. K. Mohanty and A.V.S.S. Narayana Rao, *Phys. Rev. Lett.* **84(8)** (2000) 1832-1835.

[18] L. Luo, W. Lee, L. Jia, F. Ji and L. Tsai, *Phys. Rev. E* **58(1)** (1998) 861-871.

[19] Z. G. Yu and G. Y. Chen, Rescaled range and transition matrix analysis of DNA sequences. *Comm. Theor. Phys.* **33(4)** (2000) 673-678.

[20] C. L. Berthelsen, J. A. Glazier and M. H. Skolnick, *Phys. Rev. A* **45(12)** (1992) 8902.

[21] C. M. Fraser *et al.*, The minimal gene complement of Mycoplasma genitalium, *Science*, **270** (1995) 397.

[22] Z. G. Yu and B. Wang, A time series model of CDS sequences on complete genome, *Chaos, Solitons and Fractals* **12(3)** (2001) 519-526.

[23] Z. G. Yu and V. V. Anh, Time series model based on global structure of complete genome, *Chaos, Soliton and Fractals* **12(10)** (2001) 1827-1834.

[24] Z. G. Yu, V. V. Anh and Bin Wang, Correlation property of length sequences based on global structure of complete genome, *Phys. Rev. E* **63** (2001) 11903.

[25] B. L. Hao, H. C. Lee, and S. Y. Zhang, Fractals related to long DNA sequences and complete genomes, *Chaos, Solitons and Fractals*, **11(6)** (2000) 825-836.

[26] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Research* **18(8)** 1990) 2163-2170.

[27] N. Goldman, Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences, *Nucleic Acids Research* **21(10)** (1993) 2487-2491.

[28] Z. G. Yu, B. L. Hao, H. M. Xie and G. Y. Chen, Dimension of fractals related to language defined by tagged strings in complete genome. *Chaos, Solitons and Fractals* **11(14)** (2000) 2215-2222.

[29] B. L. Hao, H. M. Xie, Z. G. Yu and G. Y. Chen, Factorizable language: from dynamics to bacterial complete genomes. *Physica A* **288** (2001) 10-20.

[30] P. Tino, Multifractal properties of Hao's geometric representation of DNA sequences, *Physica A* **304** (2002) 480-494.

[31] Z. G. Yu, V. V. Anh and K. S. Lau, Measuere representation and multifractal analysis of complete genome, *Phys. Rev. E* **64** (2001) 031903.

[32] Z. G. Yu, V. V. Anh and K. S. Lau, Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome, *Physica A* **301(1-4)** (2001) 351-361.

[33] C. Chothia, *Nature* (London) **357** (1992) 543-544.

[34] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, and H.C. Lee, *Phys. Rev. Lett.* **84(2)** (2000) 386-389.

[35] C. Anfinsen, *Science* **181** (1973) 223.

[36] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, and H.C. Lee, Geometric and statistical properties of the mean-field HP model, the LS model and real protein sequences, Los Alamos National Laboratory E-Archive physics/0104015 (http://xxx.lanl.gov).

[37] K. A. Dill, *Biochemistry* **24** (1985) 1501; H.S. Chan and K.A. Dill, *Macromolecules* **22** (1989) 4559.

[38] H. Li, R. Helling, C. Tang, and N.S. Wingreen, *Science* **273** (1996) 666.

[39] B. Wang and Z. G. Yu, *J. Chem. Phys.* **112** (2000) 6084-6088.

[40] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, *Phys. Rev. Lett.* **80** (1998) 5683.

[41] J. Wang and W. Wang, *Phys. Rev. E* **61** (2000) 6981-6986.

[42] T.A. Brown, *Genetics* (3rd Edition), CHAPMAN & Hill, London, 1998.

[43] T.G. Dewey, *J. Chem. Phys.* **98** (1993) 2250.

[44] P. Pfiefer, U. Welz, and H. Wipperman, *Chem. Phys. Lett.* **113** (1985) 535.

[45] B.A. Fedorov, B.B. Fedorov, and P.W. Schmidt, *J. Chem. Phys.* **99** (1993) 4076.

[46] M. Lewis and D.C. Rees, *Science* **230** (1985) 1163.

[47] D.A. Lidar, D. Thirumalai, R. Elber, and R.B. Gerber, *Phys. Rev. E* **59** (1999) 2231.

[48] J.S. Balafas and T.G. Dewey, *Phys. Rev. E.* **52** (1995) 880.

[49] V.S. Pande, A. Y. Grosberg, and T. Tanaka, *Proc. Natl. Acad. Sci. USA* **91** (1994) 12972.

[50] B.J. Strait and T.G. Dewey, *Phys. Rev. E.* **52** (1995) 6588.

[51] Z. G. Yu, V. V. Anh and K. S. Lau, Fractal analysis of measure representation of large proteins based on the detailed HP model, 2002, Preprint.

[52] V. V. Anh, K. S. Lau and Z. G. Yu, *J. Phys. A: Math. Gene.* **34** (2001) 7127-7139.

[53] M.F. Barnsley and S. Demko, *Proc. Roy. Soc. London A* **399** (1985) 243.

[54] B. Lewin, *Genes VI*, Oxford University Press, 1997.

[55] A. Provata and Y. Almirantis, Fractal Cantor patterns in the sequence structure of DNA. *Fractals* **8(1)** (2000) 15-27.

[56] T. Halsy, M. Jensen, L. Kadanoff, I. Procaccia, and B. Schraiman, *Phys. Rev. A* **33** (1986) 1141.

[57] E. Canessa, *J. Phys. A: Math. Gen.* **33** (2000) 3637.

[58] V.V. Anh, K.S. Lau and Z.G. Yu, Representation of complete genomes of bacteria using recurrent IFS model, 2001, Preprint.

[59] E. R. Vrscay, in *Fractal Geometry and analysis*, Eds, J. Belair, NATO ASI series, Kluwer Academic Publishers, 1991.

[60] A. L. Goldberger, C. K. Peng, J. Hausdorff, J. Mietus,S. Havlin and H. E. Stanley, Fractals and the Heart, in *Fractal Geometry in Biological Systems*, Edited by P. M. Iannaccone and M. Khokha, CRC Press, Inc, 1996, Pages 249-266.

[61] Z. G. Yu, V. V. Anh, K. S. Lau and K. H. Chu, Phylogenetic analysis of living organisms based on a fractal model of complete genomes, 2002, Preprint.

[62] W.F. Doolittle, *Science* **284** (1999) 2124-2128.

[63] R. B. Russell, in *Protein structure prediction: Methods and Protocls*, Eds, D. Webster, Humana Press Inc., Totowa, NJ, 2000.

Table 1: The estimated prameters in the IFS model of all 32 proteins selected.

Class	PDB ID	p_1	p_2	p_3	p_4
α	1AVC	0.433053	0.057476	0.360621	0.148850
	1B89	0.434701	0.090537	0.355757	0.119005
	1BJ5	0.395675	0.171289	0.263892	0.169145
	1HO8	0.425220	0.116664	0.324997	0.133119
	1IAL	0.454049	0.145905	0.279686	0.120360
	1QSA	0.429905	0.095604	0.366038	0.108453
	2BCT	0.479382	0.051937	0.343780	0.124902
	5EAS	0.438919	0.079522	0.386794	0.094765
β	1B9S	0.374272	0.055143	0.447158	0.123429
	1DAB	0.443784	0.082010	0.399380	0.074825
	1EUT	0.404940	0.086955	0.409295	0.098810
	1FNF	0.392416	0.124496	0.393389	0.089700
	1JX5	0.418789	0.121671	0.364252	0.095288
	1MAL	0.369149	0.074231	0.483407	0.073214
$\alpha + \beta$	1B90	0.412281	0.069013	0.413590	0.105117
	1BBU	0.408854	0.203032	0.238907	0.149207
	1BYT	0.419483	0.124814	0.313159	0.142543
	1CLC	0.411955	0.089417	0.393040	0.105588
	1E7U	0.407123	0.186941	0.242776	0.163161
α/β	1A8I	0.435450	0.100694	0.329504	0.134352
	1ACJ	0.437285	0.087811	0.359227	0.115677
	1AOV	0.378102	0.092808	0.390054	0.139036
	1BFD	0.503850	0.103505	0.303115	0.089530
	1CRL	0.445648	0.061138	0.432773	0.060441
Others	1DPI	0.434653	0.174507	0.229232	0.161609
	1EFG	0.463732	0.090136	0.318268	0.127863
	1EPS	0.455629	0.080760	0.366760	0.096850
	1F1O	0.438389	0.119861	0.290525	0.151225
	1KVP	0.409277	0.105865	0.364443	0.120415
	1PMD	0.384736	0.133984	0.386281	0.094999
	1TPT	0.462826	0.143851	0.272910	0.120413
	4ACE	0.437279	0.087855	0.359186	0.115681

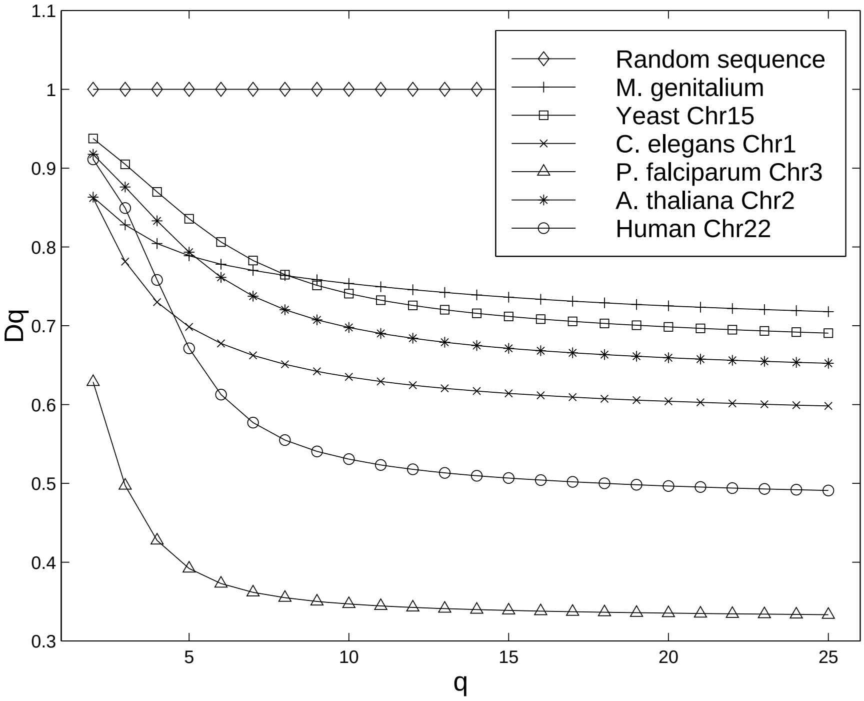

Figure 1: Dimension spectra of Chromosome 22 of Homo sapiens, Chromosome 2 of A. thaliana, Chromosome 3 of P. falciparum, Chromosome 1 of C. elegans, Chromosome 15 of S. cerevisiae and M. genitalium.

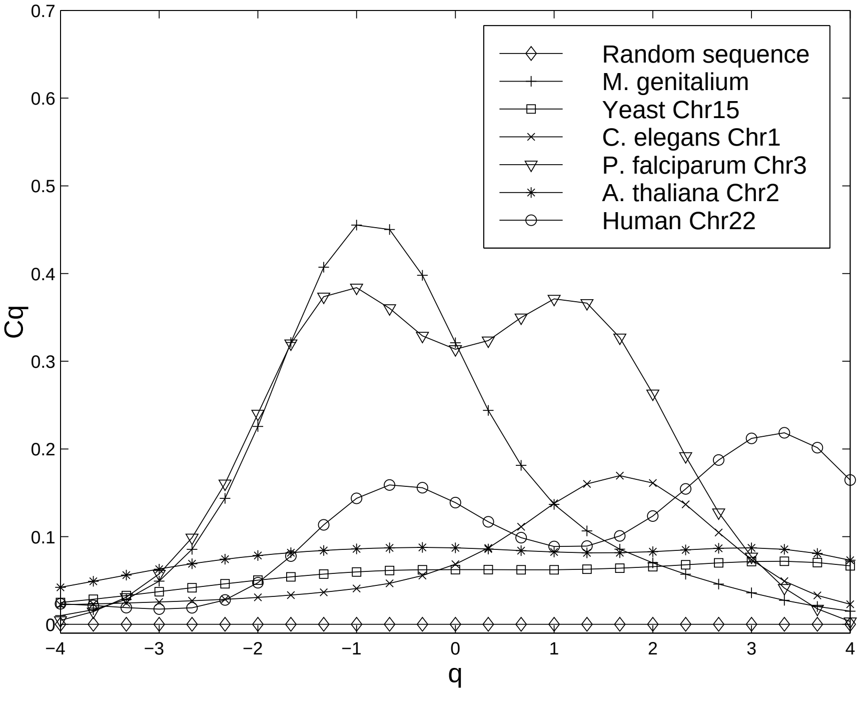

Figure 2: "Analogous" specific heat of Chromosome 22 of Homo sapiens, Chromosome 2 of A. thaliana, Chromosome 3 of P. falciparum, Chromosome 1 of C. elegans, Chromosome 15 of S. cerevisiae, M. genitalium and complete random sequence.

Figure 3: The measure representation (left) and the RIFS simulation (right) of the complete genome of Buchnera sp. APS when K=8.

Figure 4: The measure representation (left) and the IFS simulation (right) of protein *P.69 Pertactin* (PDB ID: 1DAB)

Figure 5: The phylogenetic tree of living organisms using the neighbour-joining method and the distance based on the parameters in the RIFS model.