Chaos game representation of functional protein sequences, and simulation and multifractal analysis of induced measures*

Yu Zu-Guo(喻祖国)[a)b)†], Xiao Qian-Jun(肖前军)[a)], Shi Long(石 龙)[a)], Yu Jun-Wu(余君武)[c)], and Vo Anh[b)]

[a)] *School of Mathematics and Computational Science, Xiangtan University, Xiangtan* 411105, *China*
[b)] *School of Mathematical Sciences, Queensland University of Technology, GPO Box* 2434, *Brisbane, Q* 4001, *Australia*
[c)] *Department of Mathematics and Computational Science, Hunan University of Science and Technology, Xiangtan* 411201, *China*

(Received 30 September 2009; revised manuscript received 20 November 2009)

Investigating the biological function of proteins is a key aspect of protein studies. Bioinformatic methods become important for studying the biological function of proteins. In this paper, we first give the chaos game representation (CGR) of randomly-linked functional protein sequences, then propose the use of the recurrent iterated function systems (RIFS) in fractal theory to simulate the measure based on their chaos game representations. This method helps to extract some features of functional protein sequences, and furthermore the biological functions of these proteins. Then multifractal analysis of the measures based on the CGRs of randomly-linked functional protein sequences are performed. We find that the CGRs have clear fractal patterns. The numerical results show that the RIFS can simulate the measure based on the CGR very well. The relative standard error and the estimated probability matrix in the RIFS do not depend on the order to link the functional protein sequences. The estimated probability matrices in the RIFS with different biological functions are evidently different. Hence the estimated probability matrices in the RIFS can be used to characterise the difference among linked functional protein sequences with different biological functions. From the values of the D_q curves, one sees that these functional protein sequences are not completely random. The D_q of all linked functional proteins studied are multifractal-like and sufficiently smooth for the C_q (analogous to specific heat) curves to be meaningful. Furthermore, the D_q curves of the measure μ based on their CGRs for different orders to link the functional protein sequences are almost identical if $q \geq 0$. Finally, the C_q curves of all linked functional proteins resemble a classical phase transition at a critical point.

Keywords: chaos game representation, recurrent iterated function systems, functional proteins, multifractal analysis

PACC: 8710, 4752

1. Introduction

Investigating the biological function of proteins is a key aspect of protein studies. Complete genomes provide us with an enormous amount of original information to unveil their biological functions. Almost half the biological functions of proteins encoded by genomes are unknown. For example, according to Ref. [1], about 41 percent (12809) of the gene products among the 26588 human proteins could not be classified and are termed proteins with unknown functions. Bioinformatic methods are important for studying the biological functions of proteins.[2] In this paper, the chaos game representation (CGR), the recurrent iterated function systems (RIFS) and multifractal analysis are used to analyse the features of functional protein sequences and further to study the biological functions of these proteins.

Jeffrey[3] first proposed a chaos game representation (CGR) of DNA sequences by using the four vertices of a square in a plane to represent the nucleotides a, c, g and t. The method produces a plot of a DNA sequence which displays both local and global patterns. Self-similarity or fractal structures were found in these plots. Some open questions from the biological point of view based on the CGRs were proposed.[3] Goldman[4] interpreted the CGRs in a biologically meaningful way and proposed a discrete time Markov

*Project partially supported by the National Natural Science Foundation of China (Grant No. 30570426), the Chinese Program for New Century Excellent Talents in University (Grant No. NCET-08-06867), Fok Ying Tung Education Foundation (Grant No. 101004), and Australian Research Council (Grant No. DP0559807).

†Corresponding author. E-mail: yuzg@hotmail.com

© 2010 Chinese Physical Society and IOP Publishing Ltd

chain model to simulate the CGRs of DNA sequences. Deschavanne[5] used CGRs of genomes to discuss the classification of species. Almeida[6] showed that the distribution of positions in the CGR plane is a generalisation of Markov chain probability tables that accommodates non-integer orders. Joseph and Sasikumar[7] proposed a fast algorithm for identifying all local alignments between two genome sequences using the sequence information contained in their CGRs. A CGR-walk model based on CGR coordinates for the DNA sequences[8] and for the protein sequences[9] were proposed recently.

The idea of CGR of DNA sequences proposed by Jeffrey[3] was generalized and applied for visualising and analysing protein databases by Fiser *et al.*[10] In the simplest case, the square in CGR of DNA is replaced by a 20-sided regular polygon (20-gon) for protein sequence representation. Fiser *et al.*[10] pointed out that the CGR can also be used to study three-dimensional (3D) structures of proteins. Basu *et al.*[11] (1998) proposed a new method for the CGR of different families of proteins. Using concatenated amino acid sequences of proteins belonging to a particular family and a 12-sided regular polygon, each vertex of which represents a group of amino acid residues leading to conservative substitutions, the method generates the CGR of the family and allows pictorial representation of the pattern characterizing the family. Basu *et al.*[11] found that the CGRs of different protein families exhibit distinct visually identifiable patterns. This implies that different functional classes of proteins produce specific statistical biases in the distributions of different mono-, di-, tri-, or higher order peptides along their primary sequences. In this paper we also use concatenated amino acid sequences of proteins with the same function.

Our group also proposed a CGR for protein sequences[12] which is based on the detailed HP model.[13] The HP model proposed by Dill *et al.*[14] is a well-known model of protein sequence analysis. In this model 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). But the HP model may be too simple and lacks sufficient information on the heterogeneity and the complexity of the natural set of residues.[15] According to Brown,[16] one can divide the polar class in the HP model into three subclasses: positive polar, uncharged polar and negative polar. So 20 different kinds of amino acid can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. In the detailed HP model, one considers more details than in the HP model. Based on the detailed HP model, we proposed a CGR for the linked protein sequences from the genomes.[12]

Nonlinear methods turn out to be a useful tool to study proteins. Huang and Xiao[17] made a detailed analysis of a set of typical protein sequences with a nonlinear prediction model in order to clarify their randomness. By using a modified recurrence plot, Huang *et al.*[18] showed that amino acid sequences of many multi-domain proteins had hidden repetitions. Fractal methods are important among the nonlinear methods and have been widely used in many fields such as oil pipeline[19] and surface roughness.[20] In particular, the fractal time series model was used to study the global structure[21] and CDSs[22] of the complete genome. More fractal methods for DNA sequence analysis were reviewed in Ref. [23].

RIFS in fractal theory[24,25] have been applied successfully to fractal image construction,[26] measure representation of genomes[27−30] and magnetic field data.[31,32] Yu *et al.*[33] proposed a CGR for the magnetic field data and used the two-dimensional RIFS model to simulate the CGR.

Multifractal analysis is a useful way to characterize the spatial heterogeneity of both theoretical and experimental fractal patterns.[34] A multifractal analysis based on the CGR of DNA sequences was given by Gutierrez *et al.*[35,36] Based on the measure representation of DNA sequences and the techniques of multifractal analysis, Anh *et al.*[27] discussed the problem of recognition of an organism from fragments of its complete genome. Yu *et al.*[37] used the parameters from the multifractal analysis for protein structure classification. Yang *et al.*[38] used two kinds of multifractal analyses based on the 6-letter model of amino acids to study the protein structure classification problem.

In this paper, we first give the CGR of randomly-linked functional protein sequences based on the detailed HP model, then propose to use the RIFS to simulate the measure based on their CGRs. Then multifractal analysis of the measures based on the CGR is performed. These methods can extract some features of functional protein sequences and furthermore help to understand the biological functions of these proteins.

2. Chaos game representation of linked functional protein sequences

We randomly concatenate the protein sequences with the same function one by one to obtain a long linked protein sequence. We call these sequences linked functional protein sequences. For these sequences, we outline here the way to gain their CGR from Ref. [12]. The protein sequence is formed by twenty different kinds of amino acid, namely Alanine (A), Arginine (R), Asparagine (N), Aspartic acid (D), Cysteine (C), Glutamic acid (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y) and Valine (V) (cf. page 109 of Ref. [16]). In the detailed HP model, they can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. The eight residues A, I, L, M, F, P, W, V designate the non-polar class; the two residues D, E designate the negative polar class; the seven residues N, C, Q, G, S, T, Y designate the uncharged polar class; and the remaining three residues R, H, K designate the positive polar class.

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

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

We then obtain a sequence $X(s) = a_1 \cdots a_l$, where a_i is a letter with subscript being one of the numbers in $\{0, 1, 2, 3\}$. We next define the CRG for a sequence $X(s)$ in a square $[0, 1] \times [0, 1]$, where the four vertices correspond to the four letters $0, 1, 2, 3$. The first point of the plot is placed half way between the centre of the square and the vertex corresponding to the first letter of the sequence $X(s)$; the i-th point of the plot is then placed half way between the $(i-1)$-th point and the vertex corresponding to the i-th letter. We then call the obtained plot the *CGR of the protein sequence s* based on the detailed HP model.

The CGRs of linked functional protein sequences produce clearer self-similar patterns. As an example, we show the CGR of the linked protein sequences whose biological function is the transporter in Fig. 1.

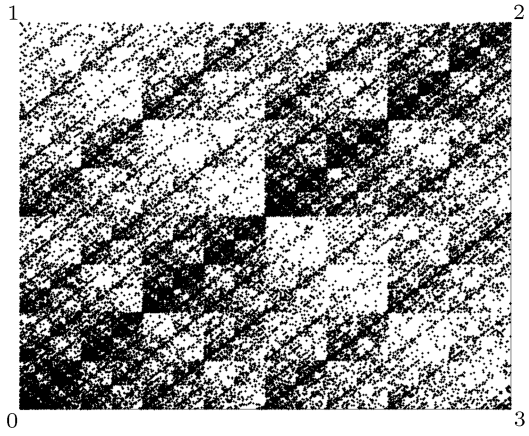


Fig. 1. Chaos game representation of the linked protein sequences whose biological function is *transporter* (with 423140 amino acids).

Considering the points in a CGR of linked functional protein sequence, we define a measure μ by $\mu(B) = \sharp(B)/N_l$, where $\sharp(B)$ is the number of points lying in a subset B of the CGR and N_l is the length of the sequence. We divide the square $[0, 1] \times [0, 1]$ into meshes of sizes 64×64, 128×128, 512×512 or 1024×1024. This results in a measure for each mesh. We then obtain a 64×64, 128×128, 512×512 or 1024×1024 matrix $\mathcal{A}$, where each element is the measure value on the corresponding mesh. We call $\mathcal{A}$ the *measure matrix* of the linked functional protein sequence. The measure μ based on a 128×128-mesh on the CGRs are considered in this paper. For example, the 128×128-mesh measure based on the CGR in Fig. 1 is shown in Fig. 2. Then we propose to use RIFS introduced in next section to simulate these measures.

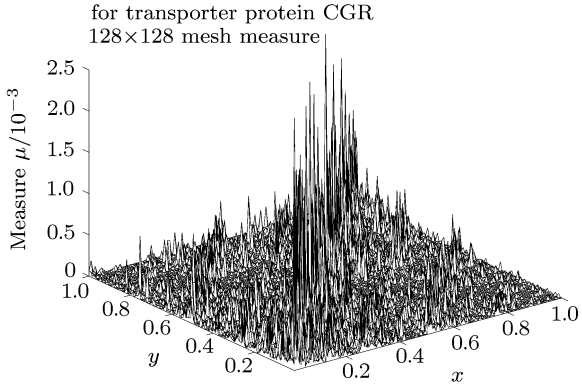


Fig. 2. The 128×128-mesh measure based on the CGR in Fig. 1.

3. Recurrent iterated function systems

Consider a system of contractive maps $S = \{S_1, S_2, \ldots, S_N\}$ and the associated matrix of probabilities $\boldsymbol{P} = (p_{ij})$ such that $\sum_j p_{ij} = 1$, $i = 1, 2, \ldots, N$. We consider a random sequence generated by a dynamical system

$$x_{n+1} = S_{\sigma_n}(x_n), \quad n = 0, 1, 2, \ldots, \tag{2}$$

where x_0 is any starting point and σ_n is chosen among the set $\{1, 2, \ldots, N\}$ with a probability that depends on the previous index σ_{n-1}: $P(\sigma_n = i) = p_{\sigma_{n-1},i}$. Then $(S, \boldsymbol{P})$ is called a RIFS. A major result for RIFS is that there exists a unique invariant measure μ of the random walk (2) whose support is the attractor of the RIFS $(S, \boldsymbol{P})$ (see Ref. [39]).

The coefficients in the contractive maps and the probabilities in the RIFS are the parameters to be estimated for the measure that we want to simulate. We now describe the method of moments to perform this task. In the two-dimensional case of our CGRs, we consider a system of N contractive maps

$$S_i = s_i \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} b_1(i) \\ b_2(i) \end{pmatrix}, \ i = 1, 2, \ldots, N.$$

If μ is the invariant measure and $\hat{A}$ the attractor of the RIFS in $\boldsymbol{R}^2$, the moments of μ are

$$g_{mn} = \int_{\hat{A}} x^m y^n \mathrm{d}\mu = \sum_{j=1}^{N} \int_{\hat{A}_j} x^m y^n \mathrm{d}\mu_j = \sum_{j=1}^{N} g_{mn}^{(j)}.$$

Using the properties of the Markov operator defined by $(S, \boldsymbol{P})$ (Vrscay, 1991), we have

$$\begin{aligned} g_{mn}^{(i)} &= \int_{\hat{A}_i} x^m y^n \mathrm{d}\mu_i = \sum_{j=1}^{N} p_{ji} \int_{\hat{A}_j} (s_j x + b_1(j))^m (s_j y + b_2(j))^n \mathrm{d}\mu_j \\ &= \sum_{j=1}^{N} p_{ji} \sum_{k=0}^{m} \sum_{l=0}^{n} \binom{m}{k} \binom{n}{l} s_j^{k+l} b_1(j)^{m-k} b_2(j)^{n-l} g_{kl}^{(j)}. \end{aligned} \tag{3}$$

When $n = 0, \ m = 0$,

$$g_{00}^{(i)} = \sum_{j=1}^{N} p_{ji} g_{00}^{(j)}, \quad \sum_{j=1}^{N} g_{00}^{(j)} = 1, \quad \sum_{j=1}^{N} (p_{ji} - \delta_{ij}) g_{00}^{(j)} = 0. \tag{4}$$

When $m = 0, n \geq 1$,

$$g_{0n}^{(i)} = \sum_{j=1}^{N} p_{ji} \sum_{l=0}^{n} \binom{n}{l} s_j^l b_2(j)^{n-l} g_{0l}^{(j)},$$

hence the moments are given by the solution of the linear equations

$$\sum_{j=1}^{N} (s_j^n p_{ji} - \delta_{ij}) g_{0n}^{(j)} = -\sum_{l=0}^{n-1} \binom{n}{l} \sum_{j=1}^{N} s_j^l b_2(j)^{n-l} p_{ji} g_{0l}^{(j)}, \ i = 1, \ldots, N. \tag{5}$$

When $n = 0, \ m \geq 1$,

$$g_{m0}^{(i)} = \sum_{j=1}^{N} p_{ji} \sum_{k=0}^{m} \binom{m}{k} s_j^k b_1(j)^{m-k} g_{k0}^{(j)},$$

hence the moments are given by the solution of the linear equations

$$\sum_{j=1}^{N} (s_j^m p_{ji} - \delta_{ij}) g_{m0}^{(j)} = -\sum_{k=0}^{m-1} \binom{m}{k} \sum_{j=1}^{N} s_j^k b_1(j)^{m-k} p_{ji} g_{k0}^{(j)}, \quad i = 1, \ldots, N. \tag{6}$$

When $m, n \geq 1$,

$$\begin{aligned} g_{mn}^{(i)} &= \sum_{j=1}^{N} p_{ji} \sum_{k=0}^{m-1} \sum_{l=0}^{n} \binom{m}{k} \binom{n}{l} s_j^{k+l} b_1(j)^{m-k} b_2(j)^{n-l} g_{kl}^{(j)} \\ &\quad + \sum_{l=0}^{n-1} \binom{n}{l} s_j^{m+l} b_2(j)^{n-l} g_{ml}^{(j)} + \sum_{j=1}^{N} p_{ji} s_j^{m+n} g_{mn}^{(j)}, \end{aligned}$$

hence the moments are given by the solution of the linear equations

$$\sum_{j=1}^{N}\left(s_j^{m+n}p_{ji}-\delta_{ij}\right)g_{mn}^{(j)} = -\sum_{k=0}^{m-1}\sum_{l=0}^{n-1}\binom{m}{k}\binom{n}{l}\sum_{j=1}^{N}s_j^{k+l}b_1(j)^{m-k}\,b_2(j)^{n-l}\,p_{ji}g_{kl}^{(j)} - \sum_{l=0}^{n-1}\binom{n}{l}\sum_{j=1}^{N}s_j^{m+l}b_2(j)^{n-l}\,p_{ji}g_{ml}^{(j)} - \sum_{k=0}^{m-1}\binom{m}{k}\sum_{j=1}^{N}s_j^{k+n}b_1(j)^{m-k}\,p_{ji}g_{kn}^{(j)}, \quad i=1,\ldots,N. \tag{7}$$

If we denote by G_{mn} the moments obtained directly from a given measure, and g_{mn} the formal expression of moments obtained from the above formulae, then solving the optimization problem

$$\min_{s_i,b_1(i),b_2(i),p_{ij}}\sum_{m,n}(g_{mn}-G_{mn})^2$$

will provide the estimates of the parameters of the RIFS.

Once the RIFS $(S_i(x), p_{ji}, i, j = 1, \ldots, N)$ has been estimated, its invariant measure can be simulated in the following way: Generate the attractor $\hat{A}$ of the RIFS via the random walk (2). Let χ_B be the indicator function of a subset B of the attractor $\hat{A}$. From the ergodic theorem for RIFS,[39] the invariant measure is then given by

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

By definition, a RIFS describes the scale invariance of a measure. Hence a comparison of the given measure with the invariant measure simulated from the RIFS will confirm whether the given measure has this scaling behaviour. This comparison can be undertaken by computing the cumulative walk of a measure visualized as intensity values on a $J \times J$ mesh; here $J = 128$ in our case. The cumulative walk is defined as $F_j = \sum_{i=1}^{j}(f_i - \overline{f})$, $j = 1, \ldots, J \times J$, where f_i is the intensity of the i-th point on the extended row formed by concatenating all the rows of the $J \times J$ mesh, and $\overline{f}$ is the average value of all the intensities on the mesh.

Returning to the CGR, a RIFS with 4 contractive maps $\{S_1, S_2, S_3, S_4\}$ is fitted to the measure obtained from the CGR using the method of moments. Here we can fix

$$S_1=\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix},\quad S_2=\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}0\\0.5\end{pmatrix},$$

$$S_3=\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}0.5\\0.5\end{pmatrix},\quad S_4=\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}0.5\\0\end{pmatrix}.$$

Hence the parameters which need to be estimated are the probabilities in the matrix $\boldsymbol{P}$. Once we have estimated the probability matrix in the RIFS, we can start from the point $(0.5, 0.5)$ and use the chaos game algorithm Eq. (2) to generate a random point sequence $\{x_i\}$ with the same length N_l of the linked functional protein sequence. Then we plot the random point sequences. The 128×128-mesh measure μ' based on the plot of the random point sequences can be regarded as a simulation of the measure μ induced from the original CGR. For example, the RIFS simulated measure of the measure in Fig. 2 is shown in Fig. 3. The cumulative walks of these two measures can then be obtained to show the performance of the simulation.

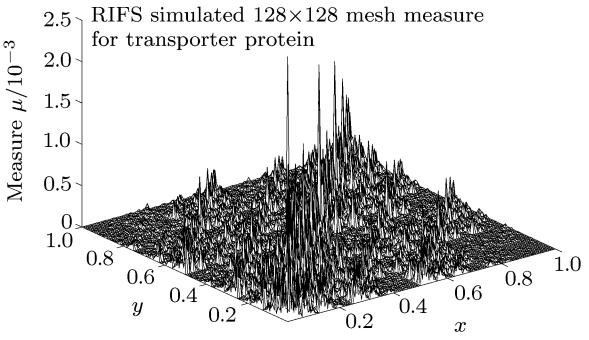


Fig. 3. The RIFS simulated measure for the measure in Fig. 2.

We determine the goodness of fit of the measure simulated from the RIFS model relative to the original measure based on the following relative standard error (RSE)[27]

$$e = \frac{e_1}{e_2},$$

where

$$e_1 = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(F_j - \hat{F}_j)^2},$$

and

$$e_2 = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(F_j - F_{\rm ave})^2}.$$

Here $N = 128 \times 128$, $(F_j)_{j=1}^{N}$ and $(\hat{F}_j)_{j=1}^{N}$ are the walks of the original measure and the RIFS simulated measure respectively. The criterion $e < 1.0$ indicates a good simulation.[27]

4. Multifractal analysis

The multifractal spectrum of a measure μ can be defined, using the box-counting method, as[40]

$$D_q^{bc} = \lim_{\varepsilon\to 0} \frac{\ln\left(\sum_i \left(\frac{M_i}{M_0}\right)^q\right)}{\ln(\varepsilon)} \frac{1}{q-1}, \qquad (8)$$

where ε is the ratio of the grid size to the linear size of the fractal, M_i the number of points falling in the i-th grid cell, M_0 the total number of points in the fractal. We randomly choose a point on the fractal, make a sandbox (a region with radius R) around it, then count the number of points of the fractal that fall in this sandbox of radius R, which is represented as $M(R)$ in the above definition. L is the linear size of the fractal, and q and M_0 have the same meaning as in the definition of D_q^{bc}. The brackets $\langle\bullet\rangle$ mean to take a statistical average over (many) randomly chosen centres of the sandboxes. Because of its dependence on statistical averaging, though the multifractal dimension is defined as $D_q = \lim_{R\to 0} D_q^{sb}(R/L)$ it is better to perform a linear fit on the logarithms of sampled data $\ln(\langle[M(R)]^{q-1}\rangle)$ and take its slope as the multifractal dimension in a practical use of the sandbox method.[41] The idea can be illustrated by rewriting Eq. (8) as

$$\ln(\langle[M(R)]^{q-1}\rangle) = D_q^{sb}(R/L) \times (q-1)\ln(R/L) + (q-1)\ln(M_0). \qquad (9)$$

First, we choose R in an appropriate range $[R_{\min}, R_{\max}]$. For each chosen R, we compute the statistical average of $[M(R)]^{q-1}$ over many radius-R sandboxes randomly distributed on the fractal, $\langle[M(R)]^{q-1}\rangle$, then plot the data on the $\ln(\langle[M(R)]^{q-1}\rangle)$ vs. $(q-1)\ln(R/L)$ plane. We next perform a linear fit on them and calculate the slope as an approximation of the multifractal dimension D_q. D_1 is called the information dimension and D_2 the correlation dimension of the measure. The D_q values for positive values of q are associated with the regions where the points are crowded. The D_q values for negative values of q are associated with the structure and properties of the most rarefied regions. In addition to the multifractal dimension D_q, there is another exponent $\tau(q)$. One can calculate $\tau(q)$ from D_q by $\tau(q) = (q-1)D_q$. Following the thermodynamic formulation of multifractal measures, Canessa[42] derived an expression for the "analogous" specific heat as

$$C_q \equiv -\frac{\partial^2\tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \qquad (10)$$

He showed that the form of C_q resembles a classical phase transition at a critical point. We will discuss the property of C_q for the measure derived from the CGR.

5. Data and result

We downloaded the functional protein sequences with 21 different functions (listed in Table 1) from the public databases at the web site http://www.rcsb.org/pdb/. First, we randomly concatenate the protein sequences with the same function one by one to attain a long linked protein sequence. Then we derive the CGR of these randomly-linked functional protein sequences. We find that the CGRs of randomly-linked functional protein sequences have clear fractal patterns (e.g. in Fig. 1). Then we use the moments of 128×128-mesh measure μ based on the CGR to estimate the parameters (probability matrix) of the RIFS. The RIFS simulation of the measure based on the original CGR is next performed using the chaos game algorithm. To show the performance of the simulation, we compare the cumulative walks of the original measure μ and its simulation μ'. For example, the cumulative walks for the measure in Fig. 2 and its RIFS simulation in Fig. 3 are given in Fig. 4. It is seen that the two walks are almost identical. This indicates that RIFS simulation fits the measure μ induced by the original CGR very well . The RSE= 0.0868 is very small, which also indicates excellent fitting. The values of the RSE of the simulation and the estimated probability matrices using RIFS for 21 different functional protein sequences are listed in Tables

2 and 3. It is seen that all the RES values are much smaller than 1.0, confirming that the RIFS model can simulate the measures of these data very well. This result indicates that we can use the estimated parameters in the RIFS for randomly-linked functional protein sequences to characterize the biological function of proteins. We also find that the estimated probability matrices of the RIFS with different biological functions are evidently different (in Tables 2 and 3).

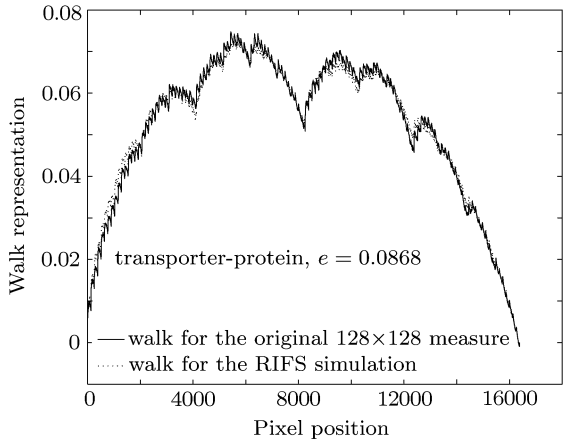


Fig. 4. The walk representation of measures in Figs. 2 and 3.

This fact implies that the CGR and estimated probability matrices in the RIFS can be used to characterize the differences among proteins with different biological functions.

Table 1. The selected functional protein sequences.

name of function	number of sequences	total of residues
transporter	748	423140
carbohydrate binding	430	378069
cofactor binding	1124	1029044
enzyme inhibitor	313	116417
hydrolase	5289	2995640
ion binding	4011	2768585
isomerase	545	373945
ligase	386	373744
lipid binding	259	95265
lyase	824	719911
metal cluster binding	228	250765
nucleic acid binding	2563	1562072
nucleotide binding	1942	1611997
oxidoreductase	2910	2530377
oxygen binding	362	158967
protein binding	1582	1165254
signal transducer	564	272711
structural molecule	488	518035
tetrapyrrole binding	915	567618
transcription factor	669	272640
transferase	2869	2298127

Table 2. The results of RIFS simulation for measures based on CGRs of first 11 linked functional protein sequences.

name of function	estimated probability matrix P	relative standard error
transporter	$\begin{pmatrix} 0.450213 & 0.146109 & 0.269893 & 0.133785 \\ 0.388836 & 0.035165 & 0.301606 & 0.274394 \\ 0.357528 & 0.143895 & 0.343036 & 0.155540 \\ 0.378738 & 0.276505 & 0.271186 & 0.073571 \end{pmatrix}$	0.0868
carbohydrate binding	$\begin{pmatrix} 0.410654 & 0.140257 & 0.319110 & 0.129978 \\ 0.360625 & 0.006062 & 0.359401 & 0.273911 \\ 0.367067 & 0.130879 & 0.380106 & 0.121948 \\ 0.357719 & 0.289410 & 0.304302 & 0.048569 \end{pmatrix}$	0.2803
cofactor binding	$\begin{pmatrix} 0.436893 & 0.158166 & 0.239309 & 0.165632 \\ 0.389684 & 0.045964 & 0.272624 & 0.291728 \\ 0.385111 & 0.129538 & 0.329393 & 0.155958 \\ 0.383246 & 0.274135 & 0.289505 & 0.053113 \end{pmatrix}$	0.1104
enzyme inhibitor	$\begin{pmatrix} 0.417343 & 0.146152 & 0.266855 & 0.169650 \\ 0.325488 & 0.041798 & 0.346359 & 0.286355 \\ 0.333169 & 0.108311 & 0.438828 & 0.119692 \\ 0.343527 & 0.260933 & 0.341574 & 0.053965 \end{pmatrix}$	0.2579

Table 2. (*Continued*).

name of function	estimated probability matrix P	relative standard error
hydrolase	$\begin{pmatrix} 0.433106 & 0.127933 & 0.310725 & 0.128237 \\ 0.344591 & 0.113996 & 0.272995 & 0.268418 \\ 0.384803 & 0.104315 & 0.391288 & 0.119594 \\ 0.340101 & 0.243037 & 0.284838 & 0.132025 \end{pmatrix}$	0.0931
ion binding	$\begin{pmatrix} 0.427150 & 0.152089 & 0.271574 & 0.149187 \\ 0.375735 & 0.062878 & 0.284718 & 0.276668 \\ 0.368963 & 0.132533 & 0.344346 & 0.154159 \\ 0.368460 & 0.269133 & 0.273180 & 0.089226 \end{pmatrix}$	0.0807
isomerase	$\begin{pmatrix} 0.438661 & 0.165248 & 0.236109 & 0.159982 \\ 0.384871 & 0.059741 & 0.277002 & 0.278387 \\ 0.398943 & 0.127263 & 0.322570 & 0.151223 \\ 0.363218 & 0.270314 & 0.272192 & 0.094275 \end{pmatrix}$	0.0756
ligase	$\begin{pmatrix} 0.432127 & 0.183405 & 0.207602 & 0.176867 \\ 0.386173 & 0.072646 & 0.265652 & 0.275529 \\ 0.393155 & 0.131294 & 0.330271 & 0.145279 \\ 0.377211 & 0.271526 & 0.272147 & 0.079116 \end{pmatrix}$	0.0658
lipid binding	$\begin{pmatrix} 0.456351 & 0.151894 & 0.212203 & 0.179552 \\ 0.376735 & 0.080904 & 0.273943 & 0.268418 \\ 0.327128 & 0.158360 & 0.354428 & 0.160085 \\ 0.387015 & 0.252199 & 0.280772 & 0.080013 \end{pmatrix}$	0.1227
lyase	$\begin{pmatrix} 0.445717 & 0.154341 & 0.233529 & 0.166413 \\ 0.381712 & 0.054147 & 0.283836 & 0.280304 \\ 0.383945 & 0.145088 & 0.313208 & 0.157759 \\ 0.378279 & 0.270513 & 0.296520 & 0.054688 \end{pmatrix}$	0.0763
metal cluster binding	$\begin{pmatrix} 0.434070 & 0.167911 & 0.236312 & 0.161706 \\ 0.389813 & 0.055780 & 0.267971 & 0.286436 \\ 0.359287 & 0.131208 & 0.353842 & 0.155664 \\ 0.381281 & 0.275748 & 0.283824 & 0.059147 \end{pmatrix}$	0.1391

Table 3. The results of RIFS simulation for measures based on CGRs of another 10 linked functional protein sequences.

name of function	estimated probability matrix P	relative standard error
nucleic acid binding	$\begin{pmatrix} 0.443988 & 0.134275 & 0.279522 & 0.142215 \\ 0.302086 & 0.161555 & 0.179193 & 0.357166 \\ 0.347288 & 0.069234 & 0.470508 & 0.112971 \\ 0.308504 & 0.303656 & 0.187827 & 0.200013 \end{pmatrix}$	0.1883
nucleotide binding	$\begin{pmatrix} 0.411430 & 0.187213 & 0.215806 & 0.185551 \\ 0.382549 & 0.081912 & 0.251593 & 0.283946 \\ 0.349295 & 0.125079 & 0.382183 & 0.143442 \\ 0.377236 & 0.274682 & 0.259434 & 0.088648 \end{pmatrix}$	0.0646
oxidoreductase	$\begin{pmatrix} 0.434337 & 0.156854 & 0.247782 & 0.161028 \\ 0.386387 & 0.044862 & 0.277748 & 0.291003 \\ 0.375481 & 0.137469 & 0.327993 & 0.159057 \\ 0.381368 & 0.278013 & 0.291883 & 0.048737 \end{pmatrix}$	0.1220

Table 3. (*Continued*).

name of function	estimated probability matrix P	relative standard error
oxygen binding	$\begin{pmatrix} 0.434534 & 0.136699 & 0.243972 & 0.184794 \\ 0.441283 & 0.039678 & 0.215024 & 0.304015 \\ 0.469058 & 0.116619 & 0.194576 & 0.219747 \\ 0.402206 & 0.232240 & 0.245078 & 0.120476 \end{pmatrix}$	0.3520
protein binding	$\begin{pmatrix} 0.420250 & 0.138550 & 0.303862 & 0.137338 \\ 0.348988 & 0.157815 & 0.231919 & 0.261278 \\ 0.371916 & 0.108368 & 0.399509 & 0.120206 \\ 0.349691 & 0.261954 & 0.233388 & 0.154967 \end{pmatrix}$	0.1118
signal transducer	$\begin{pmatrix} 0.428426 & 0.161678 & 0.256124 & 0.153772 \\ 0.380776 & 0.041920 & 0.300875 & 0.276428 \\ 0.340710 & 0.136383 & 0.356221 & 0.166686 \\ 0.365345 & 0.263556 & 0.275213 & 0.095886 \end{pmatrix}$	0.0965
structural molecule	$\begin{pmatrix} 0.447296 & 0.108242 & 0.312921 & 0.131541 \\ 0.203792 & 0.213694 & 0.271992 & 0.310522 \\ 0.382985 & 0.018085 & 0.550290 & 0.048640 \\ 0.224449 & 0.271059 & 0.246372 & 0.258120 \end{pmatrix}$	0.1779
tetrapyrrole binding	$\begin{pmatrix} 0.448680 & 0.147100 & 0.240014 & 0.164207 \\ 0.395693 & 0.046637 & 0.268870 & 0.288800 \\ 0.403180 & 0.121674 & 0.306254 & 0.168892 \\ 0.377297 & 0.259184 & 0.276206 & 0.087313 \end{pmatrix}$	0.1337
transcription factor	$\begin{pmatrix} 0.449654 & 0.136652 & 0.263659 & 0.150036 \\ 0.347839 & 0.138495 & 0.199026 & 0.314639 \\ 0.370657 & 0.091650 & 0.393514 & 0.144180 \\ 0.330393 & 0.283476 & 0.197890 & 0.188241 \end{pmatrix}$	0.1607
transferase	$\begin{pmatrix} 0.448092 & 0.135501 & 0.273635 & 0.142771 \\ 0.363655 & 0.122166 & 0.249039 & 0.265139 \\ 0.426822 & 0.109897 & 0.330417 & 0.132864 \\ 0.345461 & 0.253430 & 0.262836 & 0.138274 \end{pmatrix}$	0.0343

When we randomly concatenate the protein sequences with the same function one by one to attain a long linked protein sequence, the orders to link the sequences randomly are enormous. For example, the number of functional protein sequences with carbohydrate binding function is 430, so the number of possible orders to link these sequences are 430! Apparently, it is difficult to check the results of simulations for all the CGR of differently linked sequences. So we randomly selected 50 different linked sequences to test it. By experiment, we find that different orders give almost the same relative standard error and the same probability matrix. This means when we use RIFS to simulate the measure based on CGR of linked functional protein sequences, the relative standard error and the probability matrix are independent of the order to link the functional protein sequences.

We calculated the dimension spectra (D_q) and "analogous" specific heat (C_q) of the measure μ from their CGRs. We show the D_q curves of the measures μ from the CGRs of these 21 kinds of functional protein sequences in Fig. 5 and their C_q curves in Fig. 6. If a sequence is completely random, $D_q = 2$ for all q. It is apparent from Fig. 5 that the D_q curves are nonlinear and significantly different from those of completely random sequences. Hence the randomly-linked functional protein sequences are not completely random sequences. From the plot of D_q, the dimension spectra of the measure μ exhibit a multifractal-like form. The phase transition-like phenomenon in the C_q curves can indicate the complexity of functional proteins. From Fig. 6, the C_q curves of functional proteins resemble a classical phase transition at a critical point.

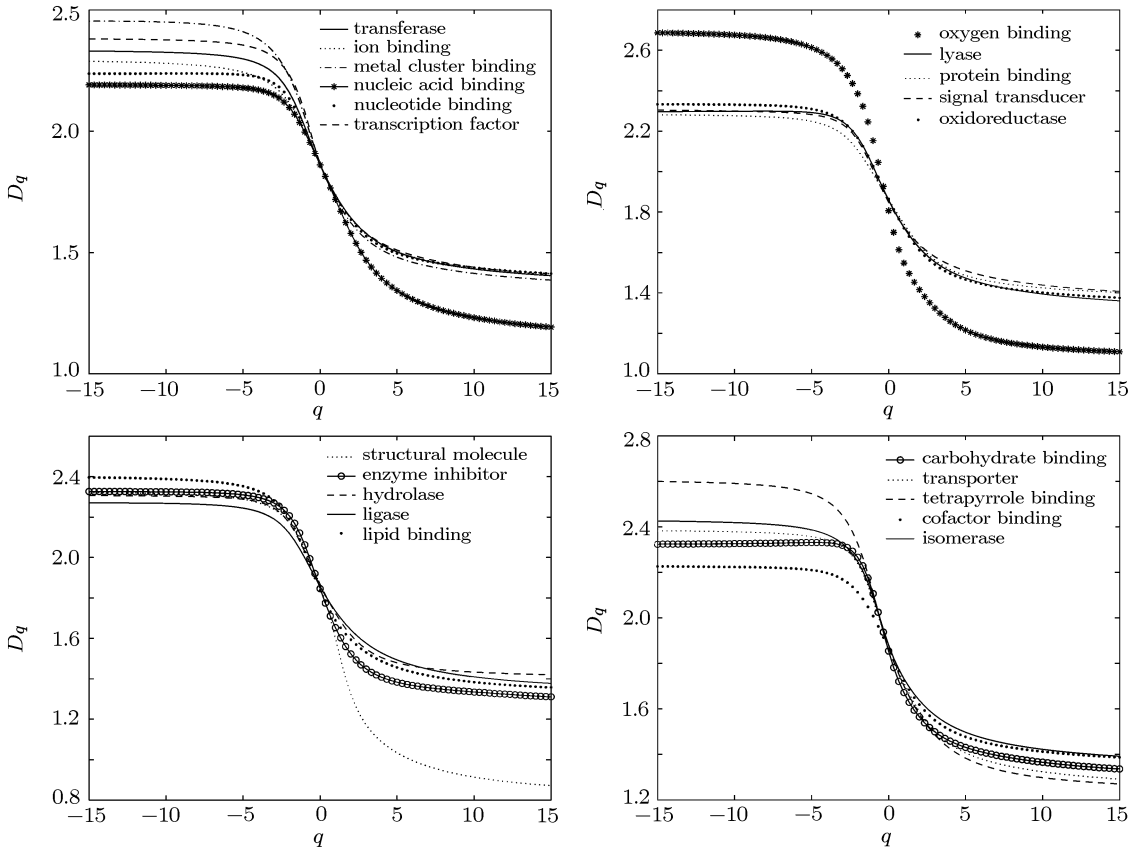


Fig. 5. The D_q curves of the measure μ induced by the CGRs of linked functional protein sequences.

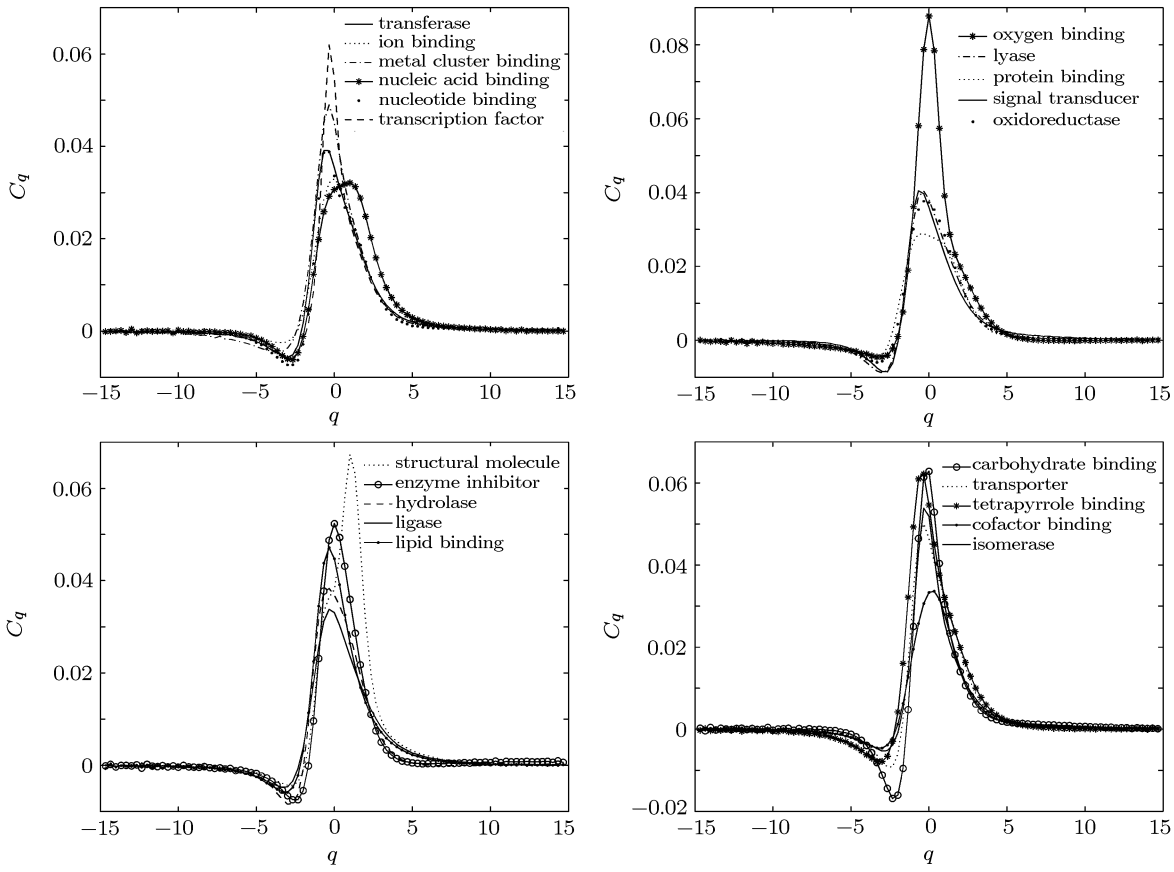
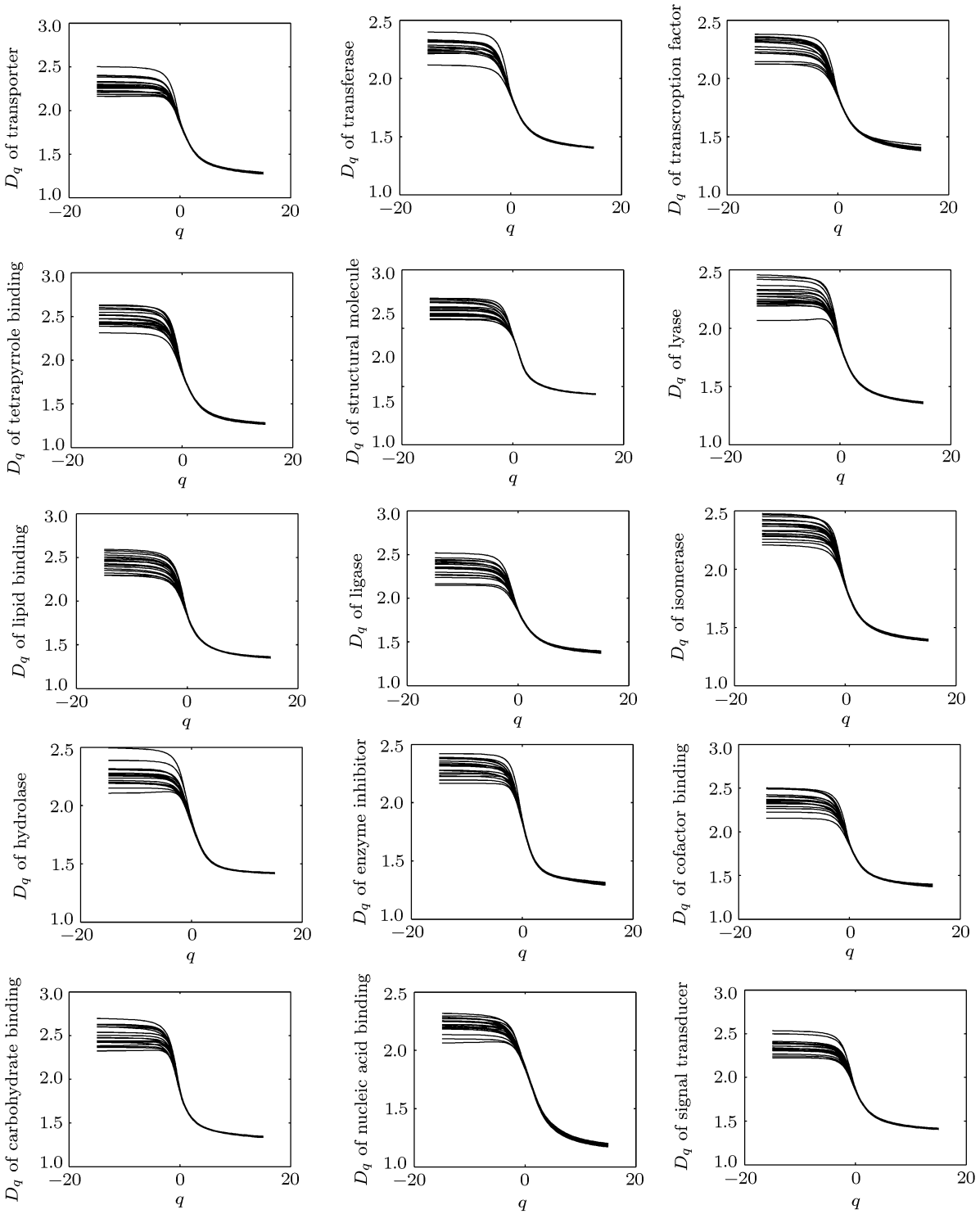


Fig. 6. The C_q curves of the measure μ induced by the CGRs of linked functional protein sequences.

We also need to test whether the D_q of the measure μ from their CGRs based on the different orders to link the sequences randomly are identical. In the same way of considering whether the results of their simulation are independent of the order to link the sequences randomly, we randomly selected 20 linked sequences with different orders to link, then produce their CGRs and calculated D_q of the measure μ from their CGRs in Fig. 7. It is apparent that the D_q spectra of the measure μ based on the CGRs of the linked sequences with different orders are almost identical for $q \geq 0$.



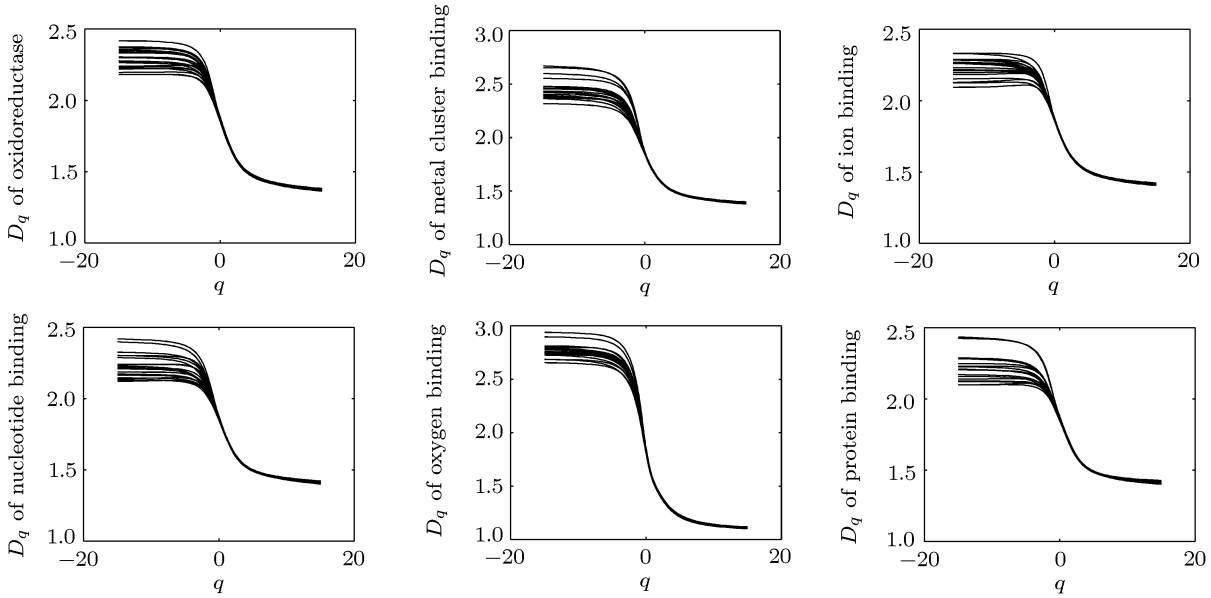


Fig. 7. The D_q curves of the measure μ based on CGRs of linked functional protein sequences using different orders to link.

6. Conclusions

The CGR based on the detailed HP model of functional protein sequences provides a simple yet powerful visualisation method to distinguish functional protein sequences themselves in more details.

The CGRs of randomly-linked protein sequences have clear fractal patterns. The RIFS can simulate the measures based on these CGRs very well. The relative standard error and the probability matrix are independent of the order to link the functional protein sequences. The estimated probability matrices of the RIFS for linked sequences with different biological functions have clear differences. This fact indicates that the CGRs and estimated probability matrices in the RIFS can be used to characterize the differences among protein sequences with different biological functions.

Multifractal analysis provides a simple yet powerful method to amplify the difference between a randomly-linked functional protein sequence and a random sequence. The D_q spectra of all linked functional protein sequences studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. The D_q spectra of the measure μ from their CGRs based on the different orders to link the functional protein sequences are almost identical for $q \geq 0$. The D_q and C_q curves indicate that the point sequences in the CGRs of all functional protein sequences considered here are not completely random. The phase transition-like phenomenon in the C_q curves indicates the complexity of functional proteins. The C_q curves of functional protein sequences resemble a classical phase transition at a critical point.

References

[1] Venter J C, Adams M D, Myers E W, *et al.* 2001 *Science* **291** 1304

[2] Pandey A and Mann M 2000 *Nature* **405** 837

[3] Jeffrey H J 1990 *Nucleic Acids Research* **18** 2163

[4] Goldman N 1993 *Nucleic Acids Research* **21** 2487

[5] Deschavanne P J, Giron A, Vilain J, Fagot G and Fertil B 1999 *Mol. Biol. Evol.* **16** 1391

[6] Almeida J S, Carrico J A, Maretzek A, Noble P A and Fletcher M 2001 *Bioinformatics* **17** 429

[7] Joseph J and Sasikumar R 2006 *BMC Bioinformatics* **7** 243(1-10)

[8] Gao J and Xu Z Y 2009 *Chin. Phys.* B **18** 370

[9] Gao J, Jiang L L and Xu Z Y 2009 *Chin. Phys.* B **18** 4571

[10] Fiser A, Tusnady G E and Simon I 1994 *J. Mol. Graphics* **12** 302

[11] Basu S, Pan A, Dutta C and Das J 1998 *J. Mol. Graphics and Modelling* **15** 279

[12] Yu Z G, Anh V V and Lau K S 2004 *J. Theor. Biol.* **226** 341

[13] Yu Z G, Anh V V and Lau K S 2004 *Physica* A **337** 171

[14] Dill K A 1985 *Biochemistry* **24** 1501
[15] Wang J and Wang W 2000 *Phys. Rev.* E **61** 6981
[16] Brown T A 1998 *Genetics* 3rd ed. (London: Chapman & Hall)
[17] Huang Y Z and Xiao Y 2003 *Chaos, Solitons and Fractals* **17** 895
[18] Huang Y Z, Li M F and Xiao Y 2007 *Chaos, Solitons and Fractals* **34** 782
[19] Feng J, Liu J H and Zhang H G 2008 *Acta Phys. Sin.* **57** 6868 (in Chinese)
[20] Chen Y P, Fu P P, Shi M H, Wu J F and Zhang C B 2009 *Acta Phys. Sin.* **58** 7050 (in Chinese)
[21] Yu Z G and Anh V V 2001 *Chaos, Solitons and Fractals* **12(10)** 1827
[22] Yu Z G and Wang B 2001 *Chaos, Solitons and Fractals* **12** 519
[23] Yu Z G, Anh V V, Gong Z M and Long S C 2002 *Chin. Phys.* **11** 1313
[24] Barnsley M F and Demko S 1985 *Proc. R. Soc. London* Ser. A **399** 243
[25] Falconer K 1997 *Techniques in Fractal Geometry* (London: John Wiley & Sons)
[26] Vrscay E R 1991 *Fractal Geometry and Analysis* ed. Belair J and Dubuc S (Dordrecht: Kluwer) pp. 405–468
[27] Anh V V, Lau K S and Yu Z G 2002 *Phys. Rev.* E **66** 031910
[28] Yu Z G, Anh V V and Lau K S 2001 *Phys. Rev.* E **64** 031903
[29] Yu Z G, Anh V V and Lau K S 2003 *Int. J. Mod. Phys.* B **17** 4367
[30] Yu Z G, Anh V V and Lau K S 2003 *J. Xiangtan Univ.* (Natural Science Edition) **25(3)** 131
[31] Wanliss J A, Anh V V, Yu Z G and Watson S 2005 *J. Geophys. Res.* **110** A08214
[32] Anh V V, Yu Z G, Wanliss J A and Watson S M 2005 *Nonlin. Processes Geophys.* **12** 799
[33] Yu Z G, Anh V V, Wanliss J A and Watson S M 2007 *Chaos, Solitons and Fractals* **31** 736
[34] Hentschel H G E and Procaccia I 1983 *Physica* D **8** 435
[35] Gutierrez J M, Iglesias A and Rodriguez M A 1998 *Chaos and Noise in Biology and Medicine* ed. Barbi M and Chillemi S (Singapore: World Scientific) pp. 315–319
[36] Gutierrez J M, Rodriguez M A and Abramson G 2001 *Physica* A **300** 271
[37] Yu Z G, Anh V V, Lau K S and Zhou L Q 2006 *Phys. Rev.* E **63** 031920
[38] Yang J Y, Yu Z G and Anh V V 2009 *Chaos, Solitons and Fractals* **40** 607
[39] Barnley M F, Elton J H and Hardin D P 1989 *Constr. Approx.* B **5** 3
[40] Halsy T, Jensen M, Kadanoff L, Procaccia I and Schraiman B 1986 *Phys. Rev.* A **33** 1141
[41] Tel T, Fulop A and Vicsek T 1989 *Physica* A **159** 155
[42] Canessa E 2000 *J. Phys.* A: *Math. Gen.* **33** 3637