Available online at www.sciencedirect.com
ScienceDirect
Chaos, Solitons and Fractals 40 (2009) 607–620
CHAOS
SOLITONS & FRACTALS
www.elsevier.com/locate/chaos

Clustering structures of large proteins using multifractal analyses based on a 6-letter model and hydrophobicity scale of amino acids

Jian-Yi Yang [a], Zu-Guo Yu [a,b,*], Vo Anh [b]

[a] School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China
[b] School of Mathematical Sciences, Queensland University of Technology, G.P.O. Box 2434, Brisbane, Q 4001, Australia

Accepted 8 August 2007

Abstract

The Schneider and Wrede hydrophobicity scale of amino acids and the 6-letter model of protein are proposed to study the relationship between the primary structure and the secondary structural classification of proteins. Two kinds of multifractal analyses are performed on the two measures obtained from these two kinds of data on large proteins. Nine parameters from the multifractal analyses are considered to construct the parameter spaces. Each protein is represented by one point in these spaces. A procedure is proposed to separate large proteins in the α, β, $\alpha+\beta$ and α/β structural classes in these parameter spaces. Fisher's linear discriminant algorithm is used to assess our clustering accuracy on the 49 selected large proteins. Numerical results indicate that the discriminant accuracies are satisfactory. In particular, they reach 100.00% and 84.21% in separating the α proteins from the $\{\beta, \alpha+\beta, \alpha/\beta\}$ proteins in a parameter space; 92.86% and 86.96% in separating the β proteins from the $\{\alpha+\beta, \alpha/\beta\}$ proteins in another parameter space; 91.67% and 83.33% in separating the α/β proteins from the $\alpha+\beta$ proteins in the last parameter space.

1. Introduction

The three-dimensional (3D) structure of a protein is determined by its amino acid sequence via the process of protein folding [1]. Four main classes of protein structures were recognized based on the types and arrangement of their secondary structural elements [2]. They are the α helices, the β strands and those with a mixture of α and β shapes called $\alpha+\beta$ and α/β. The 3D structures of a large number of protein molecules have been determined either by experimental or theoretical methods and deposited in the RCSB Protein Data Bank [3]. Richardson [4] proposed that proteins of disparate evolutionary origins could adopt similar structures. As a result, the number of protein 3D structures was predicted to be much smaller than the number of protein families defined by sequence similarity [5,6].

Hou et al. [7,8] constructed a map of the protein structure space using the pairwise structural similarity scores calculated for 1898 protein chains. Their study showed that the four main classes of protein structures cluster together

* Corresponding author. Address: School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China.
E-mail addresses: yuzg@hotmail.com, z.yu@qut.edu.au (Z.-G. Yu).

0960-0779/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chaos.2007.08.014

as four elongated arms from a common center. Yu et al. [9] used the hydrophobic free energy and solvent accessibility of proteins to construct several parameter spaces. They found that some spaces could be used to distinguish and cluster the 43 selected large proteins in the four structural classes. In this paper, we aim to classify 49 selected large proteins which include the 43 proteins studied in Ref. [9] by some new methods which produce better results.

Dill [10] proposed a simplified but well-known HP model of protein behavior. The 20 kinds of amino acids are divided into two types in this model, namely hydrophobic (H) (or nonpolar) and polar (P) (or hydrophilic). But the HP model might lack sufficient information on the heterogeneity and the complexity of the natural set of residues [11]. The polar class of the HP model can be divided into three subclasses: positive, uncharged, and negative polar according to Brown [12]. So the 20 different kinds of amino acids can be divided into four classes: nonpolar, negative polar, uncharged polar, and positive polar. This model is called *detailed HP model* by Yu et al. [9,13]. Chou and Fasman [14] analysed the frequency of amino acids in many known protein structures and divided the 20 kinds of amino acids into six classes. We call one of the classification of amino acids as *6-letter model* and use this model to cluster protein structures in this paper.

Kanzman [15] first proposed that the hydrophobic bonds within proteins are largely responsible for a protein maintaining its native globular form (i.e. its 3D structure). Then a large number of hydrophobicity scales were proposed [16–18]. Palliser and Parry [19] gave a quantitative comparison of the 127 hydrophobicity scales to recognize the surface β-strands in proteins. They found that some scales turned out to be better than others. Giuliani et al. [20–23] applied an interesting *Recurrence Quantification Analysis* (RQA) to the hydrophobicity sequence of proteins according to the Schneider and Wrede scale (SWH). They showed the general meaning and scope of the application of signal analysis methods to protein-structure relationships. The SWH of the 20 kinds of amino acids are: $A = 1.6$, $R = -12.3$, $N = -4.8$, $D = -9.2$, $C = 2$, $Q = -1.1$, $E = -8.2$, $G = 1$, $H = -3$, $I = 3.1$, $L = 2.8$, $K = -8.8$, $M = 2.4$, $F = 3.7$, $P = -0.2$, $S = 0.6$, $T = 1.2$, $W = 1.9$, $Y = -0.7$, and $V = 2.6$ [20–23]. In order to make all the values non-negative as required in our methods, we add a common value 12.30 to these 20 values (we will explain how to choose the common value below). We denote the revised hydrophobicity scale by RSWH. For example, we give the RSWH sequence of the protein 1A8I in Fig. 1.

Nonlinear methods turn out to be a useful tool in many different fields. Huang and Xiao [24] 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. [25] showed that amino acid sequences of many multi-domain proteins had hidden repetitions.

Fractal analysis first proposed by Mandelbrot [26] is one of the most popular nonlinear methods. Ordinary multifractal analysis (MFA) has been widely used in many different fields successfully [9,13,27–36]. Han et al. [34] developed a new multifractal traffic model to capture the multifractal nature of modern Internet traffic based on MFA. A multifractal spectrum was used to classify traffic behavior measurements by Li and Shang [35]. Ma et al. [36] studied the entropies and multifractal spectrum of some compact systems.

Another kind of multifractal analysis which is analogous to multiaffinity analysis [37–40] is proposed in this paper. We call it *Analogous Multifractal Analysis* (AMFA). We want to use these two kinds of multifractal analyses to analyse

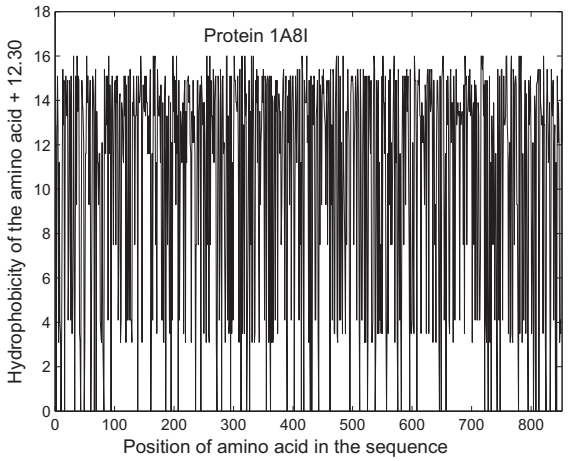

Fig. 1. The RSWH sequence of the protein 1A8I.

both the RSWH sequences and the 6-letter model of proteins. Nine parameters from the multifractal analyses are considered to construct some parameter spaces. Each protein is represented by one point in these spaces. Then a procedure is proposed to distinguish proteins from the four structural classes in these parameter spaces. Fisher's linear discriminant algorithm demonstrates that our clustering accuracies are satisfactory.

2. Models and methods

2.1. 6-Letter model and measure representation of protein sequence

According to Chou and Fasman [14], the 20 different kinds of amino acids are divided into six classes: strong β former (H_β); β former (h_β); weak β former (I_β); β indifferent (i_β); β breaker (b_β); and strong β breaker (B_β). The H_β class consists of the three residues Met, Val, IlE; the h_β class consists of the seven residues Cys, Tyr, Phe, Gln, Leu, Thr, Trp; the I_β class consists of the residue Ala; the i_β class consists of the three residues Arg, Gly, Asp; the b_β class is made up of the five residues Lys, Ser, His, Asn, Pro; and the remaining residue Glu constitutes the B_β class.

For a given protein sequence $s = s_1, \ldots, s_L$ with length L, where s_i is one of the 20 kinds of amino acids for $i = 1, \ldots, L$, similar to the map in Ref. [13], we define

$$a_i = \begin{cases} 0, & \text{if } s_i \text{ is in the } B_\beta \text{ class,} \\ 1, & \text{if } s_i \text{ is in the } b_\beta \text{ class,} \\ 2, & \text{if } s_i \text{ is in the } i_\beta \text{ class,} \\ 3, & \text{if } s_i \text{ is in the } I_\beta \text{ class,} \\ 4, & \text{if } s_i \text{ is in the } h_\beta \text{ class,} \\ 5, & \text{if } s_i \text{ is in the } H_\beta \text{ class.} \end{cases} \tag{1}$$

This results in a sequence $X(s) = a_1, \ldots, a_L$, where a_i is a letter of the alphabet $\{0,1,2,3,4,5\}$.

We now outline the method of Yu et al. [9,13] to derive the measure representation of a protein sequence. We call any string made of K letters from the set $\{0,1,2,3,4,5\}$ a K-string. In order to count the number of K-strings in a sequence $X(s)$ from a protein sequence s, 6^K counters are needed. We divide the interval $[0,1)$ into 6^K disjoint sub-intervals, and use each sub-interval to represent a counter. Letting $r = r_1, \ldots, r_K$, $r_i \in \{0,1,2,3,4,5\}$, $i = 1, \ldots, K$, be a substring with length K, we define

$$x_{\text{left}}(r) = \sum_{i=1}^{K} \frac{r_i}{6^i} \quad x_{\text{right}}(r) = \frac{1}{6^K} + \sum_{i=1}^{K} \frac{r_i}{6^i}. \tag{2}$$

We then use the subinterval $[x_{\text{left}}(r), x_{\text{right}}(r))$ to represent substring r. Let $N_K(r)$ be the number of times that a substring r with length K appears in the sequence $X(s)$ (when we count these numbers, we open a reading frame with width K and slide the frame one amino acid each time). We define $F_K(r) = N_K(r)/(L - K + 1)$ to be the frequency of substring r. It follows that $\sum_{\{r\}} F_K(r) = 1$. We can now define a measure μ_K on $[0,1)$ by $\mu_K(\mathrm{d}x) = Y_K(x)\mathrm{d}x$, where

$$Y_K(x) = 6^K F_K(r), \quad \text{when } x \in [x_{\text{left}}(r), x_{\text{right}}(r)). \tag{3}$$

It is seen that $\mu_K([0,1)) = 1$ and $\mu_K([x_{\text{left}}(r), x_{\text{right}}(r))) = F_K(r)$. We call μ_K the *measure representation* of the protein sequence corresponding to the given K. As examples, the histograms of the measure representation for protein 1A8I for $K = 2,3,4,5$ are given in Fig. 2.

2.2. Measure for the RSWH sequence of proteins

In this section, we construct a measure from the RSWH sequence of a protein with a similar method in Refs. [32,41].

Let T_t, $t = 1,2,\ldots,N$, be the RSWH sequence of a protein with length N. First, we define $F_t = T_t/(\sum_{j=1}^{N} T_j)$, $(t = 1,2,\ldots,N)$ to be the frequency of T_t. It follows that $\sum_{t=1}^{N} F_t = 1$. Now we can define a measure ν on the interval $[0,1)$ by $\nu(\mathrm{d}x) = Y_1(x)\mathrm{d}x$, where

$$Y_1(x) = N \times F_t = T_t \Big/ \left(\frac{1}{N}\sum_{j=1}^{N} T_j\right), \quad x \in \left[\frac{t-1}{N}, \frac{t}{N}\right). \tag{4}$$

We denote the interval $\left[\frac{t-1}{N}, \frac{t}{N}\right)$ by I_t. It is easy to see that $\nu([0,1)) = 1$ and $\nu(I_t) = F_t$. We call $\nu(x)$ the *measure* for the RSWH sequence of a protein.

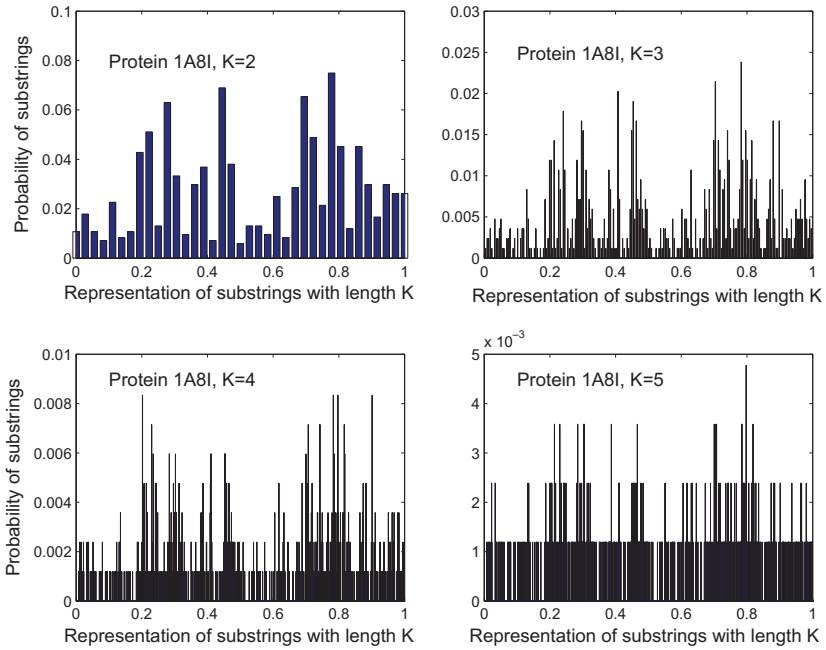

Fig. 2. The histograms of measure representation for proteins 1A8I.

2.3. Ordinary multifractal analysis (MFA)

The most common algorithms of multifractal analysis are the so called *fixed-size box-counting algorithms* [42]. In the one-dimensional case, for a given measure μ with support $E \subset \mathbf{R}$, we consider the *partition sum*

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q, \quad q \in \mathbf{R}, \tag{5}$$

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, $B = [k\varepsilon, (k+1)\varepsilon)$. The *mass exponent* $\tau(q)$ is defined [26,43] by

$$\tau(q) = \lim_{\varepsilon\to 0} \frac{\ln Z_\varepsilon(q)}{\ln \varepsilon}, \tag{6}$$

and the generalized *fractal dimensions* [26,43] of the measure are defined as

$$D_q = \tau(q)/(q-1), \quad \text{for } q \neq 1, \tag{7}$$

and

$$D_q = \lim_{\varepsilon\to 0} \frac{Z_{1,\epsilon}}{\ln \varepsilon}, \quad \text{for } q = 1, \tag{8}$$

where $Z_{1,\varepsilon} = \sum_{\mu(B)\neq 0} \ln \mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of $\ln Z_\varepsilon(q)/(q-1)$ against $\ln \varepsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\varepsilon}$ against $\ln \varepsilon$ for $q = 1$ [9,13,43]. For example, we show how to obtain the D_q spectrum using the slopes of the linear regressions in Fig. 3. The value D_1 is called the *information dimension* and D_2 the *correlation dimension* [26,43].

The concept of *phase transition* in multifractal spectra was introduced in the study of logistic maps, Julia sets, and other simple systems. Evidence of a phase transition was found in the multifractal spectrum of diffusion-limited aggregation [44]. By following the thermodynamic formulation of multifractal measures, Canessa [45] derived an expression for the analogous specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1), \tag{9}$$

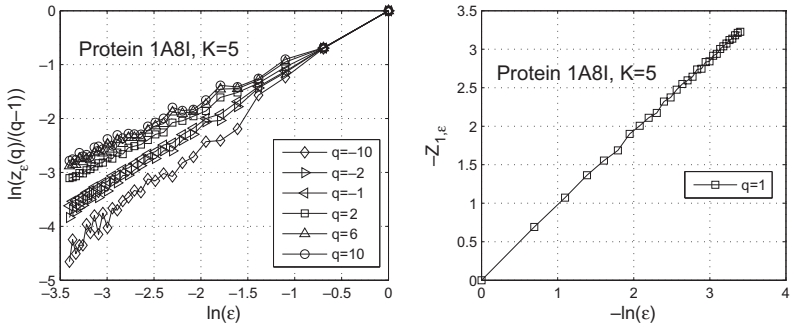

Fig. 3. The linear slopes in the D_q spectra.

and showed that the form of C_q resembles a classical phase transition at a critical point for financial time series.

The singularities of a measure are characterized by the *Lipschitz–Hölder exponent* $\alpha(q)$ [43], which is related to $\tau(q)$ by

$$\alpha(q) = \frac{\mathrm{d}}{\mathrm{d}q}\tau(q). \tag{10}$$

Substitution of Eq. (6) into Eq. (10) yields

$$\alpha(q) = \lim_{\varepsilon\to 0} \frac{\sum_{\mu(B)\neq 0}[\mu(B)]^q \ln \mu(B)}{Z_\varepsilon(q)\ln\varepsilon}. \tag{11}$$

Again the exponent $\alpha(q)$ can be estimated through a linear regression of $\sum_{\mu(B)\neq 0}[\mu(B)]^q \ln \mu(B)/Z_\varepsilon(q)$ against $\ln\varepsilon$ [23]; and the multifractal spectrum $f(\alpha)$ versus α can be calculated according to the relationship usually called *Legendre transformation* [43]:

$$f(\alpha) = q\alpha(q) - \tau(q). \tag{12}$$

2.4. Analogous multifractal analysis (AMFA)

This kind of multifractal analysis is analogous to *multiaffinity analysis* which is a useful method in many fields [37–40]. We denote a time series as $X(t)$, $t = 1,2,\ldots,N$. First, the time series is integrated as

$$y'_q(k) = \sum_{t=1}^{k}(X(t) - X_{\mathrm{ave}})^q \quad (q > 0), \tag{13}$$

$$y_q(k) = \sum_{t=1}^{k}|X(t) - X_{\mathrm{ave}}|^q \quad (q\neq 0), \tag{14}$$

where X_{ave} is the average over the whole time period. Then two quantities $M_q(L)$ and $M'_q(L)$ are defined as

$$M'_q(L) = [\langle|y'(j) - y'(j+L)|\rangle_j]^{\frac{1}{q}} \quad (q > 0), \tag{15}$$

$$M_q(L) = [\langle|y(j) - y(j+L)|\rangle_j]^{\frac{1}{q}} \quad (q\neq 0), \tag{16}$$

where $\langle\ \rangle_j$ denotes the average over j, $j = 1,2,\ldots,N-L$; L typically varies from 1 to N_1 for which the linear fit is good. From the $\ln L$ vs. $\ln M_q(L)$ and $\ln L$ vs. $\ln M'_q(L)$ planes, one can find the relations:

$$M'_q(L) \propto L^{h'(q)} \quad \text{for } q > 0, \tag{17}$$

$$M_q(L) \propto L^{h(q)} \quad \text{for } q\neq 0. \tag{18}$$

Linear regressions of $\ln M_q(L)$ and $\ln M'_q(L)$ against $\ln L$ will result in the exponents $h(q)$ and $h'(q)$, respectively.

The exponent $h(q)$ has a nonlinear dependence on q. When $q = 1$, the methods are just those reported in Refs.[46,47] and these methods are used to study the length sequences from the complete genomes by Yu et al. [48]. $M'(L)$ may get long-range correlation [37]. From Ref. [47], the linear fit to get the exponent $h(1)$ is better than that to get the exponent $h'(1)$. Our numerical results show that the exponents $h(q)$ are more robust than the exponents $h'(q)$, so we suggest to use the exponents $h(q)$.

3. Data, results and discussions

3.1. Large proteins

The methods introduced in the previous sections can only be used for long protein sequences (corresponding to the large proteins) as declared in Refs. [9,13]. The amino acid sequence of 49 large proteins which include the 43 proteins studied in Ref. [9] were selected from the RCSB Protein Data Bank [3]. These 49 proteins, which are listed in Table 1, belong to four structural classes according to the SCOP classifying standards [3]. The class information of some proteins is updated according to the Protein Data Bank comparing those given in Ref. [9].

3.2. Numerical results and discussion

Given an amino acid sequence of one protein, we first convert it into a numerical sequence according to the map (1). Then its measure representation μ_K with length K can be calculated. A question immediately arises: How to decide the value of K? If K is too small, there will not be enough combinations of length K from the set $\{0,1,2,3,4,5\}$, hence this would not yield reliable results in a statistical sense; if K is too large, the frequencies of most strings will be zero, and as a result, useful information related to the structure would not be gleaned from the measure representation. We calculate the dimension spectra of protein 1A8I for $K = 1,2,3,4,5,6$ and found that the D_q curves of $K = 4,5,6$ are very close to one another (see Fig. 4). Hence it seems appropriate to use the measure representation corresponding to $K = 5$. The D_q, τ_q, C_q curves and the multifractal spectrum $f(\alpha)$ of protein 1A8I for $K = 5$ are shown in Fig. 5. After deciding the value of K, we calculate D_q, $\tau(q)$, C_q, α and $f(\alpha)$ for the measures μ of the 49 selected proteins.

We then convert the amino acid sequences of proteins into theirs RSWH sequences according to the revised Schneider and Wrede hydrophobicity scale. We use such sequences to construct the measures ν. The ordinary multifractal analysis is then performed on these measures. The results are shown in Fig. 6.

After the times $N_K(r)$ of different K-string r in a protein sequence are counted, we get the corresponding frequency sequence $F_K(r)$ introduced in Section 2.1. We then perform the AMFA on it. Similarly, we perform AMFA on the RSWH sequences of selected proteins. The left panel of Fig. 7 gives an example in applying the AMFA to the RSWH sequence of protein 1A8I. It shows a good linear relationship between $\ln M(L)$ and $\ln(L)$. For different values of q, we get the exponents $h(q)$ from linear regressions of $\ln M(L)$ against $\ln(L)$ according to Eq. (18). The exponent curve $h(q)$ of protein 1A8I is shown in the right panel of Fig. 7.

3.3. Clustering of protein structures

For the structural classification problem of proteins, we consider nine parameters achieved from our multifractal analyses. These parameters are listed in Table 2. In the table, $\Delta\alpha$ represents the width of the multifractal spectrum $f(\alpha)$ [9,31]; q_0 is the value in $[-15,15]$ which corresponds to the maximum value of C_q. They are defined as

$$\Delta\alpha = \alpha_{\max} - \alpha_{\min}, \tag{19}$$

$$C_q \leqslant C_{q_0} \quad \forall q \in [-15,15]. \tag{20}$$

These nine parameters can be used as candidates to construct parameter spaces. In each parameter space, one point represents a protein. We want to determine whether proteins from the four structural classifications can be separated in these parameter spaces. We found that (i) in the 3D space formed by parameters 1, 5 and 8, the proteins from the α class group together and are separated from the proteins from the other classes, which is shown in Fig. 8; (ii) in the 3D space formed by parameters 5, 7 and 9, the proteins from the β class can be separated from the proteins from the $\alpha+\beta$ and α/β classes, which is shown in Fig. 9; (iii) in the 3D space formed by parameters 2, 6 and 9, the proteins from the α/β class group together and are separated from the proteins from the $\alpha+\beta$ class, which is shown in Fig. 10; (iv) in the 3D space formed by parameters 3, 4 and 8, the proteins from the α class form a group which can be separated from the proteins from the α/β class as shown in Fig. 11.

So we propose a procedure to cluster proteins which consists of three steps:

- Step 1: separating the α proteins from the $\{\beta, \alpha+\beta, \alpha/\beta\}$ proteins in the 3D space of parameters 1, 5 and 8;
- Step 2: separating the β proteins from the $\{\alpha+\beta, \alpha/\beta\}$ proteins in the 3D space of parameters 5, 7 and 9;
- Step 3: separating the α/β proteins from the $\alpha+\beta$ proteins in the 3D space of parameters 2, 6 and 9.

Table 1
Forty nine selected proteins represented by the PDB ID in the PDB database

Class	PDB ID	Protein	Length
α	1B89	Clethrin heavy chain	449
	1IAL	Importin alpha	453
	1HO8	Vacuolar ATP synthase subunit H	480
	1B8F	Histidine ammonia-lyase	509
	1DL2	Class I α-1,2-mannosidase	511
	2BCT	β-Catenin	516
	5EAS	5-Epi-aristolochene synthase	548
	1BKE	Serum albumin	581
	1BJ5	Human serum albumin	585
	1AVC	Annexin VI	673
	1ST6	Vinculin	1069
β	1A65	Laccase	504
	1A6C	Tobacco ringspot virus capsid protein	513
	1B9S	Neuramindase	390
	1DAB	P.69 pertactin	539
	1EUT	Sialidase	605
	1FNF	Fibronectin	368
	1C8F	Feline panleukopenia virus capsid	548
	1DBG	Chondroitinase B	506
	1DZL	Late major capsid protein L1	505
	1F1S	Hyaluronate lyase	814
	1KCW	Ceruloplasmin	1046
	1P2Z	Hexon protein	968
	1P30	Hexon protein	952
	1W0O	Sialidase	781
α + β	1DMT	Neutral endopeptidase	696
	1EWF	Bactericidal/permeability-increasing protein	456
	1OIE	Protein usha	532
	1W1O	Cytokinin dehydrogenase	534
	1USH	5′-Nucleotidase	550
	1AOP	Sulfite reductase hemoprotein	497
	1KA2	M32 carboxypeptidase	499
	1V0R	Phospholipase D	506
	5JDW	L-Arginine: glycine amidinotransferase	386
	1OY6	Acriflavine resistance protein B	1049
	1SIJ	Aldehyd oxidoreductase	907
	1T3T	Phosphoribosylformylglycinamidine synthase	1303
α/β	1A8I	Glycogen phosphorylase B	842
	1AOV	Apo-ovotransferin	686
	1BFD	Benzoylformate decarboxylase	528
α/β	1CRL	Lipase (triacylglycerol hydrolase)	534
	1AIV	Ovotransferrin	686
	1AK5	Inosine-5′-monophosphate dehydrogenase	503
	1AKN	Bile-salt activated lipase	579
	1AX9	Acetylcholinesterase	537
	1AXR	Glycogen phosphorylase	842
	1B1X	Lactoferrin	689
	1FA9	Glycogen phosphorylase, liver form	846
	1EJJ	Phosphoglycerate mutase	511

In order to give a quantitative assessment of our clustering on the 49 selected proteins, we use Fisher's linear discriminant algorithm [49–51] to calculate the discriminant accuracies of our method.

Fisher's discriminant algorithm is used to find a classifier in the parameter space for a training set. The given training set $H = \{x_1, x_2, \ldots, x_n\}$ is partitioned into $n_1 \leqslant n$ training vectors in a subset H_1 and $n_2 \leqslant n$ training vectors in a subset

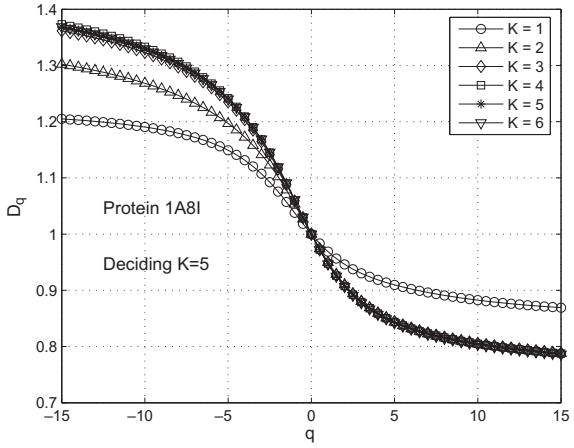

Fig. 4. Dimension spectra of measures of substrings with different lengths K in proteins 1A8I. The spectra for $K = 4, 5, 6$ are very close to one another, hence we decide that it is reasonable to set $K = 5$.

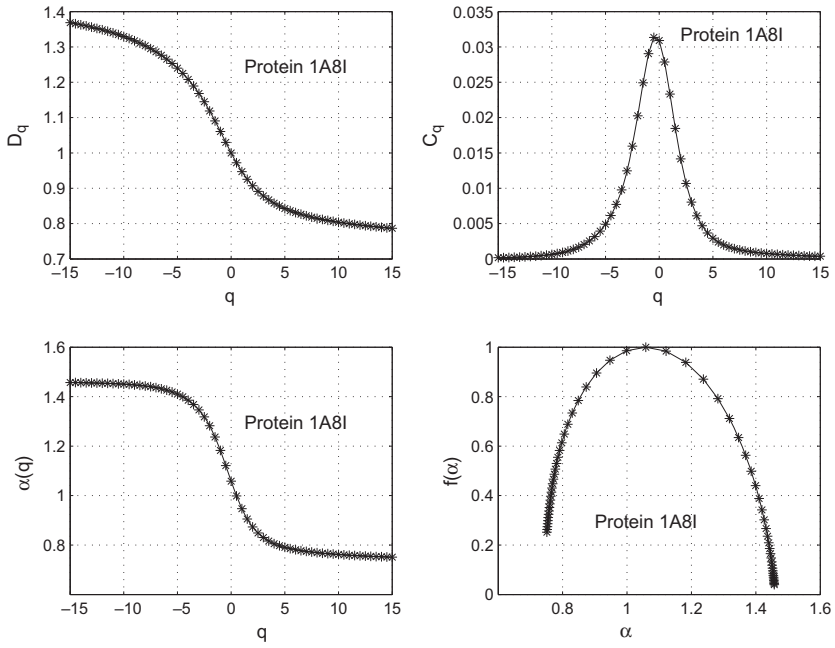

Fig. 5. The four kinds of curves for the measure representation of proteins 1A8I with $K = 5$.

H_2, where $n_1 + n_2 = n$ and each vector x_i is a point in the 3D parameter space. Then $H = H_1 \cup H_2$. We need to find a parameter vector $\mathbf{w} = (w_1, w_2, w_3)$ for the 3D space such that $\{y_i = \mathbf{w}^{\mathrm{T}} x_i\}_{i=1}^{n}$ can be classified into two classes in the space of real numbers. If we denote

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x_i} \in H_j} \mathbf{x_i}, \quad j = 1, 2, \tag{21}$$

$$\mathbf{S}_j = \sum_{\mathbf{x_i} \in H_j} (\mathbf{x_i} - \mathbf{m_j})(\mathbf{x_i} - \mathbf{m_j})^{\mathrm{T}}, \quad j = 1, 2, \tag{22}$$

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \tag{23}$$

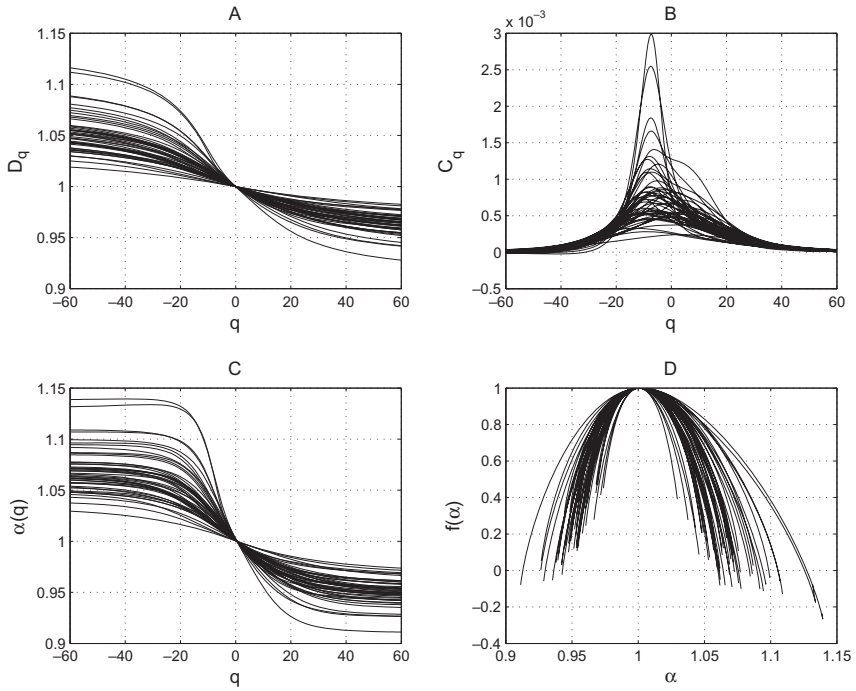

Fig. 6. The dimension spectra D_q (A), C_q curves (B), spectra of Lipschitz–Höder exponent $\alpha(q)$ (C) and the multifractal spectra $f(\alpha)$ (D) of RSWH sequences for the 49 selected proteins. In order to give a more clear image of the influence of q, here the range of q is $[-60, 60]$ rather than $[-15, 15]$.

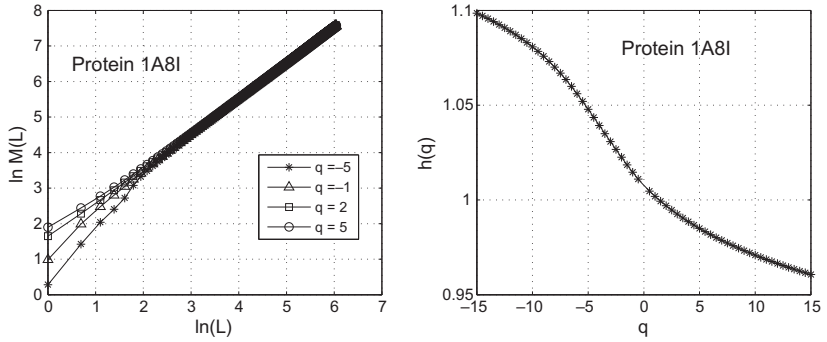

Fig. 7. The relationship between $\ln M(L)$ and $\ln(L)$ using the RSWH sequence (Left); the $h(q)$ spectra for the RSWH sequence of protein 1A8I calculated by AMFA (Right).

then the parameter vector $\mathbf{w}$ is estimated as $\mathbf{S}_w^{-1}(\mathbf{m}_1 - m_2)$ [50]. As a result, Fisher's discriminant rule becomes: assign $\mathbf{x}$ to H_1 if $(\mathbf{m_1} - \mathbf{m_2})^{\mathrm{T}}\mathbf{S}_w^{-1}\left[\mathbf{x} - \frac{1}{2}(\mathbf{m_1} + \mathbf{m_2})\right] > 0$ and to H_2 otherwise [49].

We use the whole data set as the training set because the selected protein data set is small. The discriminant accuracies for resubstitution analysis are defined as

$$P_{H_1} = \frac{n_{cH_1}}{n_1}, \tag{24}$$

$$P_{H_2} = \frac{n_{cH_2}}{n_2}, \tag{25}$$

where n_{cH_1} and n_{cH_2} denote the number of correctly discriminating H_1 elements and the number of correctly discriminating H_2 elements in the training set, respectively.

Table 2
The nine parameters from the calculating which are used into the structural classification problem of proteins

Order	Data	Method	Parameter
1	Measure representation of protein sequence	MFA	D_1
2	Measure representation of protein sequence	MFA	C_1
3	Measure representation of protein sequence	MFA	$\Delta\alpha$
4	Frequency of K-string in protein sequence	AMFA	$h(-1)$
5	Frequency of K-string in protein sequence	AFMA	$h(2)$
6	Measure of RSWH sequence	MFA	D_1
7	Measure of RSWH sequence	MFA	$\alpha_{\min}$
8	RSWH sequence	AMFA	$h(1)$
9	RSWH sequence	AMFA	$h(2)$

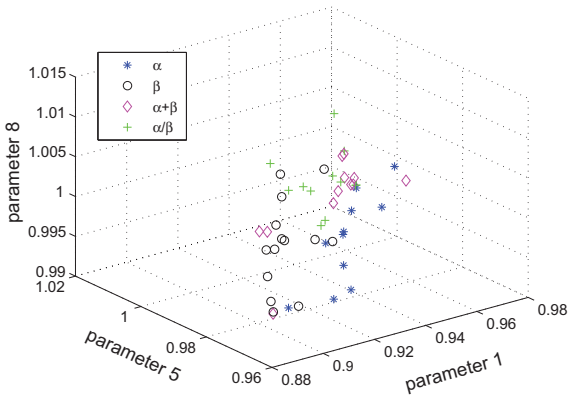

Fig. 8. The space (parameters 1, 5 and 8). In this space the α class proteins gather as a group and can be separated from the proteins from the other classes.

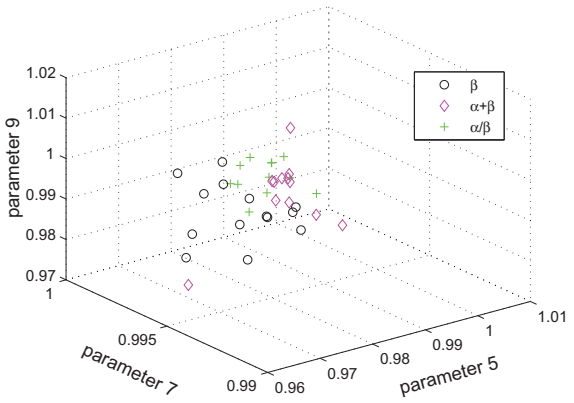

Fig. 9. The space (parameters 5, 7 and 9). In this space the β class form a group which can be separated from the proteins from the α + β and α/β classes.

We denote all the α proteins as H_2, the remaining β, α + β, α/β proteins as H_1 in the 3D space of parameters 1, 5 and 8; all the β proteins as H_2, the remaining α + β, α/β proteins as H_1 in the 3D space of parameters 5, 7 and 9; all the α/β proteins as H_2, all the α + β proteins as H_1 in the 3D space of parameters 2, 6 and 9; all the α proteins as H_2 and all the α/β proteins as H_1 in the 3D space of parameters 3, 4 and 8. The estimated parameters $\mathbf{w} = (w_1, w_2, w_3)$ in Fisher's dis-

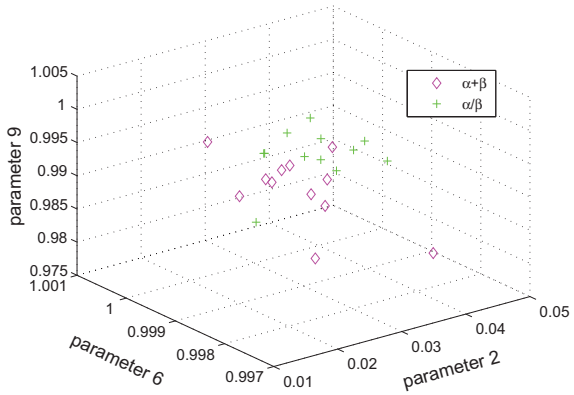

Fig. 10. The space (parameters 2, 6 and 9). In this space the proteins from the α/β class group together and are separated from the proteins from the α + β class.

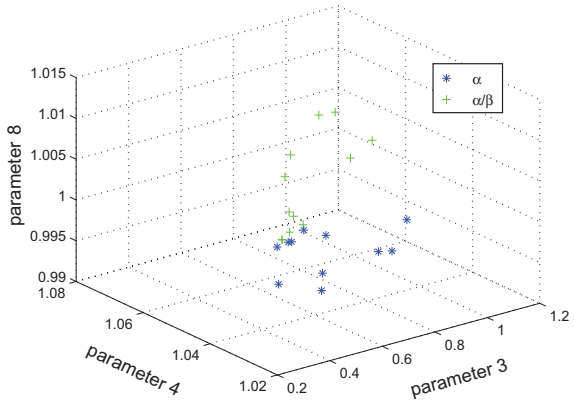

Fig. 11. The space (parameters 3, 4 and 8). In this space the proteins the α class group together and are separated from the proteins from α/β class.

criminant algorithm and the discriminant accuracies for proteins in parameter spaces shown in Figs. 8–11 are given in Table 3. From the discriminant accuracies, it is seen that our clustering is satisfactory.

The third step in our clustering procedure (i.e. to separate the α + β proteins from the α/β proteins) is the most difficult step compared with the first two steps. But it is also satisfactory when compared with some methods by other researchers [52–55]. They calculated only the fractal dimension of the protein 3D structures using the coordinates contained in the PDB database [3] and found that generally proteins from different structural classes had different fractal dimensions, which was then used in the clustering of proteins according to the four structural classes. But their methods turned out to be helpless in separating the α + β proteins from the α/β proteins as the fractal dimensions for these two structural classes were similar.

Table 3
The parameters in Fisher's discriminant and the discriminant accuracies for the selected proteins based on 6-letters model

Order	Proteins	w_1	w_2	w_3	P_{H_1} (%)	P_{H_2} (%)
1	In Fig. 8	2.9746	−5.9693	−7.4148	100.00	84.21
2	In Fig. 9	−4.4291	15.9781	−4.6495	92.86	86.96
3	In Fig. 10	−9.0266	−57.6575	−12.4420	91.67	83.33
4	In Fig. 11	0.3001	−11.7082	−24.9134	100.00	100.00

Table 4
The parameters in Fisher's discriminant and the discriminant accuracies for the selected proteins based on detailed HP model

Comparing	Proteins	w_1	w_2	w_3	P_{H_1} (%)	P_{H_2} (%)
1	In 3D space (parameters 1, 5 and 8)	1.0226	−1.2555	−5.1331	90.91	65.79
2	In 3D space (parameters 5, 7 and 9)	1.5097	−0.09414	−5.2220	64.29	73.91
3	In 3D space (parameters 2, 6 and 9)	−0.1498	−35.5716	−6.6599	58.33	75.00
4	In 3D space (parameters 3, 4 and 8)	0.0523	−8.3704	−15.8010	90.91	91.67

The number in the column identified by comparing is to compare the detailed HP model with the 6-letters model. One can return to Table 3 to compare the results in two different rows characterized by the same number in the first column.

Table 3 shows that separating the α proteins from the α/β proteins is very satisfactory achieving both 100% accuracies. This (i.e. separating the α proteins from the α/β proteins) has been reported to be a difficult task in Ref. [9], but it is quite satisfactory here. Therefore, we believe that our results are better than those in Ref. [9] in this aspect.

In order to demonstrate that the 6-letter model can give us more information than the detailed HP model, we perform our methods based on the detailed HP model and furthermore plotted the figures such as Figs. 8–11 using the parameters based on the detailed HP model. From these figures, we can see the clustering of the protein structures is much worse than those in Figs. 8–11. We considered five-strings in the detailed HP model as the same as that used by Yu et al. [9]. Fisher's discriminant algorithm is also used to give us quantitative comparison. The results are shown in Table 4 for all the four situations. From Tables 3 and 4, we see that the 6-letter model yields a better clustering than the detailed HP model.

In RSWH, we add a common value 12.30 to the Schneider and Wrede scale to make all the 20 values non-negative. We use δ to represent such a common value. One may ask how to choose the value of δ? If we choose different values for δ, will they affect the curves D_q, C_q, $\alpha(q)$, $f(\alpha)$ and $h(q)$? Furthermore, will they affect the clustering results? It is easy to see that the values of δ have no effect on the values of $h(q)$ from Eqs. (14), (16) and (18). The effect of δ on D_q, C_q and $f(\alpha)$ can be seen from Fig. 12, which indicates that the multifractality is not obvious when δ is larger. Consequently, it might affect our results. The results in the first step of the classifying procedure do not change due to different values of

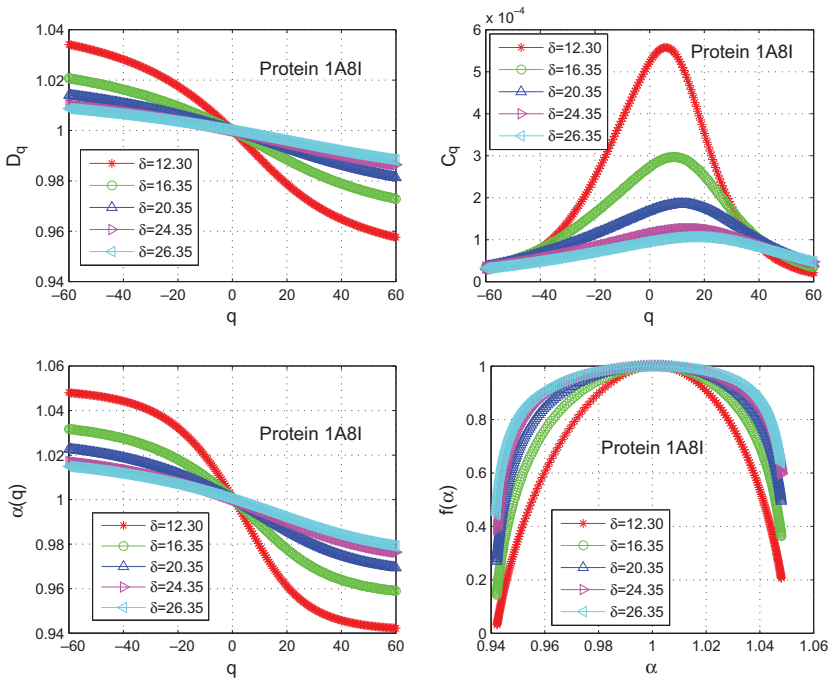

Fig. 12. The influence of δ on D_q, C_q, $\alpha(q)$ and $f(\alpha)$ curves. As δ increases, the multifractality is not obvious because the curves become flatter for larger value of δ.

δ because the parameters are not affected by δ. We checked the effect of δ in the second step. The classifying accuracies are all 85.71% and 91.30% for the three conditions: $\delta = 20.35$, 50.35 and 100.35 in Step 2. They are not as good as those corresponding to $\delta = 12.30$. We also set $\delta = 12.31$, 12.35 which are a little bit larger than 12.30 (the smallest hydrophobicity value of residues is -12.30); we find that the results are the same. So $\delta = 12.30$ is a suitable choice.

4. Conclusions

The ordinary and analogous multifractal analyses of both measure representation and the RSWH sequence of proteins provide useful information and visualization of their secondary structure classification. If a protein sequence is completely random, then the measure representation yields a uniform measure. From the measure representation and the curves D_q, C_q, $\alpha(q)$, $f(\alpha)$ and $h(q)$, it is seen that there is a clear difference between the protein sequences considered here and a completely random sequence. Hence we can conclude that these protein sequences possess correlations. In fact, it has been previously recognized that a protein sequence is not a completely random sequence (for example, see Pande et al. [56]).

From the D_q curves, it is seen that they are multifractal-like and sufficiently smooth so that the C_q curves can be meaningfully estimated. The C_q curves resemble a classical phase transition at a critical point, while the $f(\alpha)$ and $h(q)$ curves indicate that the hydrophobicity displays multifractal scaling.

Some parameter spaces can be constructed using the parameters from the ordinary and analogous multifractal analyses to distinguish and cluster proteins. Each protein can be represented by a point in these spaces. Numerical results indicate that the α proteins can be separated from the $\{\beta, \alpha+\beta, \alpha/\beta\}$ proteins in the 3D space of parameters 1, 5 and 8. Then the β proteins can be separated from the $\{\alpha+\beta, \alpha/\beta\}$ proteins in the 3D space of parameters 5, 7 and 9, and finally the α/β proteins from the $\alpha+\beta$ proteins in the 3D space of parameters 2, 6 and 9. Fisher's linear discriminant algorithm is used to give a quantitative assessment of our clustering of the 49 selected proteins. The discriminant accuracies are satisfactory. In particular, they reach 100.00% and 84.21% in separating the α proteins from the $\{\beta, \alpha+\beta, \alpha/\beta\}$ proteins in the 3D space of parameters 1, 5 and 8.

Compared with the results in Refs. [52–55], our methods are better in the sense that they give higher accuracies in distinguishing the proteins of the α/β class from the $\alpha+\beta$ class. Also, the 6-letter model of protein contains much more information than the detailed HP model [9,13] and the comparison between the results from the two models shows that the former is more helpful in the clustering of protein structures.

These two kinds of multifractal analyses are useful in classifying proteins. Our clustering algorithm is fast and can be evaluated in many combinations if more large proteins are available in the protein database. Once validated, it is easy to be used to perform the secondary structural classification of a protein.

The global mapping of protein structures into some spaces was reported by Hou et al. [7,8]. Our clustering method can also be regarded as a global mapping of protein structures into parameter spaces. This method of protein structure classification seems capable of yielding useful results.

Acknowledgements

Financial support was provided by the Chinese National Natural Science Foundation (Grant No. 30570426), Fok Ying Tung Education Foundation (Grant No. 101004) and the Youth Foundation of Educational Department of Hunan Province in China (Grant No. 05B007) (Z.-G. Yu) and the Australian Research Council (Grant No. DP0559807) (V.V. Anh).

References

[1] Anfinsen C. Science 1973;181:223.
[2] Levitt M, Chothia C. Nature 1976;261:552.
[3] http://www.rcsb.org/pdb/index.html.
[4] Rchardson JS. Adv Protein Chem 1981;34:167.
[5] Chothia C. Nature 1992;357:543.
[6] Zhang C, DeLisi C. J Mol Biol 1998;284:1301.
[7] Hou J, Jun S-R, Zhang C, Kim S-H. Proc Natl Acad Sci USA 2005;102:3651.
[8] Hou J, Gregory ES, Zhang C, Kim S-H. Proc Natl Acad Sci USA 2003;100:2386.

[9] Yu ZG, Anh V, Lau KS, Zhou LQ. Phys Rev E 2006;73:031920.
[10] Dill KA. Biochemistry 1985;24:1501.
[11] Wang J, Wang W. Phys Rev E 2000;61:6981.
[12] Brown TA. Genetics. third ed. London: Chapman & Hall; 1998.
[13] Yu ZG, Anh V, Lau KS. Physica A 2004;337:171.
[14] Chou PY, Fasman GD. Biochemistry 1974;13:222.
[15] Kanzmann W. Adv Protein Chem 1959;14:1.
[16] Kyte J, Doolittle RF. J Mol Biol 1982;157:105.
[17] Cornette JL, Cease KB, Margulit H, Spouge JL, Berzofsky JA, Delisi C. J Mol Biol 1987;195:659.
[18] Cid H, Bunster M, Canales M, Gaziua F. Protein Eng 1992;5:373.
[19] Pslliser C, Parry AD. Protein: Struct Funct Genet 2001;42:243.
[20] Zbiult JP, Giuliani A, Webber CL, Colosimo A. Protein Eng 1998;11:87.
[21] Giuliani A, Benigni R, Sirabella P, Zbiult JP, Colosimo A. Biophys J 2000;78:136.
[22] Giuliani A, Tomasi M. Protein: Struct Funct Genet 2002;46:171.
[23] Giuliani A, Benigni R, Zbiult JP, Webber CL, Sirabella P, Colosimo A. Chem Rev 2002;102:1471.
[24] Huang YZ, Xiao Y. Chaos, Solitons, & Fractals 2003;17:895.
[25] Huang YZ, Li MF, Xiao Y. Chaos, Solitons, & Fractals 2007;34:782.
[26] Mandelbrot BB. The fractal geometry of nature. New York: Academic Press; 1983.
[27] Balafas JS, Dewey TG. Phys Rev E 1995;52:880.
[28] Enright MB, Leitner DM. Phys Rev E 2005;71:011912.
[29] Anh V, Lau KS, Yu ZG. J Phys A: Math Gen 2001;34:7127.
[30] Li H, Ding ZJ, Wu ZQ. Phys Rev B 1995;51:554.
[31] Sun X, Xiong G, Wu ZQ. Acta Physica Sinica 2000;49:854.
[32] Zhou LQ, Yu ZG, Deng JQ, Anh V, Long SC. J Theor Biol 2005;232:559.
[33] Tian YC, Yu ZG, Flidge C. Physics Letters A 2007;361:103.
[34] Han LX, Cen ZW, Chu CB, Gau CS. Chaos, Solitons, & Fractals 2002;13:507.
[35] Li XW, Shang PJ. Chaos, Solitons, & Fractals 2007;31:1089.
[36] Ma D, Wu M, Liu C. Chaos, Solitons, & Fractals 2008;38:840–51.
[37] Bunde A, Havlin S, editors. Fractals in science. Berlin: Springer-Verlag; 1979.
[38] Nakaoa H, Kuramoto Y. Eur Phys J B 1999;11:345.
[39] Dasgupta C, Kim JM, Dutta M, Sarma SD. Phys Rev E 1997;55:2235.
[40] Katsuragi H. Physica A 2000;278:275.
[41] Yu ZG, Anh V, Lau KS. Physica A 2001;301:351.
[42] Yu ZG, Anh V, Lau KS. Phys Rev E 2001;64:031903.
[43] Feder J. Fractals. New York: Plenum; 1988.
[44] Lee J, Stanley HE. Phys Rev Lett 1988;61:2945.
[45] Canessa E. J Phys A 2000;33:3637.
[46] Dunki RM, Keller E, Meier PF, Ambuhl B. Physica A 2000;276:596.
[47] Dunki RM, Ambuhl B. Physica A 1996;230:544.
[48] Yu ZG, Anh V, Wang B. Phys Rev E 2001;63:011903.
[49] Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press; 1979.
[50] Duda RO, Hart PE, Stork DG. Pattern classification. second ed. New York: John Wiley & Sons; 2001.
[51] Sneath PHA, Sokal RR. Numerical taxonomy. San Francisco: Freeman; 1973.
[52] Torrens F. Molecules 2002;7:26.
[53] Isogai Y, Itoh T. J Phys Soc Jpn 1984;53:2162.
[54] Wang CX, Shi YY, Huang FH. Phys Rev A 1990;41:7043.
[55] Daniel M, Baskar S, Latha MM. Phys Scripta 1999;60:270.
[56] Pande VS, Grosberg AY, Tanaka T. Proc Natl Acad Sci USA 1994;91:12972.