Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins

Z. G. Yu,[1,2,*] V. V. Anh,[1] K. S. Lau,[3] and L. Q. Zhou[2]
[1]*Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia*
[2]*School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China*
[3]*Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong, China*
(Received 22 November 2005; revised manuscript received 18 January 2006; published 21 March 2006)

The hydrophobic free energy and solvent accessibility of amino acids are used to study the relationship between the primary structure and structural classification of large proteins. A measure representation and a Z curve representation of protein sequences are proposed. Fractal analysis of the measure and Z curve representations of proteins and multifractal analysis of their hydrophobic free energy and solvent accessibility sequences indicate that the protein sequences possess correlations and multifractal scaling. The parameters from the fractal and multifractal analyses on these sequences are used to construct some parameter spaces. Each protein is represented by a point in these spaces. A method is proposed to distinguish and cluster proteins from the α, β, $\alpha+\beta$, and α/β structural classes in these parameter spaces. Fisher's linear discriminant algorithm is used to give a quantitative assessment of our clustering on the selected proteins. Numerical results indicate that the discriminant accuracies are satisfactory. In particular, they reach 94.12% and 88.89% in separating β proteins from $\{\alpha, \alpha+\beta, \alpha/\beta\}$ proteins in a three-dimensional space.

DOI: 10.1103/PhysRevE.73.031920 PACS number(s): 87.10+e, 47.53+n

I. INTRODUCTION

The molecular function of a protein can be inferred from its structure information [1]. The three-dimensional (3D) structure of a protein is determined by its amino acid sequence via the process of protein folding [2–5]. The prediction of protein structure and function from amino acid sequences is one of the most important problems in molecular biology. There were some arguments that protein structures could not be accurately predicted directly from sequences [6]. On the other hand, protein secondary structure, which is a summary of the general conformation and hydrogen bonding pattern of the amino acid backbone [7,8], provides some knowledge to further simplify the complicated 3D structure prediction problem. Hence an intermediate but useful step is to predict the protein secondary structure. Since the 1970s, many methods have been developed for predicting protein secondary structure such as neural networks [9], hidden Markov models [10], multiple sequence alignments [11], advanced machine learning techniques [12], and support vector machines [6]. More recent references in this field are provided by Adamczak *et al.* [13].

Based on their secondary structures, proteins are known to group into four main classes: the α helices, the β strands, and those with a mixture of α and β shapes called $\alpha+\beta$ and α/β. In fact, Hou *et al.* [1] (see also a short report recently published in *Science* [14]) constructed a map of the protein structure space using the pairwise structural similarity of 1898 protein chains. They found that the space has a defining feature showing these four classes clustered together as four elongated arms emerging from a common center. In this paper, we aim to identify certain parameters that are characteristic of these four classes, and develop tools to estimate these parameters, which then form certain parameter spaces of protein structures. These tools, which are based on the concepts of fractal geometry and multifractal analysis, are capable of distinguishing proteins in these classes. The parameters are obtained from the detailed hydrophobic-polar (HP) model of protein behavior, the hydrophobic free energy of amino acids, and the solvent accessibility of the side chain of a protein. The latter two parameters are two chemical properties related to the protein folding process according to physical and chemical principles. These are two significant parameters which provide useful information toward the protein folding problem. Researchers are still looking for the universal architectural logic hidden in protein sequences.

A simplified but well-known model of protein behavior is the HP model proposed by Dill [15]. In this model, 20 kinds of amino acids are divided into two types, hydrophobic (H) (or nonpolar) and polar (P) (or hydrophilic). By studying the model on lattices, Li *et al.* [16] found that there are a 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 proteinlike secondary structures [5,16–18]. But the HP model lacks sufficient information on the heterogeneity and the complexity of the natural set of residues [19]. According to Brown [20], the polar class of the HP model can be divided into three subclasses: positive, uncharged, and negative polar. So the 20 different kinds of amino acids can be divided into four classes: nonpolar, negative polar, uncharged polar, and positive polar. In this model, more details can be considered than in the HP model. We call this model a *detailed HP model* [21]. In this paper, information on the

*Corresponding author. Email address: yuzg1970@yahoo.com; z.yu@qut.edu.au

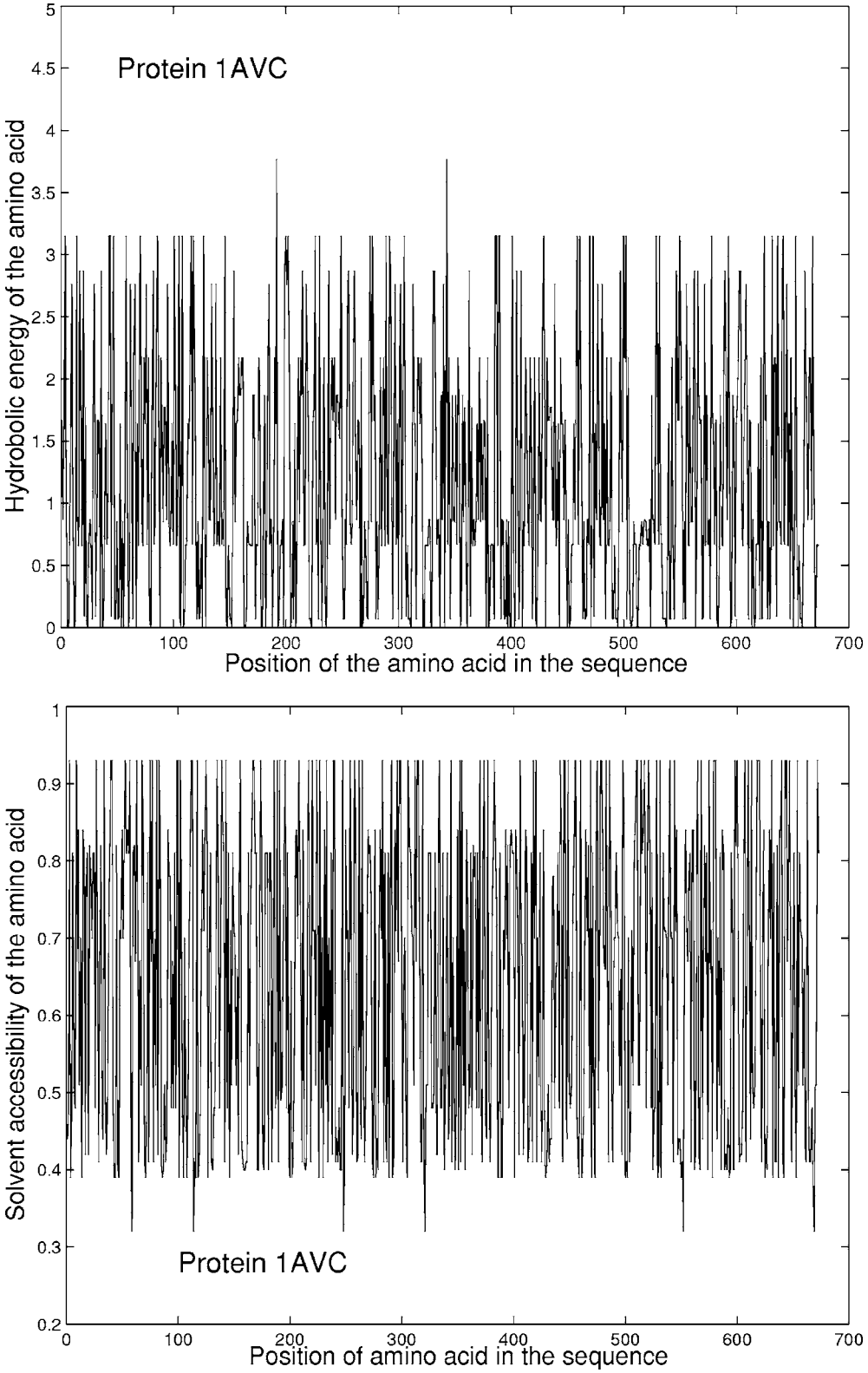

FIG. 1. The hydrophobic free energy sequence (top) and the solvent accessibility sequence (bottom) of the protein annexin VI.

detailed HP model is used to construct the measure representation and the Z curve representation of a protein.

The hydrophobic free energy of amino acids has been used to study protein structure via wavelet analysis [22–25]. Measured in kcal/mol, the hydrophobic free energies of the 20 amino acids are A=0.87, R=0.85, N=0.09, D=0.66, C =1.52, Q=0.0, E=0.67, G=0.0, H=0.87, I=3.15, L=2.17, K=1.65, M=1.67, F=2.87, P=2.77, S=0.07, T=0.07, W=3.77, Y=2.76, and V=1.87 [24]. For example, we give the hydrophobic free energy sequence of the protein annexin VI [Protein Data Base (PDB) ID: 1AVC] in the top panel of Fig. 1.

The solvent accessibility of the side chain of a protein has also been used to study the secondary structure prediction

[13,26] and structural classification of proteins [27]. The solvent accessibilities for solvent exposed area larger than 30 Å^2 are S=0.70, T=0.71, A=0.48, G=0.51, P=0.78, C=0.32, D=0.81, E=0.93, Q=0.81, N=0.82, L=0.41, I=0.39, V=0.40, M=0.44, F=0.42, Y=0.67, W=0.49, K=0.93, R=0.84, and H=0.66 [28]. The solvent accessibility sequence of the protein annexin VI is shown in the bottom panel of Fig. 1 as an example.

Fractal geometry provides a mathematical formalism for describing complex spatial and dynamical structures [29,30]. Multifractal analysis is a useful way to characterize the spatial heterogeneity of both theoretical and experimental fractal patterns [31]. In recent years it has been applied successfully in many different fields including time series analysis and financial modeling [32]. Some applications of fractal methods to DNA sequences are provided in [32–34] and the references therein.

Fractal methods have also been used to study proteins. These include fractal analysis of the proton-exchange kinetics [35], chaos game representation of protein structures [36] and sequences based on the detailed HP model [37], multifractal analysis of the measure representation of protein sequences [38], fractal dimension of protein mass [39], and fractal properties of protein chains [40]. But there has not been much work related to the secondary structure using fractal methods. We have found only a few existing studies: simulation of the measure representation of proteins using iterated function systems [21], and multifractal analysis of the solvent accessibility of protein [27] and the fractal dimensions of protein secondary structure elements [41]. In this paper, we will obtain certain parameters from fractal analysis of the hydrophobic free energy and solvent accessibility sequences, and use them to construct some parameter spaces. Each protein is represented by a point in these spaces. A method is proposed to distinguish proteins from the α, β, $\alpha+\beta$, and α/β structural classes in these parameter spaces.

II. DETAILED HP MODEL AND MEASURE REPRESENTATION OF PROTEIN SEQUENCES

A. Measure representation

We first outline the definition of the detailed HP model proposed in our previous work [21]. Twenty different kinds of amino acids are found in proteins. In the detailed HP model, they are divided into four classes: nonpolar, negative polar, uncharged polar, and positive polar. The nonpolar class consists of the eight residues ALA, ILE, LEU, MET, PHE, PRO, TRP, VAL; the negative polar class consists of the two residues ASP, GLU; the uncharged polar class is made up of the seven residues ASN, CYS, GLN, GLY, SER, THR, TYR; and the remaining three residues ARG, HIS, LYS constitute 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 20 kinds of amino acids for $i=1,\ldots,L$, we define

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

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

We call any string made of K letters from the set $\{0,1,2,3\}$ a K string. For a given K, there are in total 4^K different K strings. In order to count the number of K strings in a sequence $X(s)$ from a protein sequence s, 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 $r=r_1\cdots r_K$, $r_i\in\{0,1,2,3\}$, $i=1,\ldots,K$, be a substring with length K, we define

$$x_{left}(r)=\sum_{i=1}^{K}\frac{r_i}{4^i} \tag{2}$$

and

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

We then use the subinterval $[x_{left}(r),x_{right}(r)[$ 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

$$F_K(r)=N_K(r)/(L-K+1) \tag{4}$$

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

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

It is seen that $\int_0^1 d\mu_K(x)=1$ and $\mu_K\{[x_{left}(r),x_{right}(r)[\}=F_K(r)$. 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 proteins human serum albumin (1BJ5), sialidase (1EUT), neutral endopeptidase (1DMT), and apo-ovotransferin (1AOV) for $K=4$ and 5 are given in Fig. 2.

B. Iterated function system model

In order to simulate the measure representation of a protein sequence, we propose to use the iterated function system (IFS) model (see [21], [32], and [42]). IFS is the name given by Barnsley and Demko [43] originally to a system of contractive maps $w=\{w_1,w_2,\ldots,w_N\}$. 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

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

Then $E=\cap_{n=1}^{\infty}E_n$ is called the *attractor* of the IFS. The attractor is usually a fractal set and the IFS is a relatively

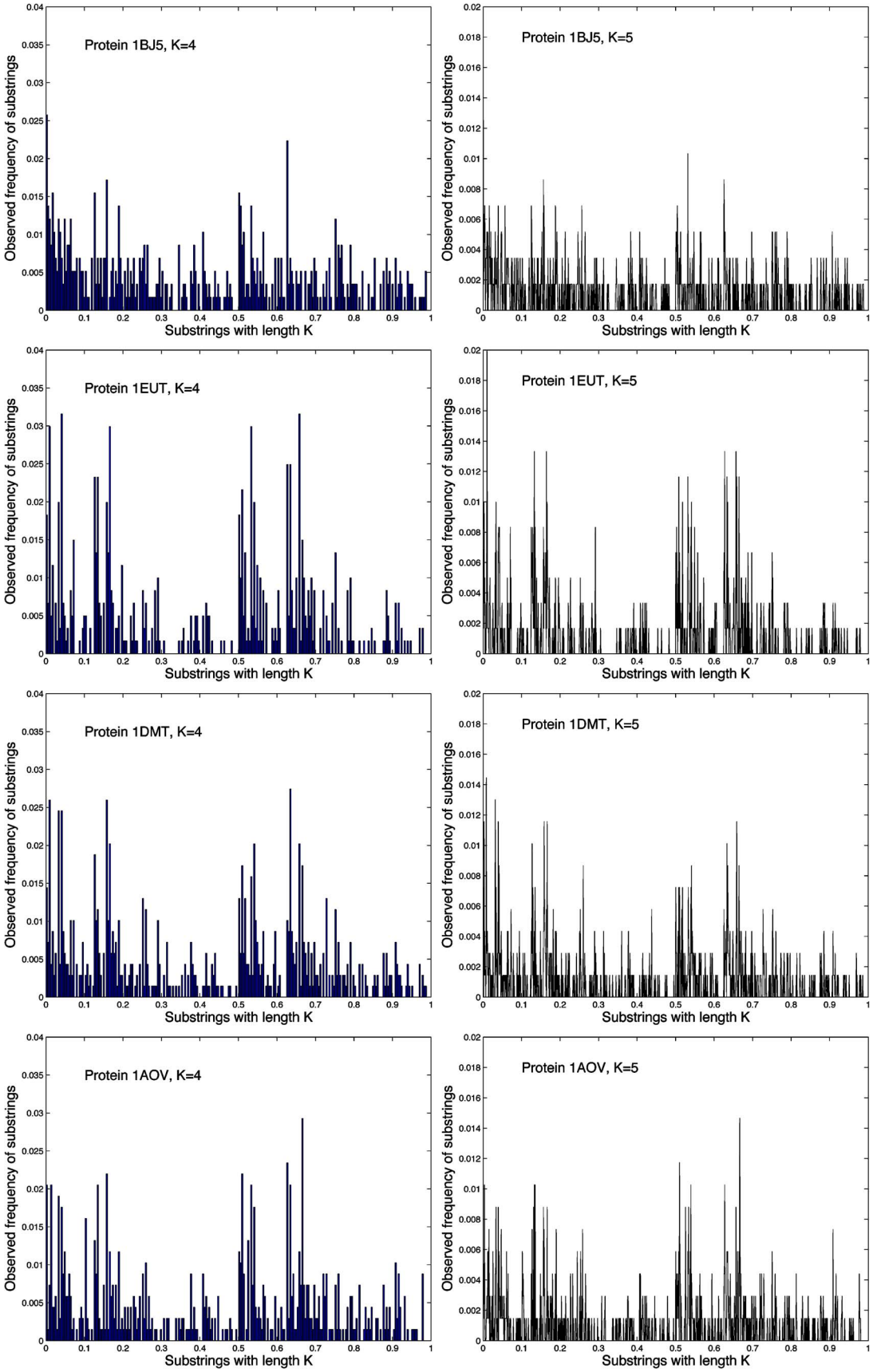

FIG. 2. (Color online) The histograms of measure representation for proteins human serum albumin, sialidase, neutral endopeptidase, and apo-ovotransferin for $K=4$ (left) and 5 (right).

general model to generate many well-known fractal sets such as the Cantor set and the Koch curve. Given a set of probabilities $P_i>0$, $\Sigma_{i=1}^{N}P_i=1$, pick an $x_0 \in E$ and define the iteration sequence

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

where the indices σ_n are chosen randomly and independently from the set $\{1,2,\ldots,N\}$ with probabilities $P(\sigma_n=i)=P_i$. Then every orbit $\{x_n\}$ is dense in the attractor E [43]. For n large enough, we can view the orbit $\{x_0,x_1,\ldots,x_n\}$ as an approximation of E. This process is called the chaos game.

Let μ be the invariant measure on the attractor of the IFS and χ_B the characteristic function for the Borel subset $B \subset E$; then from the ergodic theorem for the IFS [43]

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

In other words, $\mu(B)$ 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.

C. Moment method to estimate the parameters in the IFS model

The coefficients in the contractive maps and the probabilities in the IFS model are the parameters to be estimated for a real measure which we want to simulate. Vrscay [44] described a moment method to perform this task. If μ is the invariant measure and E the attractor of the IFS in $\mathbf{R}$, the moments of μ are

$$g_i = \int_E x^i d\mu, \quad g_0 = \int_E d\mu = 1. \tag{6}$$

If $w_i(x)=c_ix+d_i$, $i=1,\ldots,N$, then the following well-known recursion relations hold [44]:

$$\left(1-\sum_{i=1}^{N}P_ic_i^n\right)g_n = \sum_{j=1}^{n}\binom{n}{j}g_{n-j}\left(\sum_{i=1}^{N}P_ic_i^{n-j}d_i^j\right). \tag{7}$$

Thus, setting $g_0=1$, the moments g_n, $n\geqslant 1$, may be computed recursively from a knowledge of $g_0,\ldots,g_{n-1}$. If we denote by G_k the moments obtained directly from the real measure using (6) and g_k the formal expression of moments obtained from (7), then through solving the optimization problem

$$\min_{c_i,d_i,P_i}\sum_{k=1}^{n}(g_k-G_k)^2 \quad \text{for some chosen } n, \tag{8}$$

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

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

$$w_1(x) = x/4, \quad w_2(x) = x/4 + 1/4,$$

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

in the IFS model. For a given measure representation of a protein sequence, we obtain the estimated values of the probabilities P_1,P_2,P_3,P_4 by solving the optimization problem (8). Based on the estimated values of the probabilities, we can use the chaos game to generate a histogram approximation of the invariant measure of the IFS, which can be compared with the real measure representation of the protein sequence.

III. Z CURVE REPRESENTATION OF PROTEINS AND DETRENDED FLUCTUATION ANALYSIS

The concept of the Z curve representation of a DNA sequence was first proposed by Zhang and Zhang [45], and was used to distinguish coding and noncoding DNA sequences [46,47]. We propose a similar concept for proteins in the present paper. Once we get the sequence $X(s)=a_1\cdots a_L$ for a protein as in Sec. II A, where a_i is a letter of the alphabet $\{0,1,2,3\}$, we can define the *Z curve representation* of this protein as follows. This Z curve consists of a series of nodes Q_i, $i=0,1,\ldots,L$, whose coordinates are denoted by x_i, y_i, and z_i. These coordinates are defined as

$$x_i = 2(v_i^0 + v_i^2) - i,$$

$$y_i = 2(v_i^0 + v_i^1) - i,$$

$$z_i = 2(v_i^0 + v_i^3) - i, \quad i = 0,1,2,\ldots,L, \tag{9}$$

where v_i^0, v_i^1, v_i^2, v_i^3 denote the number of occurrences of the symbols $0,1,2,3$ in the prefix $a_1a_2\cdots a_i$ respectively, and $v_0^0=v_0^1=v_0^2=v_0^3=0$. The connection of nodes $Q_0,Q_1,\ldots,Q_L$ one by one by lines is defined as the *Z curve representation* of this protein. We then define

$$\Delta x_i = x_i - x_{i-1},$$

$$\Delta y_i = y_i - y_{i-1},$$

$$\Delta z_i = z_i - z_{i-1}, \quad i = 1,2,\ldots,L, \tag{10}$$

where Δx_i, Δy_i, and Δz_i can only have values 1 and -1. For example, we show the Z curve representation of the protein annexin VI in Fig. 3.

The exponent in a detrended fluctuation analysis can be used to characterize the correlation of a time series [33,48]. We view Δx_i, Δy_i, and Δz_i, $i=1,2,\ldots L$, as time series. We denote a time series by $F(t),t=1,\ldots,L$. First, the time series is integrated as $T(k)=\Sigma_{t=1}^{k}[F(t)-F_{av}]$, where F_{av} is the average over the whole time period. Next, the integrated time series is divided into boxes of equal length n. In each box of length n, a linear regression is fitted to the data by least squares, representing the trend in that box. The T coordinate of the straight line segments is denoted by $T_n(k)$. We then detrend the integrated time series $T(k)$ by subtracting the local trend $T_n(k)$ in each box. The root-mean-square fluctuation of this integrated and detrended time series is computed as

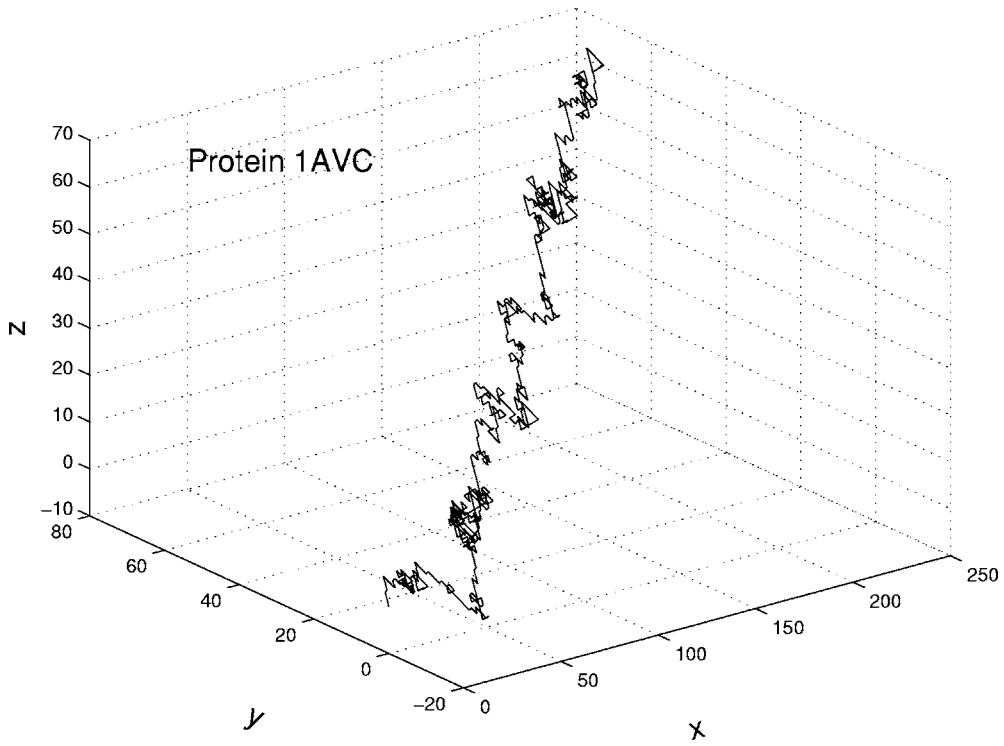

FIG. 3. The Z curve representation of the protein annexin VI.

$$\mathcal{F}(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[T(k) - T_n(k)]^2}. \tag{11}$$

Typically, $\mathcal{F}(n)$ increases with box size n. A linear relationship on a log-log graph indicates the presence of scaling

$$\mathcal{F}(n) \propto n^{\lambda}. \tag{12}$$

Under such conditions, the fluctuations can be characterized by the scaling exponent λ, the slope of the line relating $\ln \mathcal{F}(n)$ to $\ln n$. For uncorrelated data, the integrated value $T(k)$ corresponds to a random walk, and therefore $\lambda = 0.5$. A value of $0.5 < \lambda < 1.0$ indicates the presence of long memory so that, for example, a large value is likely to be followed by a large value. In contrast, the range $0 < \lambda < 0.5$ indicates a different type of power-law correlation such that positive and negative values of a time series are more likely to alternate. The exponents λ for the Δx_i, Δy_i, and Δz_i, $i = 1, 2, \ldots, L$, of the Z curve representation of DNA sequences were used to construct a parameter space to distinguish coding and noncoding sequences [46]. We consider the exponents λ for the Δx_i, Δy_i, and Δz_i, $i = 1, 2, \ldots, L$, of the Z curve representation of protein sequences as candidates to construct parameter spaces for proteins in this paper. These exponents are denoted by λ_x, λ_y, and λ_z, respectively.

IV. MULTIFRACTAL ANALYSIS OF HYDROPHOBIC FREE ENERGY AND SOLVENT ACCESSIBILITY OF PROTEINS

In this section, we view the hydrophobic free energy and solvent accessibility sequences of proteins as time series. Using a similar method to positive time series proposed in our previous paper [49], we can get a measure from the hydrophobic free energy sequence or solvent accessibility sequence of a protein. The most common algorithms of multifractal analysis are the so-called fixed-size box-counting algorithms [50]. In the one-dimensional case, for a given measure μ with support $E \subset \mathbf{R}$, we consider the partition sum

$$Z_\epsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q, \tag{13}$$

$q \in \mathbf{R}$, 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,

$$B = [k\epsilon, (k+1)\epsilon[. \tag{14}$$

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

$$\tau(q) = \lim_{\epsilon\to 0} \frac{\ln Z_\epsilon(q)}{\ln \epsilon} \tag{15}$$

and the generalized fractal dimensions of the measure are defined as

$$D_q = \tau(q)/(q-1) \quad \text{for } q \neq 1 \tag{16}$$

and

$$D_q = \lim_{\epsilon\to 0} \frac{Z_{1,\epsilon}}{\ln \epsilon} \quad \text{for } q = 1, \tag{17}$$

where $Z_{1,\epsilon} = \Sigma_{\mu(B)\neq 0}\mu(B)\ln \mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of $[\ln Z_\epsilon(q)]/(q-1)$ against $\ln \epsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\epsilon}$ against $\ln \epsilon$ for $q = 1$. The value D_1 is called the information dimension and D_2 the correlation dimension.

The concept of phase transitions 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 [51]. By following the thermodynamic formulation of multifractal measures, Canessa [52] derived an expression for the analogous specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \quad (18)$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. In a later section, we will discuss the property of C_q for measures from the hydrophobic free energy and solvent accessibility sequences of proteins.

The singularities of a measure are characterized by the Lipschitz-Hölder exponent α, which is related to $\tau(q)$ by

$$\alpha(q) = \frac{d}{dq}\tau(q). \quad (19)$$

Substitution of Eq. (15) into Eq. (19) yields

$$\alpha(q) = \lim_{\epsilon \to 0} \frac{\sum_{\mu(B)\neq 0} [\mu(B)]^q \ln \mu(B)}{Z_\epsilon(q) \ln \epsilon}. \quad (20)$$

Again the exponent $\alpha(q)$ can be estimated through a linear regression of $\Sigma_{\mu(B)\neq 0}[\mu(B)]^q \ln \mu(B)/Z_\epsilon(q)$ against $\ln \epsilon$ [27]. And the multifractal spectrum $f(\alpha)$ versus α can be calculated according to the relationship

$$f(\alpha) = q\alpha(q) - \tau(q). \quad (21)$$

V. RESULTS AND DISCUSSION

The methods introduced in the previous sections can only be used for long protein sequences (corresponding to large proteins). The amino acid sequences of 43 large proteins were selected from the RCSB Protein Data Bank [58]. These 43 proteins, which are listed in Table I, belong to four structural classes [53] according to their secondary structures.

First, we converted the amino acid sequences of these proteins into their measure representations with $K=5$ according to the method of Sec. II A. If K is too small, there will not be enough combinations of length K from the set $\{0,1,2,3\}$, hence this would not yield reliable results in a statistical sense. If K is too large, the frequencies of most substrings will be zero and, as a result, useful biological information would not be gleaned from the measure representation. Although the difference between the histograms of four-strings and five-strings is not significant (as shown in Fig. 2), on balance we selected $K=5$ with the view that consideration of more sample points in the measure representation is better in the statistical sense, and a length of less than 1000 of the selected proteins would not allow a higher value of K. We found that the IFS model corresponding to $K=5$ is a good model to simulate the measure representation of protein sequences, and the estimated value of the probability P_1 from the IFS model contains information useful for the secondary structural classification of proteins [21]. We performed an IFS simulation for the proteins selected and adopted the estimated parameter P_1 as one parameter to construct the parameter space for proteins.

Second, we converted the amino acid sequences of these proteins to their Z curve representations and performed their detrended fluctuation analysis. The exponents λ_x, λ_y, and λ_z were estimated and used as candidate parameters to construct the parameter space.

Third, the generalized fractal dimensions D_q and the related spectra C_q, multifractal spectra $f(\alpha)$ of hydrophobic free energy sequences and solvent accessibility sequences of all 43 proteins were computed. For examples, the D_q curves for the hydrophobic free energy sequences of four proteins are shown in the top panel of Fig. 4 and their related C_q curves are shown in the bottom panel of Fig. 4; the multifractal spectra $f(\alpha)$ for the hydrophobic free energy sequences and solvent accessibility sequences of the four proteins are shown in Fig. 5.

Last, for the structural classification problem of proteins, we consider the following parameters: P_1 from the IFS estimations of the measure representations; the exponents λ_x, λ_y, λ_z from the detrended fluctuation analysis of the Z curve representations; the range of D_q (that is, the value $D_{-15} - D_{15}$ in our frame); the maximum value of C_q (denoted $C_{max\ q}$); the value q_0 of q which corresponds to the maximum value of C_q; the maximum value of α (denoted α_{max}), the minimum value of α (denoted α_{min}), and $\Delta\alpha$ (defined by $\alpha_{max} - \alpha_{min}$) from the multifracal analysis of the hydrophobic free energy (HE) sequences and solvent accessibility (SA) sequences of proteins as candidates to construct parameter spaces. In a parameter space, one point represents a protein. We want to determine whether the proteins can be separated from four structural classifications in these parameter spaces. We found that in the 2D space (q_0 for HE, P_1) and the 3D space (q_0 for HE, P_1, $C_{max\ q}$ for SA), the proteins from the β class group together and are separated from the proteins from the other classes. These results are shown in Figs. 6 and 7. Then, in the 3D space ($\Delta\alpha$ for SA, P_1, λ_z for the Z curve), the proteins from the $\alpha+\beta$ class form a group which can be separated from the proteins from the α and α/β classes as shown in Fig. 8. Finally, in the 3D space (α_{max} for SA, α_{min} for SA, P_1), the proteins from the α and α/β classes can be separated as shown in Fig. 9.

So we propose a method which consists of the following three components to cluster proteins: (i) separating β proteins from $\{\alpha, \alpha+\beta, \alpha/\beta\}$ proteins in the 2D space (q_0 for HE, P_1) and the 3D space (q_0 for HE, P_1, $C_{max\ q}$ on SA); (ii) separating $\alpha+\beta$ proteins from $\{\alpha, \alpha/\beta\}$ proteins in the 3D space ($\Delta\alpha$ for SA, P_1, λ_z for the Z curve); (iii) separating α proteins from α/β proteins in the 3D space (α_{max} for SA, $\alpha_{min}$for SA, P_1).

In order to give a quantitative assessment of our clustering on the selected proteins, we use Fisher's linear discriminant algorithm [54–56] 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

TABLE I. Properties of the 43 proteins selected.

Class	PDB ID	Protein	length
α	1AVC	Annexin VI	673
	1B89	Clethrin heavy chain	449
	1BJ5	Human serum albumin	585
	1HO8	Vacuolar ATP synthase subunit H	480
	1IAL	Importin alpha	453
	1QSA	Soluble lytic transglycosylase SH70	618
	2BCT	β-catenin	516
	5EAS	5-epi-aristolochene synthase	548
	1A65	Laccase	504
	1A6C	Tobacco ringspot virus capsid protein	513
	1B8F	Histidine ammonia-lyase	509
	1BKE	Serum albumin	581
	1DL2	Class I α-1,2-mannosidase	511
β	1B9S	Neuramindase	390
	1DAB	P.69 pertactin	539
	1EUT	Sialidase	605
	1FNF	Fibronectin	368
	1JX5	Integrin α-Iib	452
	1MAL	Maltoporin	421
	1C8F	Feline panleukopenia virus capsid	548
	1DBG	Chondroitinase B	506
	1DZL	Late major capsid protein L1	505
$\alpha+\beta$	1B90	β-amylase	516
	1BBU	Lysyl-tRNA synthetase	504
	1BYT	Lioxygenase-3	857
	1CLC	Endoglaeanase celd	639
	1E7U	Phosphatidylinositol 3-kinase catalytic subunit	961
	1DMT	Neutral endopeptidase	696
	1EWF	Bactericidal/permeability-increasing protein	456
	1BP1	Bactericidal/permeability-increasing protein	456
	1KA2	M32 carboxypeptidase	499
α/β	1A8I	Glycogen phosphorylase B	841
	1ACJ	Acetylcholinesterase complexed with tacrine	537
	1AOV	Apo-ovotransferin	686
	1BFD	Benzoylformate decarboxylase	528
	1CRL	Lipase (Triacylglycerol hydrolase)	534
	1ACL	Acetylcholinesterase complexed with decamethonium	537
	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

set $H=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ is partitioned into $n_1 \leqslant n$ training vectors in a subset H_1 and $n_2 \leqslant n$ training vectors in a subset H_2, where $n_1+n_2=n$ and each vector $\mathbf{x}_i$ is a point in the 2D or 3D parameter space. Then $H=H_1 \cup H_2$. We need to find a parameter vector $\mathbf{w}=(w_1,w_2)$ for the 2D space and $\mathbf{w}=(w_1,w_2,w_3)$ for the 3D space such that $\{y_i=\mathbf{w}^T\mathbf{x}_i\}_{i=1}^n$ can be classified into two classes in the space of real numbers. If we denote

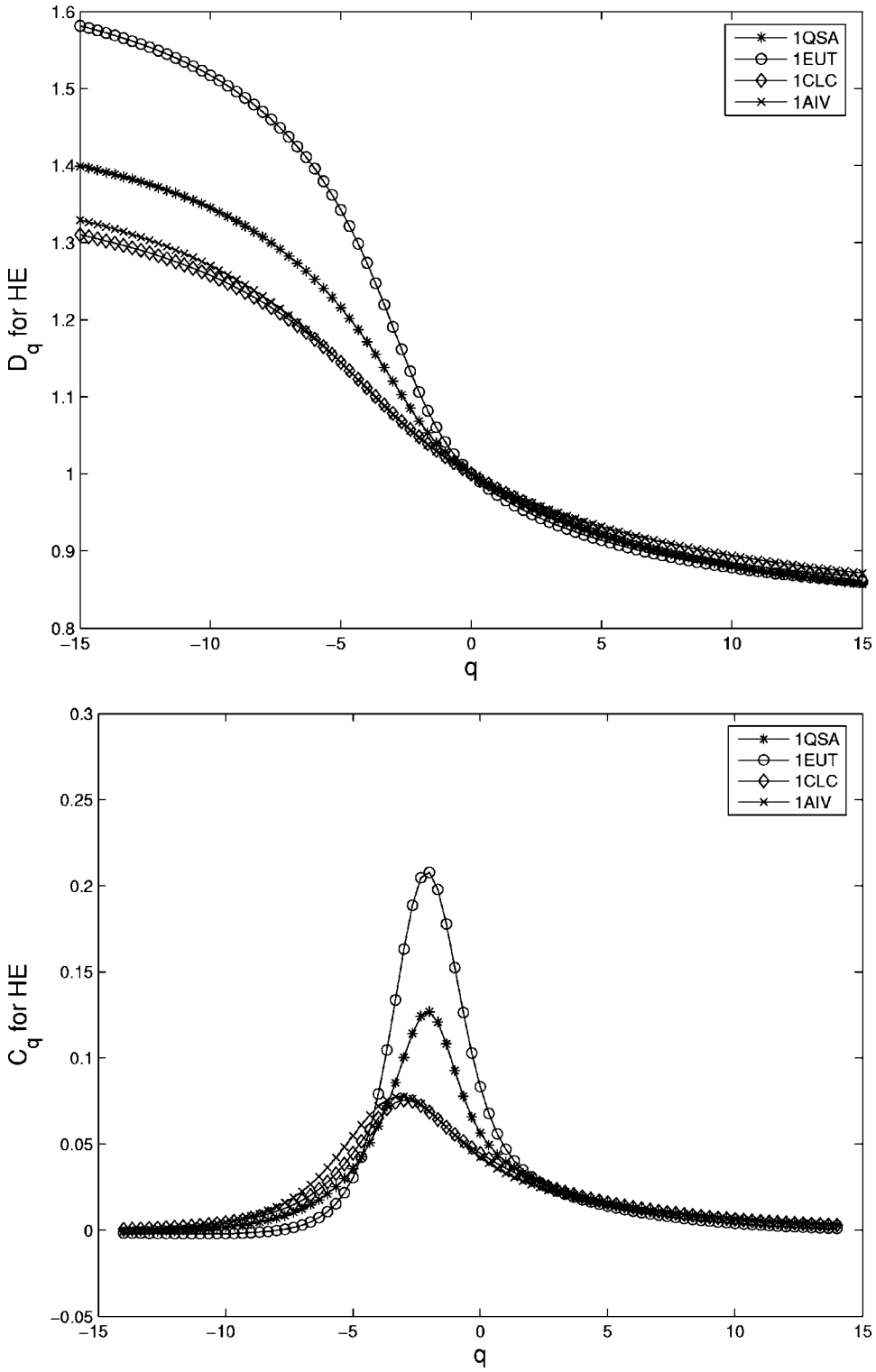

FIG. 4. The D_q curves for the hydrophobic free energy sequences of the four proteins (top) and their related C_q curves (bottom).

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in H_j} \mathbf{x}_i, \quad j = 1,2, \tag{22}$$

$$\mathbf{S}_j = \sum_{\mathbf{x}_i \in H_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T, \quad j = 1,2, \tag{23}$$

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \tag{24}$$

then the parameter vector $\mathbf{w}$ is estimated as $\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ [55]. As a result, Fisher's discriminant rule becomes *assign* $\mathbf{x}$ *to* H_1 *if* $(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{C}_w^{-1}\left[\mathbf{x} - \frac{1}{2}(\mathbf{m}_1 + \mathbf{m}_2)\right] > 0$ *and to* H_2 *otherwise*. [54].

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

$$p_{H1} = n_{cH1}/n_1, \tag{25}$$

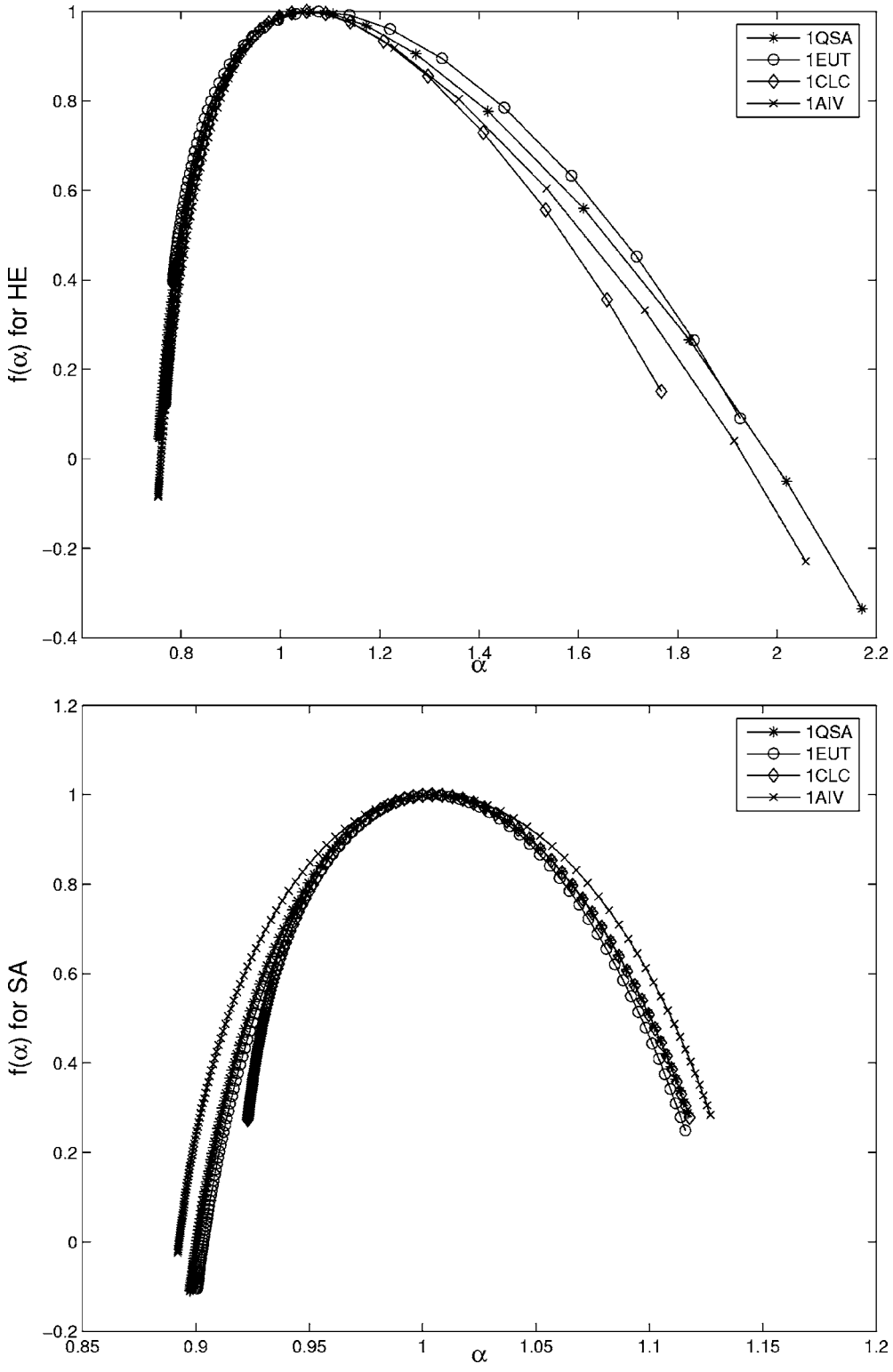

FIG. 5. The multifractal spectra $f(\alpha)$ for the hydrophobic free energy sequences (top) and solvent accessibility sequences (bottom) of the four proteins.

$$p_{H2} = n_{cH2}/n_2, \tag{26}$$

where n_{cH1} and n_{cH2} denote the number of correctly discriminated H_1 elements and the number of correctly discriminated H_2 elements in the training set, respectively.

We denote all β proteins as H_2, the left $\{\alpha, \alpha+\beta, \alpha/\beta\}$ proteins as H_1 in the 2D space (q_0 for HE, P_1) and the 3D space (q_0 for HE, P_1, $C_{max\ q}$ on SA); all $\alpha+\beta$ proteins as H_2, the left $\{\alpha, \alpha/\beta\}$ proteins as H_1 in the 3D space ($\Delta\alpha$ for SA, P_1, λ_z for the Z curve); all α proteins as H_1, all α/β proteins as H_2 in the 3D space (α_{max} for SA, α_{min} for SA, P_1). The estimated parameters $\mathbf{w}=(w_1, w_2, w_3)$ in Fisher's discriminant algorithm and the discriminant accuracies for proteins in parameter spaces shown in Figs. 6–9 are given in Table II. From the discriminant accuracies, it is seen that our clustering is satisfactory and the step to separate α proteins from

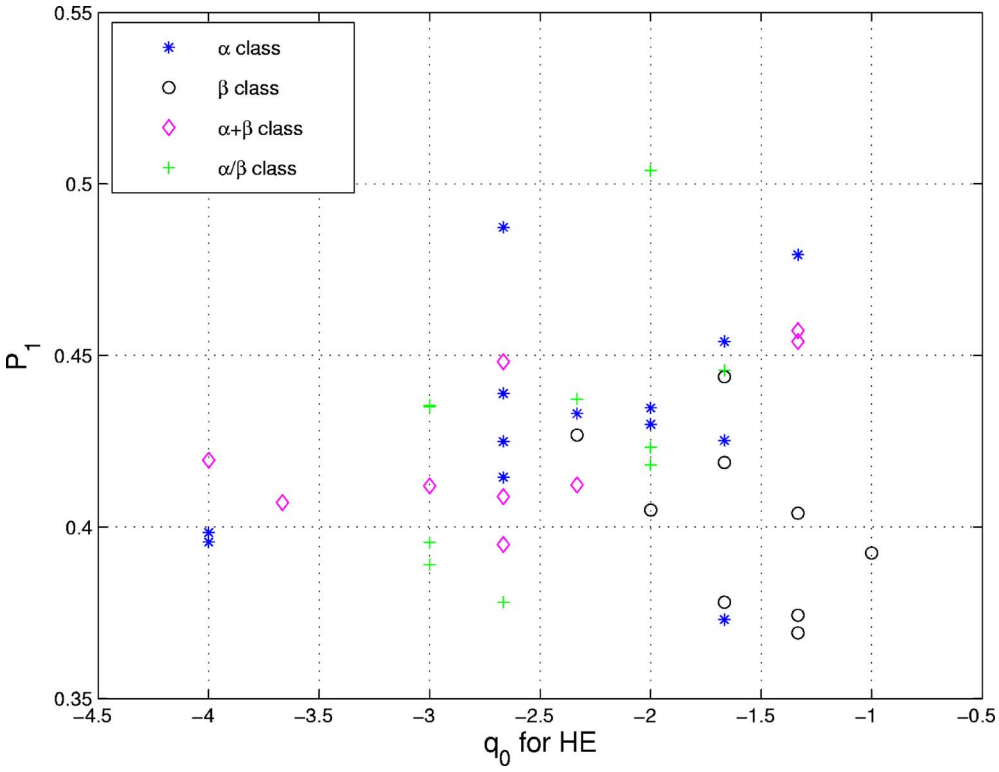

FIG. 6. (Color online) The two-dimensional space (q_0 for HE, P_1) for proteins. In this space, the β class proteins gather as a group and can be separated from the proteins from the other classes.

α/β proteins is the most difficult step. The discriminant accuracies in 3D space (q_0 for HE, P_1, $C_{max\ q}$ on SA) are higher than those in 2D space (q_0 for HE, P_1).Hence the dimension added by $C_{max\ q}$ on SA does add more information to separate β proteins from $\{\alpha, \alpha+\beta, \alpha/\beta\}$ proteins.

The scaling of the solvent accessibility has been used to study structural classification of proteins by Balafas and Dewey [27]. We tried this method for the selected 43 proteins, but the method does not work well for this data set.

VI. CONCLUSIONS

The measure representation, Z curve representation, hydrophobic free energy sequence, and solvent accessibility se-

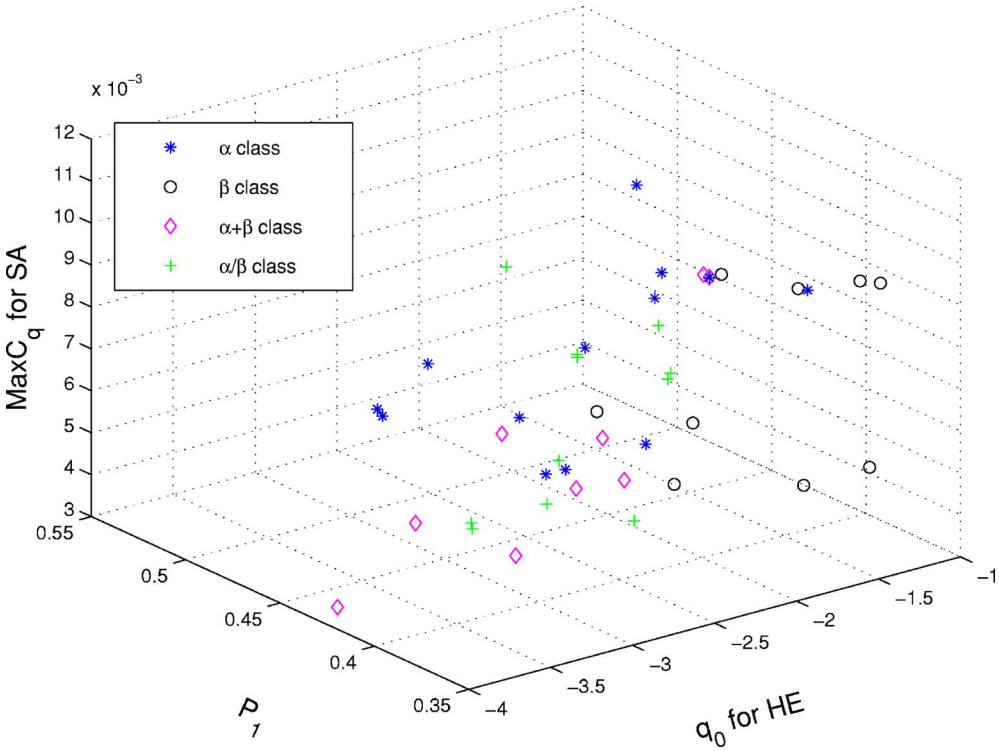

FIG. 7. (Color online) The space (q_0 for HE, P_1, $C_{max\ q}$ for SA). In this space the β class proteins gather as a group and can be separated from the proteins from the other classes.

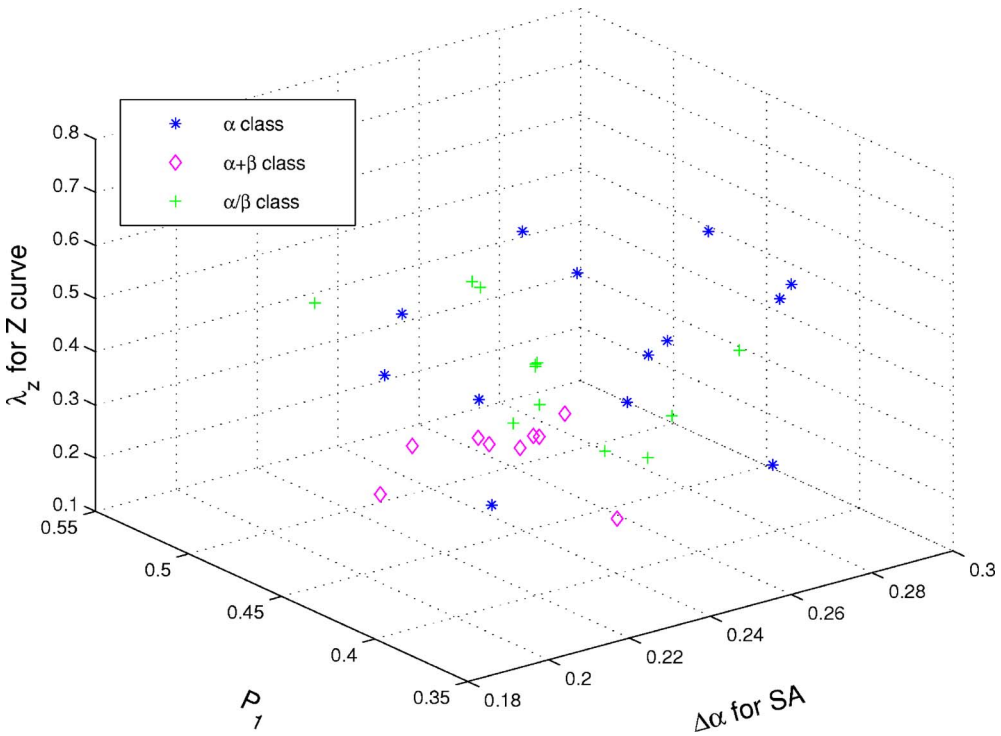

FIG. 8. (Color online) The space ($\Delta\alpha$ for SA, P_1, λ_z for Z curve). In this space, the proteins from the $\alpha+\beta$ class form a group which can be separated from the proteins from the α and α/β classes.

quence of proteins provide useful information and visualization of their secondary and three-dimensional structures.

If a protein sequence is completely random, then the measure representation yields a uniform measure. From the measure representation, the values of the exponent λ of the Z curve representation, and the values of D_q, C_q, and $f(\alpha)$ on the hydrophobic free energy sequence and solvent accessibility sequence, 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 is widely recognized that a protein sequence is not a completely random sequence (for example, see Pande *et al.* [57]).

From the D_q curves of all hydrophobic free energy sequences and solvent accessibility sequences for proteins selected, 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 transi-

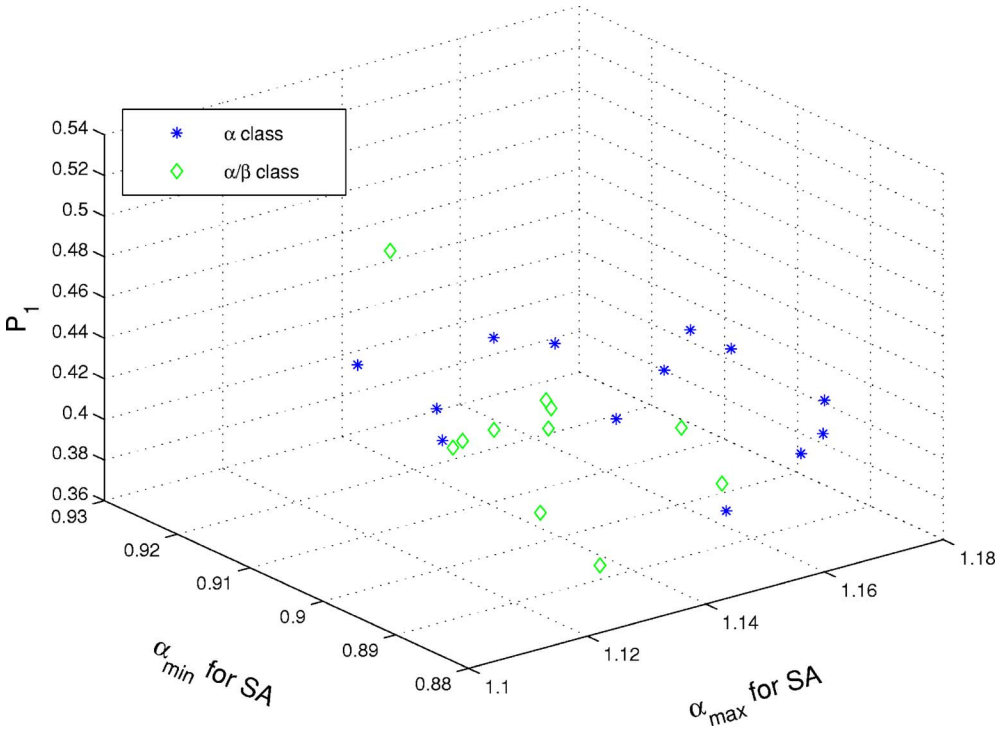

FIG. 9. (Color online) The space (α_{max} for SA, α_{min} for SA, P_1). In this space, the proteins from the α and α/β classes can be separated.

TABLE II. The parameters in Fisher's discriminant and the discriminant accuracies for the selected proteins.

Proteins	w_1	w_2	w_3	p_{H1}	p_{H2}
In Fig. 6	−0.0660	1.3387		91.18%	77.78%
In Fig. 7	−0.0757	1.4263	8.5523	94.12%	88.89%
In Fig. 8	1.5974	0.3458	0.3548	72.00%	100.00%
In Fig. 9	3.2282	2.0229	0.9026	61.54%	75.00%

tion at a critical point, while the $f(\alpha)$ curves indicate that the hydrophobic free energy and solvent accessibility display multifractal scaling.

Some parameter spaces can be constructed using the parameters from the IFS, detrended fluctuation, and multifractal analyses to distinguish and cluster proteins. Each protein can be represented by a point in these spaces. Numerical results indicate that β proteins can be separated from $\{\alpha, \alpha+\beta, \alpha/\beta\}$ proteins in the 2D space (q_0 for HE, P_1) and the 3D space (q_0 for HE, P_1, $C_{max\ q}$ on SA). Then $\alpha+\beta$ proteins can be separated from $\{\alpha, \alpha/\beta\}$ proteins in the 3D space ($\Delta\alpha$ for SA, P_1, λ_z for the Z curve), and finally α proteins from α/β proteins in the 3D space (α_{max} for SA, α_{min} for SA, P_1). Fisher's linear discriminant algorithm is used to give a quantitative assessment of our clustering of the selected proteins. The discriminant accuracies are satisfactory. In particular, they reach 94.12% and 88.89% to separate β proteins from $\{\alpha, \alpha+\beta, \alpha/\beta\}$ proteins in the 3D space (q_0 for HE, P_1, $C_{max\ q}$ on SA).

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 use to perform the secondary structural classification of a protein.

The global mapping of protein structures into some spaces was recently reported by Hou *et al.* [1]. 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.

ACKNOWLEDGMENTS

This research was partially supported by Natural Science Foundation of China (NSFC, Grant No. 30570426), the Youth Foundation of Educational Department of Hunan province in China (Grant No. 05B007), and the Australian Research Council Grant No. DP0559807.

[1] J. Hou, S.-R. Jun, C. Zhang, and S.-H. Kim, Proc. Natl. Acad. Sci. U.S.A. **102**, 3651 (2005).
[2] C. Chothia, Nature (London) **357**, 543 (1992).
[3] G. E. Crooks, J. Wolfe, and S. E. Brenner, Proteins **57**, 804 (2004).
[4] C. Anfinsen, Science **181**, 223 (1973).
[5] C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh, and H. C. Lee, Phys. Rev. Lett. **84**, 386 (2000).
[6] J. Guo, H. Chen, Z. Sun, and Y. Lin, Proteins **54**, 738 (2004).
[7] D. Frishman and P. Argos, Proteins **23**, 566 (1995).
[8] G. E. Crooks and S. E. Brenner, Bioinformatics **20**, 1603 (2004).
[9] D. T. Jones, J. Mol. Biol. **292**, 195 (1999).
[10] K. Karplus, C. Barrett, and R. Hughey, Bioinformatics **14**, 846 (1998).
[11] J. A. Cuff and G. J. Barton, Proteins **40**, 502 (1999).
[12] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, Proteins **47**, 228 (2002).
[13] R. Adamczak, A. Porollo, and J. Meller, Proteins **59**, 467 (2005).
[14] R. Service, Science **307**, 1555 (2005).
[15] K. A. Dill, Biochemistry **24**, 1501 (1985).
[16] H. Li, R. Helling, C. Tang, and N. S. Wingreen, Science **273**, 666 (1996).
[17] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **80**, 5683 (1998).
[18] B. Wang and Z. G. Yu, J. Chem. Phys. **112**, 6084 (2000).
[19] J. Wang and W. Wang, Phys. Rev. E **61**, 6981 (2000).
[20] T. A. Brown, *Genetics*, 3rd ed. (Chapman & Hall, London, 1998).
[21] Z. G. Yu, V. V. Anh, and K. S. Lau, Physica A **337**, 171 (2004).
[22] A. J. Mandell, K. A. Selz, and M. F. Shlesinger, Proc. Natl. Acad. Sci. U.S.A. **94**, 13576 (1997).
[23] A. J. Mandell, K. A. Selz, and M. F. Shlesinger, Physica A **244**, 254 (1997).
[24] K. A. Selz, A. J. Mandell, and M. F. Shlesinger, Biophys. J. **75**, 2332 (1998).
[25] H. Hirakawa, S. Muta, and S. Kuhara, Bioinformatics **15**, 141 (1999).
[26] J. R. MacDonald and W. C. Johnson, Protein Sci. **10**, 1172 (2001).
[27] J. S. Balafas and T. G. Dewey, Phys. Rev. E **52**, 880 (1995).
[28] D. Bordo and P. Argos, J. Mol. Biol. **217**, 721 (1991).
[29] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Academic Press, New York, 1983).
[30] J. Feder, *Fractals* (Plenum, New York, 1988).
[31] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[32] Z. G. Yu, V. Anh, and K. S. Lau, Phys. Rev. E **64**, 031903 (2001).
[33] Z. G. Yu, V. V. Anh, and B. Wang, Phys. Rev. E **63**, 011903 (2001).
[34] C. K. Peng, S. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[35] T. G. Dewey, Proc. Natl. Acad. Sci. U.S.A. **91**, 12101 (1994).
[36] A. Fiser, G. E. Tusnady, and I. Simon, J. Mol. Graphics **12**, 302 (1994).
[37] Z. G. Yu, V. Anh, and K. S. Lau, J. Theor. Biol. **226**, 341 (2004).
[38] Z. G. Yu, V. Anh, and K. S. Lau, Phys. Rev. E **68**, 021913

(2003).
[39] M. B. Enright and D. M. Leitner, Phys. Rev. E **71**, 011912 (2005).
[40] M. A. Moret, J. G. V. Miranda, E. Nogueira, M. C. Santana, and G. F. Zebende, Phys. Rev. E **71**, 012901 (2005).
[41] Y. S. Pavan and C. K. Mitra, Indian J. Biochem. Biophys. **42**, 141 (2005).
[42] V. V. Anh, K. S. Lau, and Z. G. Yu, Phys. Rev. E **66**, 031910 (2002).
[43] M. F. Barnsley and S. Demko, Proc. R. Soc. London, Ser. A **399**, 243 (1985).
[44] E. R. Vrscay, in *Fractal Geometry and Analysis*, edited by J. Belair and S. Dubuc NATO Advanced Studies Institute, Series C: Mathematical and Physical Science (Kluwer, Dordrecht, 1991).
[45] R. Zhang and C. T. Zhang, J. Biomol. Struct. Dyn. **11**, 767 (1994).
[46] C. T. Zhang, Z. S. Lin, M. Yan, and R. Zhang, J. Theor. Biol. **192**, 467 (1998).
[47] M. Yan, Z. S. Lin, and C. T. Zhang, Bioinformatics **14**, 685 (1998).
[48] A. L. Goldberger, C. K. Peng, J. Hausdorff, J. Mietus, S. Havlin, and H. E. Stanley, in *Fractal Geometry in Biological Systems*, edited by P. M. Iannaccone and M. Khokha (CRC Press, Boca Raton, FL, 1996), pp. 249–266.
[49] Z. G. Yu, V. Anh, and K. S. Lau, Physica A **301**, 351 (2001).
[50] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).
[51] J. Lee and H. E. Stanley, Phys. Rev. Lett. **61**, 2945 (1988).
[52] E. Canessa, J. Phys. A **33**, 3637 (2000).
[53] R. B. Russell, in *Protein Structure Prediction: Methods and Protocols*, edited by D. Webster (Humana Press, Totowa, NJ, 2000).
[54] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis* (Academic Press, London, 1979).
[55] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (John Wiley & Sons, New York, 2001).
[56] P. H. A Sneath and R. R. Sokal, *Numerical Taxonomy* (Freeman, San Francisco, 1973).
[57] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12972 (1994).
[58] http://www.rcsb.org/pdb/index.html