Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 337 (2004) 171–184
PHYSICA A
www.elsevier.com/locate/physa

Fractal analysis of measure representation of large proteins based on the detailed HP model

Zu-Guo Yu[a,b,*], Vo Anh[a], Ka-Sing Lau[c]

[a] *Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane Q4001, Australia*
[b] *Department of Mathematics, Xiangtan University, Hunan 411105, PR China*
[c] *Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong*

Received 16 September 2003; received in revised form 12 January 2004

Abstract

The notion of measure representation of protein sequences is introduced based on the detailed HP model. Multifractal analysis and detrended fluctuation analysis are then performed on the measure representations of a large number of long protein sequences. It is concluded that these protein sequences are not completely random sequences through the measure representations and the values of the D_q spectra and related C_q curves. The values of the exponent from the detrended fluctuation analysis show that the K-strings with the ordering in the measure representation exhibit strong long-range correlation. For substrings with length $K = 5$, the D_q spectra of all proteins studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. The C_q curves of all proteins resemble a classical phase transition at a critical point. An IFS model is found to simulate the measure representation of protein sequences very well. From the estimated values of parameters in the IFS model, we think the non-polar residues and uncharged polar residues play a more important role than other kinds of residues in the protein folding process.

PACS: 87.10.+e; 47.53.+n

Keywords: Measure representation; Detrended fluctuation analysis; Multifractal analysis; Analogous specific heat; IFS model

* Corresponding author. School of Mathematical Science, Queensland University of Technology, Garden Point Campus, GPO Box 2434, Brisbane, Q4001, Australia. Tel.: +61-7-38645194; fax: +61-7-38642310.
E-mail addresses: yuzg@hotmail.com, z.yu@qut.edu.au (Z.-G. Yu).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2004.01.057

1. Introduction

The three-dimensional structure of proteins is a complex physical and mathematical problem of prime importance in molecular biology, medicine and pharmacology [1,2]. A protein is composed of one or more chains that are covalently joined. The chain of amino acids are called *polypeptides*. Twenty different kinds of amino acids are found in proteins. It is believed that the dynamical folding process and stable structure, or native conformation, of a protein are determined by its primary structure, namely its amino acid sequence [3,4]. The 20 different amino acids in natural polypeptides can be in any number and any order. Because the number of amino acids in a polypeptide molecule usually ranges from 100 to 1000, the number of different protein molecules that is possible is enormous. Once an amino acid sequence is known, the number of possible space structures it can fold to is also enormous. How to predict the high-level structures (secondary and space structures) from the amino acid sequence is a challenging problem in science, in particular to the large proteins. A number of coarse-grained models have been proposed to provide insight to these very complicated issues [4]. A well-known model in this class is the HP model proposed by Dill et al. [5]. In this model 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). In last decade the HP model has been extensively studied by several groups (e.g. [2,6,7]). After studying the model on lattices, Li et al. [6] found 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 protein-like secondary structures [2,6,8]. But the HP model may be too simple and lacks enough information on the heterogeneity and the complexity of the natural set of residues Ref. [9]. According to Brown [10], in the HP model, one can divide the polar class into three classes: positive polar, uncharged polar and negative polar. So 20 different kinds of amino acids can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. In this model, one considers more details than in the HP model. We call this model a *detailed HP model*. In this paper, we will adopt the detailed HP model.

Fractal geometry provides a mathematical formalism for describing complex spatial and dynamical structures [11,12]. The fractal method has been successfully used to study many problems in Physics, Mathematics, Engineering and Biology in the past 2 decades or so. Multifractal analysis is a useful way to characterise the spatial inhomogeneity of both theoretical and experimental fractal patterns [13]. Multifractal analysis was initially proposed to treat turbulence data. In recent years, it has been applied successfully in many different fields including time series analysis and financial modelling [14]. For the applications of fractal method to DNA sequences, one can refer to Refs. [14–16] and the references therein. The fractal method has been used to study the protein backbone [17], the accessible surface of protein [17–20] and protein potential energy landscapes [21]. The multifractal analysis of solvent accessibilities in proteins was done by Balafas and Dewey [22]. In Ref. [22], the model used to fit the multifractal spectrum is also discussed. But the parameters derived in their multifractal analysis cannot be used to predict the structural classification of a protein from its amino acid sequence.

The amino acid sequence of a protein is also called a *protein sequence* in this paper. Based the idea of DNA walk model and different mapping, a decoded walk model was proposed to study the correlation property of protein sequences by Pande et al. [23] using "Bridge analysis" and Strait and Dewey [24] using multifractal analysis. Deviations of the decoded walk from random behaviour provides evidence of memory.

Inspired by the idea of measure representation of DNA sequence [14], in this paper we propose a visual representation—measure representation of protein sequences based on the detailed HP model. The *Detrended Fluctuation Analysis* (DFA) [15,25] is used to study the correlation property when the measure representation of protein is viewed as a time series. The multifractal analysis of the measure representation of protein will follow. To our knowledge [26], it is much harder to simulate a measure than to fit its multifractal spectrum (because different measures may have the same multifractal spectrum). The iterated function systems (IFS) model proposed by Barnsley and Demko [27] is a powerful tool in fractal theory (many fractals such as the Cantor set can be generated by the IFS model). Here we find the IFS model can be used to simulate the measure representation of protein sequences.

2. Detailed HP model and measure representation of protein sequences

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

For a given protein sequence with length L, $s = s_1 \dots s_L$ where s_i is one of the 20 kinds of amino acids for $i = 1, \dots, L$,, we define

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

So we can obtain a sequence $X(s) = a_1 \dots a_L$, 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 each kind of K-strings in a sequence $X(s)$ from 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 $r = r_1 \dots r_K, r_i \in \{0, 1, 2, 3\}, i = 1, \dots, K$, 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 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 $\sum_{\{r\}} F_K(r) = 1$. Now we can define a measure μ_K on $[0, 1[$ by $\mathrm{d}\mu_K(x) = Y(x)\mathrm{d}x$, where

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

It is easy to see $\int_0^1 \mathrm{d}\mu_K(x) = 1$ 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.

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

3. Detrended fluctuation analysis

The exponent in the detrended fluctuation analysis can be used to characterise the correlation of a time series [15,25]. We can order all the $F(r)$ according to the increasing order of $x_{left}(r)$. We then obtain a sequence of real numbers consisting of 4^K elements which we denote as $F(t), t = 1, \ldots, 4^K$. We can view the sequence $\{F(t)\}_{t=1}^{4^K}$ as a time series. First the time series is integrated as $y(k) = \sum_{t=1}^{k} [F(t) - F_{ave}]$, where F_{ave} 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 least-squares line is fit to the data, representing the trend in that box. The y-coordinate of the straight line segments is denoted by $y_n(k)$. We then detrend the integrated time series, $y(k)$, by subtracting the local trend, $y_n(k)$, in each box. The root-mean-square fluctuation of this integrated and detrended time series is calculated as

$$\mathscr{F}(n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [y(k) - y_n(k)]^2} . \tag{6}$$

Typically, $\mathscr{F}(n)$ will increase with box size n. A linear relationship on a double log graph indicates the presence of scaling

$$\mathscr{F}(n) \propto n^{\lambda} . \tag{7}$$

Under such conditions, the fluctuations can be characterised by the scaling exponent λ, the slope of the line relating $\ln \mathscr{F}(n)$ to $\ln n$. For uncorrelated data, the integrated value $y(k)$ 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 a large value is more likely to be followed by a large value and vice versa. In contrast, $0 < \lambda < 0.5$ indicates a different type of power-law correlations such that large and small values of time series are more likely to alternate.

4. Multifractal analysis

The most common algorithms of multifractal analysis are the so-called *fixed-size box-counting algorithms* [28]. In the one-dimensional case, for a given measure μ with support $E \subset \mathbf{R}$, we consider the *partition sum*

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q \,, \tag{8}$$

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

$$B = [k\varepsilon, (k+1)\varepsilon[\,. \tag{9}$$

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

$$\tau(q) = \lim_{\varepsilon\to 0} \frac{\ln Z_\varepsilon(q)}{\ln \varepsilon} \tag{10}$$

and the generalized fractal dimensions of the measure are defined as

$$D_q = \tau(q)/(q-1) \quad \text{for } q \neq 1 \tag{11}$$

and

$$D_q = \lim_{\varepsilon\to 0} \frac{Z_{1,\varepsilon}}{\ln \varepsilon} \quad \text{for } q = 1 \,, \tag{12}$$

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

$$\frac{1}{q-1} \ln Z_\varepsilon(q)$$

against $\ln \varepsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\varepsilon}$ against $\ln \varepsilon$ for $q = 1$. D_1 is called *information dimension* and D_2 is called *correlation dimension*. The D_q of the positive values of q give relevance to the regions where the measure is large, i.e., to the K-strings with high probability. The D_q of the negative values of q deal with the structure and the properties of the most rarefied regions of the measure.

Some sets of physical interest have a non-analytic dependence of D_q on q. Moreover, this phenomenon has a direct analogy to the phenomenon of phase transitions in condensed-matter physics [29]. The existence and type of phase transitions might turn out to be a worthwhile characterisation of universality classes for the structures [30]. The concept of phase transition in multifractal spectra was introduced in the study of

logistic maps, Julia sets and other simple systems. Evidence of phase transition was found in the multifractal spectrum of diffusion-limited aggregation [31]. By following the thermodynamic formulation of multifractal measures, Canessa [32] derived an expression for the 'analogous' specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1)\,. \tag{13}$$

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 our measure representations of protein sequences.

5. IFS model and moment method

5.1. IFS model

In order to simulate the measure representation of the complete genome, Anh et al. [33] proposed the *iterated function systems* (IFS) model and the recurrent IFS model. IFS is the name given by Barnsley and Demko [27] 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, $E_{\sigma_1\sigma_2\ldots\sigma_n} = w_{\sigma_1} \circ w_{\sigma_2} \circ \cdots \circ w_{\sigma_n}(E_0)$ and

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

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

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

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 [27]. 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 *chaos game*.

Let μ be the invariant measure on the attractor of the IFS, χ_B the characteristic function for the Borel subset $B \subset E$, then from the ergodic theorem for IFS Ref. [27],

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

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.

5.2. Moment method to estimate the parameters in 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 [34] introduced a moment method to perform this task. If μ is the invariant measure and E the attractor of IFS in $\mathbf{R}$, the moments of μ are

$$g_i = \int_E x^i \mathrm{d}\mu, \quad g_0 = \int_E \mathrm{d}\mu = 1 \ . \tag{14}$$

If $w_i(x) = c_i x + d_i,\ i = 1, \ldots, N$, then the following well-known recursion relations hold [34]:

$$\left[1 - \sum_{i=1}^{N} p_i c_i^n\right] g_n = \sum_{j=1}^{n} \binom{n}{j} g_{n-j} \left(\sum_{i=1}^{N} p_i c_i^{n-j} d_i^j\right) \ . \tag{15}$$

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 (14), and g_k the formal expression of moments obtained from (15), then through solving the optimisation problem

$$\min_{c_i, d_i, p_i} \sum_{k=1}^{n} (g_k - G_k)^2 \quad \text{for some chosen } n \ , \tag{16}$$

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

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

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 optimisation problem (16). Based on the estimated values of the probabilities, we can use the chaos game to generate a histogram approximation of the invariant measure of IFS which we can compare with the real measure representation of the protein sequence. In order to clarify how close the simulation measure is to the original measure representation, we convert the measure to its walk representation. If $t_j,\ j = 1, 2, \ldots, 4^K$, is the histogram of a measure and t_{ave} is its average, then we define $T_j = \sum_{k=1}^{j} (t_k - t_{ave}),\ j = 1, 2, \ldots, 4^K$. So we can plot the two walks of the real measure representation and the measure generated by chaos game of IFS model on the same figure.

6. Data and numerical results

The methods introduced in the previous sections can only be used for long protein sequences (corresponding to large proteins). The amino acid sequences of 32 large proteins are selected from RCSB Protein Data Bank (PDB) (http://www.rcsb.org/pdb/index.html). These 32 proteins belong to five structure classes [35] according to their

Table 1
The properties and the exponent λ in detrended fluctuation analysis of all 32 proteins selected

Class	PDB ID	Protein	Length	Exponent λ
	1AVC	Annexin VI	673	0.7668386
	1B89	Clethrin heavy chain	449	0.8144683
	1BJ5	Human serum albumin	585	0.7427572
α	1HO8	Vacuolar ATP synthase subunit H	480	0.7721964
	1IAL	Importin alpha	453	0.7197123
	1QSA	Soluble lytic transglycosylase SH70	618	0.7482758
	2BCT	β-catenin	516	0.7193562
	5EAS	5-Epi-Aristolochene synthase	548	0.8192653
	1B9S	Neuramindase	390	0.7716008
	1DAB	P.69 pertactin	539	0.7557988
β	1EUT	Sialidase	605	0.7632415
	1FNF	Fibronectin	368	0.7087377
	1JX5	Integrin α-Iib	452	0.7224556
	1MAL	Maltoporin	421	0.7831249
	1B90	β-Amylase	516	0.7781916
	1BBU	Lysyl-tRNA synthetase	504	0.7936335
$\alpha+\beta$	1BYT	Lioxygenase-3	857	0.7693996
	1CLC	Endoglaeanase celd	639	0.7655830
	1E7U	Phosphatidylinositol 3-kinase Catalytic subunit	961	0.7795467
	1A8I	Glycogen phosphorylase B	841	0.8215442
α/β	1ACJ	Acetylcholinesterase complexed with tacrine	537	0.7390218
	1AOV	Apo-ovotransferin	686	0.7503767
	1BFD	Benzoylformate decarboxylase	528	0.7581296
	1CRL	Lipase (triacylglycerol hydrolase)	534	0.7358587
	1DPI	DNA Ploymerase I dCMP complex-chian	605	0.7550018
	1EFG	Elongation factor G complexed with guanosine 5'-diphosphate chain A	691	0.7986233
	1EPS	5-enol-pyruvyl-3-phosphate synthase chain	427	0.7521843
Others	1F1O	Adenylosuccinate lyase	431	0.7942725
	1KVP	Capsid protein chimera	497	0.7643545
	1PMD	Peptidoglycan synthesis	675	0.7451864
	1TPT	Thymidine phosphorylase chain	440	0.7486859
	4ACE	Acetylcholinesterase	537	0.7388495

secondary structures: α, β, $\alpha+\beta$ (α,β alternate), α/β (α,β segregate) and others (no α and no β) proteins. The properties of these proteins are given in Table 1. First we convert the amino acid sequences of these proteins to their measure representations with $K=5$ according to the method introduced in Section 2. If K is too small, there are not enough combinations of letters from set $\{0,1,2,3\}$, therefore there is no statistical sense. And if K is too big, the frequencies of most substrings are zero. So we cannot obtain any biological information from the measure representation. Considering the

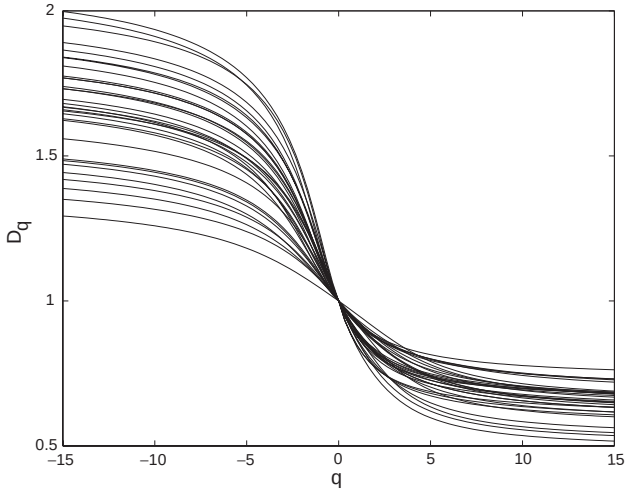

Fig. 1. The multifractal spectra D_q of measure representations of 32 proteins selected.

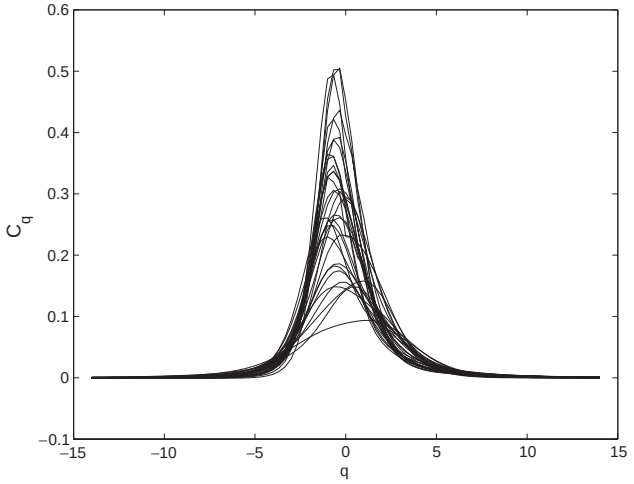

Fig. 2. C_q curves of measure representations of 32 proteins selected.

length of the selected proteins which ranges from 350 to 1000, we think it is suitable to choose $K = 5$. Then the detrended fluctuation analysis of these proteins was performed. The values of the exponent λ in the detrended fluctuation analysis are also given in Table 1.

The multifractal spectra D_q and the related spectra C_q of the measure representations of all 32 proteins are calculated and showed in Figs. 1 and 2, respectively.

Table 2
The estimated parameters in the IFS model of all 32 proteins selected

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

Finally, we simulated the measure representations of all 32 proteins using the IFS model and moment method introduced in Section 5. The estimated parameters in the IFS model are given in Table 2. For examples, we show the histograms of measure representation and simulated measures of protein *Human Serum Albumin* (PDB ID: 1BJ5) in Fig. 3 and their walk representations in Fig. 4; those measures of protein *P.69 Pertactin* (PDB ID: 1DAB) in Fig. 5 and their walk representations in Fig. 6.

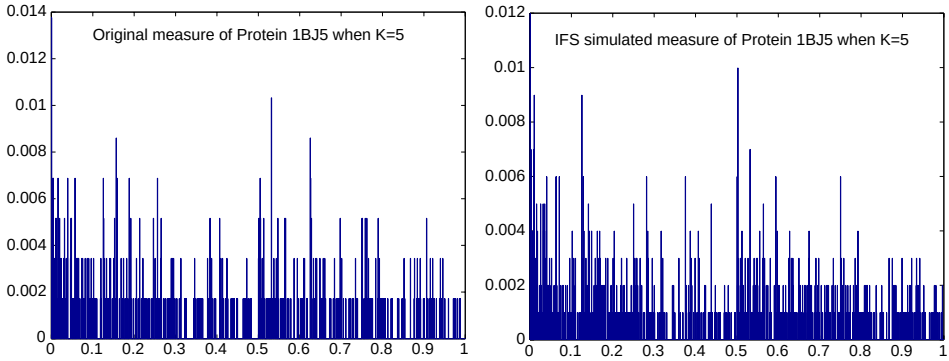

Fig. 3. The measure representation (left) and the IFS simulation (right) of protein *human serum albumin* (PDB ID: 1BJ5).

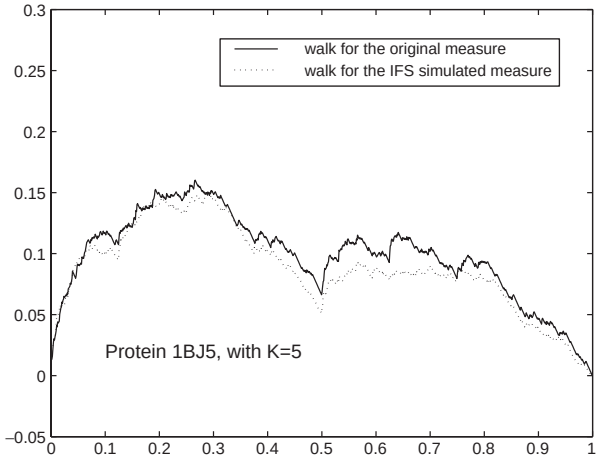

Fig. 4. The walk representations of measures in Fig. 3.

7. Discussion and conclusions

The idea of our measure representation of protein is similar to the measure representation of complete genome [14]. It provides a powerful visualisation method for protein sequences in more details than the HP model. If a protein sequence is completely random, then our measure representation yields a uniform measure ($D_q = 1,\ C_q = 0$).

From the measure representation and the values of D_q and C_q, it is seen that there exists a clear difference between the protein sequences considered here and completely random sequence. Hence we can conclude that these protein sequences are not random sequences. This result coincides with the result of Pande et al. [23].

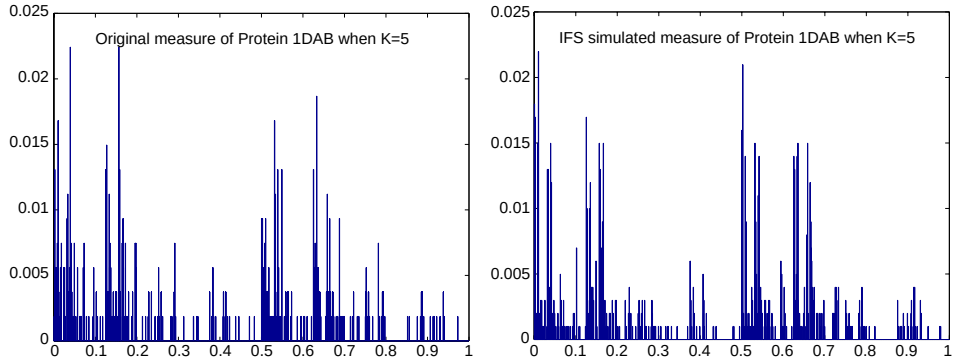

Fig. 5. The measure representation (left) and the IFS simulation (right) of protein *P.69 Pertactin* (PDB ID: 1DAB).

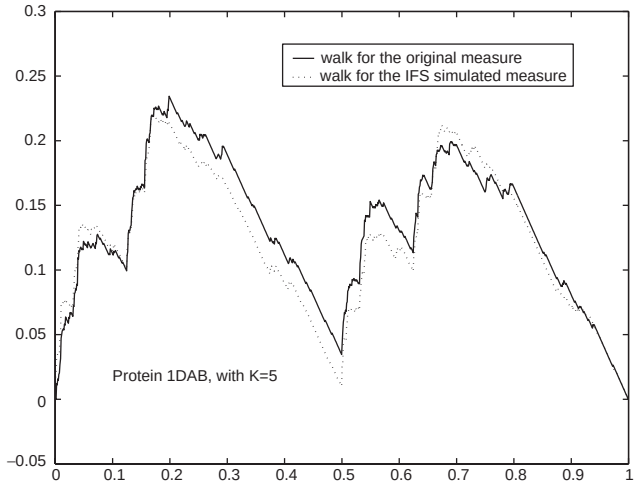

Fig. 6. The walk representations of measures in Fig. 5.

From Fig. 1, it is seen that the D_q spectra of all protein sequences are multifractal-like and sufficiently smooth so that the C_q curves can be meaningfully estimated. From Fig. 2, one can see that the C_q curves of all protein sequences resemble a classical phase transition at a critical point.

Through the detrended fluctuation analysis and from Table 1, the values of exponent λ range from 0.70 to 0.83. These values are far from 0.5. Hence when we view our measure representations of protein sequences as time series, they are far from being random time series, and in fact exhibit strong correlation. Here the long-range correlation is for the K-strings with ordering in the measure representation, and it is different from the residue correlations introduced by other people.

Figs. 4 and 6 indicate that the difference between the walk representations of measure representation and IFS simulated measure is very small. We find that IFS is a good model to simulate the measure representation of protein sequences. From above, once the probabilities are determined, the IFS model is obtained. Hence the probabilities obtained from the IFS model can be used to characterise the measure representation of the protein sequences. From Table 2, we find the probability p_3 (which corresponds to the uncharged polar property) can be used to distinguish the structural class of proteins from α class and β class (values of p_3 of proteins in class α are less than those of proteins in class β), and the probability p_1 (which corresponds to the non-polar property) can be used to distinguish the structural class of proteins from class $\alpha+\beta$ and class α/β (values of p_1 of proteins in class α/β are less than those of proteins in class $\alpha+\beta$). Hence we believe that the non-polar residues and uncharged residues play a more important role than other kind of residues in the protein folding process. This information is useful for protein structure prediction.

We also tried replacing the detailed HP model by the classification of residues used in Ref. [9] in our frame. But it cannot improve the results obtained from the detailed HP model.

The detailed HP model can also be used in the chaos game representation of linked protein sequences from the complete genome [36].

Acknowledgements

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

References

[1] C. Chothia, Nature (London) 357 (1992) 543–544.
[2] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, H.C. Lee, Phys. Rev. Lett. 84 (2) (2000) 386–389.
[3] C. Anfinsen, Science 181 (1973) 223.
[4] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, H.C. Lee, Phys. Rev. E 65 (2002) 041923.
[5] K.A. Dill, Biochemistry 24 (1985) 1501;
H.S. Chan, K.A. Dill, Macromolecules 22 (1989) 4559.
[6] H. Li, R. Helling, C. Tang, N.S. Wingreen, Science 273 (1996) 666.
[7] B. Wang, Z.G. Yu, J. Chem. Phys. 112 (2000) 6084–6088.
[8] C. Micheletti, J.R. Banavar, A. Maritan, F. Seno, Phys. Rev. Lett. 80 (1998) 5683.
[9] J. Wang, W. Wang, Phys. Rev. E 61 (2000) 6981–6986.
[10] T.A. Brown, Genetics, 3rd Edition, Chapman & Hill, London, 1998.
[11] B.B. Mandelbrot, The Fractal Geometry of Nature, Academic Press, New York, 1983.
[12] J. Feder, Fractals, Plenum Press, New York, 1988.
[13] P. Grassberger, I. Procaccia, Phys. Rev. Lett. 50 (1983) 346.
[14] Z.G. Yu, V.V. Anh, K.S. Lau, Phys. Rev. E 64 (2001) 031903.
[15] Z.G. Yu, V.V. Anh, B. Wang, Phys. Rev. E 63 (2001) 011903.
[16] C.K. Peng, S. Buldyrev, A.L.Goldberg, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168.

[17] T.G. Dewey, J. Chem. Phys. 98 (1993) 2250.
[18] P. Pfiefer, U. Welz, H. Wipperman, Chem. Phys. Lett. 113 (1985) 535.
[19] B.A. Fedorov, B.B. Fedorov, P.W. Schmidt, J. Chem. Phys. 99 (1993) 4076.
[20] M. Lewis, D.C. Rees, Science 230 (1985) 1163.
[21] D.A. Lidar, D. Thirumalai, R. Elber, R.B. Gerber, Phys. Rev. E 59 (1999) 2231.
[22] J.S. Balafas, T.G. Dewey, Phys. Rev. E 52 (1995) 880.
[23] V.S. Pande, A.Y. Grosberg, T. Tanaka, Proc. Natl. Acad. Sci. USA 91 (1994) 12972.
[24] B.J. Strait, T.G. Dewey, Phys. Rev. E 52 (1995) 6588.
[25] A.L. Goldberger, C.K. Peng, J. Hausdorff, J. Mietus, S. Havlin, H.E. Stanley, Fractals and the Heart, in: P.M. Iannaccone, M. Khokha (Eds.), Fractal Geometry in Biological Systems, CRC Press, Inc., Boca Raton, 1996, pp. 249–266.
[26] V. Anh, K.S. Lau, Z.G. Yu, J. Phys. A: Math. Gene. 34 (2001) 7127–7139.
[27] M.F. Barnsley, S. Demko, Proc. R. Soc. London A 399 (1985) 243.
[28] T. Halsey, M. Jensen, L. Kadanoff, I. Procaccia, B. Schraiman, Phys. Rev. A 33 (1986) 1141.
[29] D. Katzen, I. Procaccia, Phys. Rev. Lett. 58 (1987) 1169.
[30] T. Bohr, M. Jensen, Phys. Rev. A 36 (1987) 4904.
[31] J. Lee, H.E. Stanley, Phys. Rev. Lett. 61 (1988) 2945.
[32] E. Canessa, J. Phys. A: Math. Gen. 33 (2000) 3637.
[33] V.V. Anh, K.S. Lau, Z.G. Yu, Recognition of an organism from fragments of its complete genome, Phys. Rev. E 66 (2002) 031910.
[34] E.R. Vrscay, in: J. Belair (Ed.), Fractal Geometry and Analysis, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.
[35] R.B. Russell, in: D. Webster (Ed.), Protein Structure Prediction: Methods and Protocols, Humana Press Inc., Totowa, NJ, 2000.
[36] Z.G. Yu, V.V. Anh, K.S. Lau, Chaos game representation, and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model, J. Theor. Biol. 226 (3) (2004) 341–348.