Protein structural classification and family identification by multifractal analysis and wavelet spectrum*

Zhu Shao-Ming(朱少茗)[a)b)], Yu Zu-Guo(喻祖国)[a)b)†], and Ahn Vo[a)]

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

(Received 9 April 2010; revised manuscript received 10 September 2010)

Family identification is helpful for predicting protein functions. It has been known from the literature that longer sequences of base pairs or amino acids are required to study patterns in biological sequences. Since most protein sequences are relatively short, we randomly concatenate or link the protein sequences from the same family or superfamily together to form longer protein sequences. The 6-letter model, 12-letter model, 20-letter model, the revised Schneider and Wrede scale hydrophobicity, solvent accessibility and stochastic standard state accessibility are used to convert linked protein sequences into numerical sequences. Then multifractal analyses and wavelet analysis are performed on these numerical sequences. The parameters from these analyses can be used to construct parameter spaces where each linked protein is represented by a point. The four classes of proteins, namely the α, β, $\alpha+\beta$ and α/β classes, are then distinguished in these parameter spaces. The Fisher linear discriminant algorithm is used to assess the discriminant accuracy. Numerical results indicate that the discriminant accuracies are satisfactory in separating these classes. We find that the linked proteins from the same family or superfamily tend to group together and can be separated from other linked proteins. The methods are helpful for identifying the family of an unknown protein.

Keywords: protein family, multifractal analysis, wavelet spectrum
PACS: 05.45.Df, 64.60.al, 87.15.bd **DOI:** 10.1088/1674-1056/20/1/010505

1. Introduction

The two key steps in the analysis of genomic data are the identification of genes in the raw DNA sequence and the prediction of the function of the corresponding open reading frames (ORFs).[1] For the second step, once a gene sequence is identified, the protein sequence can be derived directly from the DNA sequence, but the mapping of protein sequence to structure and function is more complex.[2−4] The structure of a protein is difficult to obtain with experimental methods. On the other hand, computational structure prediction methods provide the information about large sequences whose structures are not determined experimentally. It is well known that an amino acid sequence partially determines the eventual three-dimensional structure of a protein, and the similarity at the protein sequence level implies similarity of function.[5] Although there are some notable exceptions to this general rule such as proteins that are known to have different functions depending on their cellular location,[6] most successful protein function prediction methods are based on homology between the protein sequence of interest and that of a protein with known structure and function.[2] Abascal and Valencia[1] proposed an automatic annotation of protein function based on family identification. Weisser and Klein-Seetharaman[2] presented algorithms to test membrane protein on the G-protein coupled receptor (GPCR) family. They focused on the study of proteins from one family and its subfamilies. These proteins are amenable to experimental studies and can provide useful information for the prediction methods of protein structure and function. Proteins in the same family usually have similar functions. Therefore family identification is an important problem in the study of proteins. This study focuses on the classification of protein families and structural classes.

Classification of protein structural classes is another important issue in bioinformatics.[4,7−13] Some prediction methods used features such as pseudo AA composition,[14] complexity measure factors [15,16] and

*Project supported by the Australian Research Council (Grant No. DP0559807), a Research Capacity Building Award at QUT, Scientific Research Fund of Hunan Provincial Education Department of China (Grant No. 06C826), the Chinese Program for New Century Excellent Talents in University (Grant No. NCET-08-06867), the Hunan Provincial Natural Science Foundation of China (Grant No. 10JJ7001), the Program for Furong Scholars of Hunan Province of China, and the Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province of China.
†Corresponding author. E-mail: yuzg@hotmail.com
© 2011 Chinese Physical Society and IOP Publishing Ltd

some complex feature vectors.[8,17−19] Many classification algorithms were applied to the prediction of protein structural classes such as rough sets,[10] neural network,[20] support vector machine[21,22] and boosting.[23] A review by Chou[24] detailed the prediction of protein structural classes systematically and introduced the development of some existing methods. The mentioned methods can also be applied to the prediction of the subcellular positions[11] and secondary structures of proteins.[18] The proteins used for prediction can be found in the structural classification of proteins (SCOP) database[25,26] and the class, architecture, topology and homologous (CATH) superfamily database.[27] Yu *et al.*[28] used hydrophobic free energy and solvent accessibility of proteins to construct discriminant parameter spaces of large proteins. Yang *et al.*[29] studied 49 large proteins with hydrophobicity scale of amino acids and a 6-letter model. Zhou *et al.*[30] proposed to cluster protein structures using recurrence quantification analysis. Yang *et al.*[31] proposed to predict protein structural classes by recurrence quantification analysis based on chaos game representation (CGR).

Fractal methods are important among the nonlinear methods and have been widely used in many fields such as oil pipeline[32] and surface roughness.[33] Fractal analysis proposed by Mandelbrot[34] provides a mathematical formalism for describing complex spatial and dynamical structures of sequences. Multifractal analysis (MFA) is a method to study a sequence with the box-counting technique to obtain fractal characteristics, such as the generalized fractal dimension D_q, the related spectrum C_q and the multifractal spectrum $f(\alpha)$. It is useful to characterize the spatial heterogeneity of both theoretical and experimental fractal patterns. In recent years the multifractal analysis has been applied to the analysis of genome and protein sequences by our group.[28,29,35−39] Some fractal methods of analysing DNA sequence were reviewed by Yu *et al.*[40] Yang *et al.*[39] applied multifractal analysis to human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. Jeffrey[41] first proposed a chaos game representation (CGR) of DNA sequences, self-similarity or fractal structures were found in these plots. Han and Fu[42] proposed a wavelet-based multifractal analysis of DNA sequences by using chaos-game representation. The idea of CGR of DNA sequences proposed by Jeffrey[41] was generalized and applied for visualizing and analysing protein databases by Fiser *et al.*[43] Basu *et al.*[44] proposed a new method for the CGR of different families of proteins. A CGR-walk model based on CGR coordinates for the DNA sequences[45] and for the protein sequences[46] were proposed recently. Yu *et al.*[47] gave a CGR of randomly-linked functional protein sequences, then proposed to use the recurrent iterated function systems (RIFS) in fractal theory to simulate the measure based on their chaos game representations.

The wavelet analysis is a useful tool in many applications such as noise reduction, image processing, information compression, synthesis of signals and the study of biological data. Wavelets are mathematical functions that decompose data into different frequency components, and then each component is studied with a resolution matched to its scale. The wavelet transform of a function can be viewed as an approximation of the function. The wavelet approach has been used for predicting secondary structures of proteins.[48−50] Chen *et al.*[48] predicted the secondary structure of a protein by the continuous wavelet transform (CWT) and Chou–Fasman method. Marsolo and Ramamohanarao[49] combined the pair-wise distance with wavelet decomposition to generate a set of coefficients and capture some features of proteins. Qiu *et al.*[50] used the continuous wavelet transform to extract the positions of α-helices and short peptides with special scales.

The development in proteomic analyses requires to determine protein functions. The first step is to classify proteins and predict families. Jeffrey[41] reported that at least 2000 bases are required to generate fractal patterns. Because the number of amino acids in the majority of proteins is much less than this threshold value, Basu *et al.*,[44] Deschavanne and Tuffery[51] proposed to concatenate the amino acid sequences of proteins which belong to a particular family into a single sequence and then extract patterns or features from the linked protein sequences to study the problems of protein family identification and structural classification. Basu *et al.*[44] found that different protein families exhibit distinct patterns in their CGRs; therefore, it is possible to use the features of such protein families for identifying new members of the families. Deschavanne and Tuffery[51] discussed how to perform functional class assignment using family signature. Recently, the study focus has changed from a single protein to the link of proteins.[52] This motivates the approach of using linked proteins to study the structure of proteins from the same family.

The proteins in the same family normally share similar structural and functional characteristics. In the present study, we also link protein sequences from the same family or superfamily together to form longer protein sequences, then use multifractal analysis to characterize the linked protein sequences, and apply wavelet analysis to capturing the details. We believe that it is the first time to combine the multifractal analysis and wavelet method in the analysis of the structure classes of linked proteins and their families.

Since some families have many proteins with homologous amino acid sequences, we arbitrarily choose some families and superfamilies, and select those proteins with less homologous sequences. We intend to classify the structural classes and families of proteins listed in Tables 1 to 4. We arbitrarily link the proteins from the same family or the same superfamily to form 120 long protein sequences, and each structural class contains 30 linked protein sequences. Firstly, these proteins are converted into numerical sequences with six different methods, namely the revised Schneider and Wrede scale hydrophobicity (RSWH), the solvent accessibility (SA), the stochastic standard state accessibility (SSSA), the 6-letter model, 12-letter model, and 20-letter model. Then ordinary multifractal analysis (MFA), analogous multifractal analysis (AMFA) and wavelet spectrum (WS) are used to analyse the resulting six kinds of sequences. Nine parameters from these methods are selected to construct parameter spaces where each linked protein is represented by one point. In addition, we use Fisher's linear discriminant algorithm to distinguish the linked protein sequences from the four structural classes in these parameter spaces. The results show that the accuracies are satisfactory. We find that in these parameter spaces, the proteins from the same family tend to group together and can be separated from other proteins in different families. Those proteins having similar amino acid sequences cluster together to reveal information about their families.

2. Models and chemical features of amino acids

2.1. 6-letter model, 12-letter model and 20-letter model

According to Ref. [53], the twenty kinds of amino acids can be divided into six groups: 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 group consists of three residues Met, Val, IlE. The h_β group includes seven residues Cys, Tyr, Phe, Gln, Leu, Thr, Trp. The I_β group contains the residue Ala. The i_β group has three residues Arg, Gly, Asp. The b_β group has five residues Lys, Ser, His, Asn, Pro. The residue Glu belongs to the B_β group. Given a protein sequence $s = s_1, \cdots, s_L$ with length L, Yang *et al.*[29] defined a 6-letter model map as $a_i = 0, 1, 2, 3, 4, 5$ when $s_i \in B_\beta, b_\beta, i_\beta, I_\beta, h_\beta$ and H_β respectively. Hence, the protein sequence can be represented by a numerical sequences $\{a_i\}_{i=1}^{L}$; here $a_i \in \{0, 1, 2, 3, 4, 5\}$ for each $i = 1, 2, \ldots, L$.

We now analogously define a 12-letter model and a 20-letter model. With the idea of chaos game representation (CGR) based on a 12-sided regular polygon in Ref. [44], we define the 12-letter model as $A = 0$, $G = 0$, $P = 1$, $S = 2$, $T = 2$, $H = 3$, $Q = 4$, $N = 5$, $D = 6$, $E = 6$, $R = 7$, $K = 7$, $I = 8$, $L = 8$, $V = 8$, $M = 8$, $W = 9$, $F = 10$, $Y = 10$, and $C = 11$ according to the order of vertices on the 12-sided regular polygon to represent these amino acids.

The 20-letter model can be similarly defined as $A = 0$, $R = 1$, $N = 2$, $D = 3$, $C = 4$, $Q = 5$, $E = 6$, $G = 7$, $H = 8$, $I = 9$, $L = 10$, $K = 11$, $M = 12$, $F = 13$, $P = 14$, $S = 15$, $T = 16$, $W = 17$, $Y = 18$, and $V = 19$ according to the dictionary order of the 3-letter representation of each amino acid listed in Ref. [54].

These three models can be used to convert a protein sequence into three different numerical sequences.

2.2. Revised Schneider and Wrede scale hydrophobicity, solvent accessibility, stochastic standard state accessibility

It is important to select the physico-chemical features of amino acids. These features can influence the results of experiments. We tested some properties, and found that these properties are significant in our numerical experiments. The RSWH, solvent accessibility of the side chain of a protein have been used in other studies. Giuliani *et al.*[55] applied recurrence quantification analysis (RQA) to the hydrophobicity sequence of proteins with Schneider and Wrede scale (SWH). The SWHs of the twenty 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 $V = 2.6$. Yang *et al.*[29] added a constant 12.30 to each of these values to make all the twenty values non-negative, yielding the so-called RSWH.

The solvent accessibility of the amino acids can also be used for the prediction of secondary structures and structural classification.[28,56,57] The solvent accessibilities for solvent exposed areas larger than 30 Å^2 (1 Å=0.1 nm) are:[28] $S = 0.7$, $T = 0.71$, $A = 0.48$, $G = 0.51$, $P = 0.78$, $C = 0.32$, $D = 0.81$, $E = 0.93$, $Q = 0.81$, $N = 0.82$, $L = 0.41$, $I = 0.39$, $V = 0.40$, $M = 0.44$, $F = 0.42$, $Y = 0.67$, $W = 0.49$, $K = 0.93$, $R = 0.84$, $H = 0.66$.

Rose *et al.*[58] proposed different measures of hydrophobicity of proteins. They gave four kinds of values for surface area and hydrophobicity of each amino acid. We use A^0, the stochastic standard state accessibility, i.e. the solvent accessible surface area of a residue in standard state. Physically, the stochastic standard state reflects the degree to which residues are buried by backbone atoms from covalent neighbours in the ensemble. The SSSA of the 20 kinds of amino acids are: $A = 118.1$, $R = 256.0$, $N = 165.5$, $D = 158.7$, $C = 146.1$, $Q = 193.2$, $E = 186.2$, $G = 88.1$, $H = 202.5$, $I = 181.0$, $L = 193.1$, $K = 225.8$, $M = 203.4$, $F = 222.8$, $P = 146.8$, $S = 129.8$, $T = 152.5$, $W = 266.3$, $Y = 236.8$, and $V = 164.5$.

We convert each amino acid according to RSWH, SA and SSA along the protein sequence into three different numerical sequences.

3. Methods

3.1. Ordinary multifractal analysis

The common algorithm of multifractal analysis is the fixed-size box-counting algorithm.[28,29,59] In the one-dimensional case, for a given measure μ with support $E \subset \boldsymbol{R}$, we may consider the partition sum

$$Z_\epsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q, \quad q \in \boldsymbol{R}, \tag{1}$$

where the sum runs over all nonempty boxes B with a given side ϵ in a grid covering the support E, i.e., $B = [k\epsilon, (k+1)\epsilon)$. The mass exponent $\tau(q)$ is defined by

$$\tau(q) = \lim_{\epsilon\to 0} \frac{\ln Z_\epsilon(q)}{\ln \epsilon} \tag{2}$$

and the generalized fractal dimensions of the measure are given by

$$D(q) = \tau(q)/(q-1) \ \text{ for } q \neq 1, \tag{3}$$

and

$$D_q = \lim_{\epsilon\to 0} \frac{Z_{1,\epsilon}}{\ln \epsilon} \ \text{ for } \ q = 1, \tag{4}$$

where $Z_{1,\epsilon} = \sum_{\mu(B)\neq 0} \mu(B) \ln \mu(B)$. The generalized fractal dimensions are calculated by a linear regression of $(\ln Z_\epsilon(q))/(q-1)$ against $\ln \epsilon$ for $q \neq 1$, and a linear regression of $Z_{1,\epsilon}$ against $\ln \epsilon$ for $q = 1$.

The concept of phase transitions in multifractal spectra was proposed in the study of logistic maps, Julia sets and other simple systems. With the ideas of the thermodynamic formulation of multifractal measures, Canessa [60] obtained an expression for the analogous specific heat as

$$C_q = -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \tag{5}$$

Canessa[60] found that a classical phase transition resembles the form of C_q at the critical point in financial time series.

The singularities of a measure can be characterized by the Lipschitz–Hölder exponent $\alpha(q)$ obtained from $\tau(q)$ as

$$\alpha(q) = \frac{\mathrm{d}}{\mathrm{d}q}\tau(q). \tag{6}$$

From Eqs. (2) and (6), $\alpha(q)$ becomes

$$\alpha(q) = \lim_{\epsilon\to 0} \frac{\sum_{\mu(B)\neq 0}[\mu(B)]^q \ln \mu(B)}{Z_\epsilon(q) \ln(\epsilon)}. \tag{7}$$

The exponent $\alpha(q)$ can be obtained through a linear regression of $\sum_{\mu(B)\neq 0}[\mu(B)]^q \ln \mu(B)/Z_\epsilon(q)$ against $\ln \epsilon$.[61] The multifractal spectrum $f(\alpha)$ can then be obtained via the Legendre transform:

$$f(\alpha) = \min_q \{q\ \alpha(q) - \tau(q)\}. \tag{8}$$

3.2. Analogous multifractal analysis

The analogous multifractal analysis (AMFA) is similar to multiaffinity analysis, and can be briefly sketched as follows.[29] First, for a given time series $x(i), i = 1, 2, \ldots, N$, one calculates the integrated series

$$y_q(k) = \sum_{t=1}^{k} |x(t) - x_{\mathrm{ave}}|, \ q \neq 0, \tag{9}$$

where $x_{\rm ave}$ is the average over the whole period. Then one defines a quantity $M_q(L)$ as

$$M_q(L) = [\langle |y_q(j) - y_q(j+L)| \rangle_j^q]^{\frac{1}{q}}, \ (q \neq 0), \quad (10)$$

where $\langle \cdot \rangle_j$ denotes the average over j, $j = 1, 2, \ldots, N - L$. L varies from 1 to N_1 for which the linear fit is good. From the $\ln L$ vs. $\ln M_q(L)$ plane, one can find the power-law scaling within a significant range of L:

$$M_q(L) \propto L^{h(q)}, \ \text{for } q \neq 0. \quad (11)$$

The linear regression of $\ln M_q(L)$ against $\ln L$ will produce exponent $h(q)$.

3.3. Wavelet spectrum

As the wavelet transform of a function can be considered as an approximation of the function, wavelet algorithms process data at different scales or components. At each scale, many coefficients can be obtained and the wavelet spectrum is calculated based on these coefficients. Hence the wavelet spectrum provides some useful information for analysing data. Given a function $f(t)$, one defines its wavelet transform as[62]

$$W_f(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-b}{a}\right) \mathrm{d}t, \quad (12)$$

where b is the position and a is the scale. The scale a in wavelet transform means the ath resolution of the data. Taking $a = j, j = 2^0, 2^1, 2^2, \ldots$, and $b = k, k \in \mathbb{R}$, we obtain the wavelet power spectrum as

$$\text{spectrum } [j] = \sum_k C_{j,k}^2, \ k = 2^0, 2^1, 2^2, \ldots,$$

where $C_{j,k} = W_f(j,k)$.

For simplicity, the scale j can be selected as $j = 1, 3/2, 2, \ldots, 20$, which are more adjacent and can be used to capture more details of the data. The wavelet spectrum is calculated by summing the squares of the coefficients in each scale j. The local wavelet spectrum is defined through the modulus maxima of the coefficients[63] as

$$\text{local spectrum } [j] = \sum_k \tilde{C}_{j,k}^2,$$

where

$$\tilde{C}_{j,k} = \begin{cases} |C_{j,k}|, & \text{if } |C_{j,k}| > |C_{j,k-1}| \ \text{ and } |C_{j,k}| > |C_{j,k+1}|, \\ 0, & \text{otherwise.} \end{cases}$$

The maximum of the wavelet spectrum and the maximum of the local wavelet spectrum will be used in the following for the prediction of structural classes and families of proteins.

In this work, we choose the Daubechies wavelet which is commonly used as a signal processing tool. These wavelet functions are compactly supported wavelets with extremal phase and highest number of vanishing moments for a given support width. They are orthogonal and biorthogonal functions. The Daubechies wavelets can improve the frequency-domain characteristics of other wavelets.[62]

4. Data analysis and discussion

The proteins are selected from Protein Data Bank. The structural classes and families of these proteins are listed in Tables 1 to 4. We link the proteins from the same family or the same superfamily to form 120 long protein sequences, and each structural class contains 30 linked protein sequences. The link order is also listed in these tables. The lengths of all linked proteins vary from 523 to 2281 amino acids.

We use all of the six methods (6-letter model, 12-letter model, 20-letter model, RSWH, SA, SSSA) on twenty kinds of amino acids to convert the amino acid sequences of linked proteins into numerical sequences.

Then we calculate the generalized fractal dimensions D_q, the relevant spectra C_q, the multifractal spectra $f(\alpha)$, and the $h(q)$ curves of the six kinds of numerical sequences of all 120 linked proteins. For example, the D_q curves for the 20-letter model of four proteins are shown in Fig. 1. Their C_q curves are shown in Fig. 2, $f(\alpha)$ curves are shown in Fig. 3, and $h(q)$ curves are shown in Fig. 4. From the curves of $D_q, C_q, f(\alpha)$ and $h(q)$, it can be seen that the numerical sequences from linked proteins are multifractal-like and sufficiently smooth. The C_q curves resemble the phase transition at a certain point, while the $f(\alpha)$ and the $h(q)$ curves indicate the multifractal scaling features of proteins.

Table 1. Proteins from α class and connection order.

fold	superfamily	PDB ID chains	length
long	Ribosomal	1jj2(1)-1yhq(B)-1vq8(E)-1m90(1)	637
alpha-	protein	1yi2(A)-1kqs(C)-1n8r(D)	823
hairpin	L29	1yjw(D)-1s72(G)-1k9m(G)	702
		2otl(C)-1k8a(N)-1vqm(H)	611
		1q86(J)-2b9n(K)-1r73(A)-2j01(C)-1vsa(U)-2gya(3)	807
	Chaperone	1fpo(A)-1hdj-1xbl-1nz6(A)-1iur-1faf(A)-1wjz	717
	J-domain		
	tRNA-	1set(A)-2iy5(A)	773
	binding arm	1eiy(B)-1lrz-1ivs(A)	2075
	Fe,Mn	1ids(A)-1dt0(A)-2nyb(A)-1myb(A)	795
	superoxide	1coj-1wb8(A)-1b06(A)	632
	dismutase	1ma1(A)-1p7g(A)-1unf(X)	665
	N-terminal	2p4k-1kkc(A)-1ixb(A)-1mng(A)	827
	domain	1jr9-1gv3(A)-1y67(A)-1uer(A)	870
DNA/RNA	Homeo-	1p7i(A)- 1akh(B)-1mh3-1k61(A)-1mnm(A)-1lfb	800
binding	domain-	1ic8(A)-1e3o(C)-1gt0(D)-1o4x(A)-1pog-1au7(A)-	933
3-helical	like	-1ftt-1hdp-1ocp	
bundle		1b72(A)-1puf(A)-1lfu(P)-1du6-1bw5-1ig7-9ant(A)-	1027
		-1nk2(P)-1fjl(A)-1mij-1b8i(A)-1jgg(A)-1ftz	
		1xpx-1uhs-2hi3-1wh5-1s7e-1x41-2e1o-1yz8(P)-	1101
		-1x2n-1zq3(P)-1x58-2cuf-2cra	
	'Winged	1bia-1jhf-1aoy-1b4a(A)	750
	helix'	1lea-1f9n(A)-2dt5(A)-1jgs(A)-1lnw(A)-1lj9(A)-1hsj(A)	1360
	DNA-	1i5z(A)-1i1g(A)-2cg4(A)-2cfx(A)-1r1t(A)-1u2w(A)	890
	binding	1o3r-1b9m(A)-1bja(A)-1q1h(A)-1mzb	806
	domain	1o3q-1ft9(A)-2bgc(A)	660
		1omi(A)-2gau-1zyb-2coh-2cyy-1ri7-2cg4(A)	1388
		2cfx(A)-1r1t(A)-1r1u(A)-1y0u(B)-1u2w(a)-1mkm(A)-	1124
		-1jgs(A)-1lnw(A)	
		1lj9(A)-1hsj(A)-1p4x-2frh(A)-2fnp(A)-1s3j(A)-1ub9	1387
		2eth(A)-2hr3(A)-2fbi-1z91(A)-2fbh-2fxa(A)-2a61(A)-2fbk(A)	1269
spectrin	spectrin	2spc(A)-1u5p-1cun(A)-1u4q(A)-1aj3-1owa-1s35(A)	1338
repeat-	Succinate	1chu-1nek(A)	1128
like	dehydrog-	1jnr-2bs2(A)	1303
	enase	1kf6(A)-1l0v(B)	845

The components in a linked protein are selected from the same families or superfamilies.

Table 2. Proteins from β class and connection order.

fold	superfamily	PDB ID chains	length
immunoglobulin-	E set domains	1p7h(L)-2as5(M)-1a02(N)-1owr(M)	1158
like		1s9k(C)-1pzu(B)-1imh(C)-1u3y	968
beta-sandwich		1svc(P)-1nfi(A)-1u42-1ooa(A)	1089
		1nfk(A)-2i9t(A)-1a3q	889
		1bft(A)-1oy3(B)-1k3z(D)-1ikn(D)	755
		1gji(A)-1ttu	752
		1gof-1k3i	1295
		1w8o(A)-1clc	1240
		1ut9-1f1s	1423
		1qba-1cxl	1546
		1cyg-1ciu	1363
		1qho-1gvi(A)	1274
		1ea9(C)-1wzl	1168
		1j0h-1h3g	1189
		1bf2-1eh9	1308
		1svb-1oke(A)	789
		1pjw-1ztx(H)-2gg1-2ala	825
		1im3(A)-2a9d-1ogp(A)	1040
	immunoglobulin	1bww(A)-1rzg(A)-1eeq(A)-2nxy(A)-	998
		-1mju(H)	
		1q9r(A)-1j05(B)-1ors(A)-1clo(H)	776
		1rup(H)-2dlf(H)-1mj8(H)-1qyg(H)-	1138
		-1n4x(H)-1f3d(H)	
common	carbohydrate-binding	1exh-1e5b-1nbc(A)-1g43-1js4(A)	1117
fold of	domain	2ccl(A)-1zv9-1tyj-2bm3	666
diphtheria		2bm3-2b59(A)-1tyj	523
toxin/		1g87(A)-1zv9	788
transcription		1nbc(A)-1g43-1tf4(A)	920
factors/	p53-like	2ac0(A)-1hu8(A)-1t4w-2h1l(A)	952
cytochrome	transcription	1p7h-1imh(C)-1svc-1nfk(A)	1257
f	factors	1a3q-1xbr-1nfi(A)-1gji(A)	1045
		1ljm(A)-1eaq(A)-1mnn	611

The components in a linked protein are selected from the same families or superfamilies.

Table 3. Proteins from $\alpha + \beta$ class and connection order.

fold	superfamily	PDB ID chains	length
Ferredoxin-	4Fe-4S	1dur-2fdn-1clf-1blu-1rgv-7fd1-1bc6-1h98	588
like	ferredoxins	1jb0(A)-1k0t	835
		1xer-1vjw-1fxr(A)-1fxd-1sj1-1feh	925
		1hfe(L)-2c42(A)	1652
	dimeric	1lq9(A)-1q4r-1rjj(A)-1tro(A)-1iuj(A)-1sqe(A)-1tz0(A)	772
	alpha	1xbw(A)-1q8b(A)-1mwq(A)-1mli(A)-1i1g(A)-	1403
	+beta	-2cyy-1ri7-2cfx(A)-2cg4(A)-1tuw-1s7i	
	barrel	1t0t(V)-1vdh(A)-1x7v(A)-1tuv-1y0h(A)-2hiq(A)-	1287
		-1wd6-2asy(A)-1vqs(A)	
		2ap6(A)-1vqy(A)-2gvk-2d3q(A)-2iiz	1299
	Gln-like	2pii-1hwu(A)-1qy7(A)-1vfj(A)-2gnk-2ns1	976
		2nuu(A)-2cz4(A)-1nza(A)-1naq(A)-1p1l-1kr4-	1328
		-1vhf-1o5j-1osc(A)	
		1xk8(A)-1o51-1nh8-1h3d	864
	beta-Grasp	1xd3(A)-1ogw-2j7q(A)-1wrd(A)-2hd5(A)-1uzx(A)-	1891
	(ubiquitin-	-1s1q(A)-2g45(A)-2c7n(A)-2ibi(B)	
	like)	1nbf(A)-2fif(B)-2gmi(B)-1yd8(H)-2hth(B)-2dx5(A)-	990
		-1gjz(A)	
		1zgu(A)-1fxt(A)-2ayo(A)-2fuh(A)-1sif	926
		2uyz(B)-1wyw(B)-2bf8(A)-2pe6(A)-2iy1(B)-1tgz(A)-	1566
		-2iy0(A)-2g4d(A)-1z5s(A)-2io2(C)	
		1wm3-1wm2(A)-2d07(B)-2io0(B)-2ckh(A)-2iyd(B)-	895
		-2io3(A)-1wz0-2awt	

Table 3. (*Continued*).

cystatin-like	cystatin/ monellin	1eqk-1cew(I)-1yvb(A)-1nb5(A)-1nb3(I)	769
		1stf(E)-1r4c(A)-1g96-1roa-1kwi-1pfp	770
Ferredoxin-like	FAD-linked oxidases, C-terminal domain	1qlt(A)-1diq(A)	1081
		1dii(C)-1f0x(A)	651
		1i19(A)-1w10	1095
		1lq9(A)-1iuj(A)-1sqe(A)-1tz0(A)-1xbw(A)-1mwq(A)--1mli(A)-1q8b-2q3b-1rjj(A)	1075
		1tuw-1s7i-1t0t(X)-1vdh(A)-1r6y-1yoh(A)-2asy(A)	1169
		1wd6-2hiq(A)-1vqy(A)-2ap6(A)-2gvk-2hag-2d3q(A)	1535
		2ftr(A)-2fiu(A)-2cb2-1x8d-2op5(A)-2od6(A)	856
	GHMP Kinase, C-terminal domain	1pie-1s4e(A)-1wuu(A)	1170
		1kkh-1kvk-1k47(A)	1049
		1fi4-1uek-1oj4(A)-1mg7(A)	1391
cystatin-like	NTF2-like	1gy6(A)-1jkg(A)-1jn5(B)-1gy7(A)-1of5(A)-1q40(B)	1082
		1o7n(A)-2hmm(B)-2b1xA)-1uli(A)	1573

The components in a linked protein are selected from the same families or superfamilies.

Table 4. Proteins from α/β class and connection order.

fold	superfamily	PDB ID chains	length
The "swivelling" beta/beta /alpha domain	Phosphohistidine domain	1kbl(A)	873
		2fm4-1h6z	1041
		1vbg	876
	RraA-like	1xnj(A)-1q5x(A)-1vi4-1j3l(A)	682
	LeuD/ IlvD-like	7acn-1l5j(A)	1619
		1v7l(A)-2b3y(A)	1051
		2gp4(A)-2hi6(A)	769
	swiveling	1nbw(A)-2d0o(A)	1217
TIM beta/alpha-barrel	ribulose-phoshate binding barrel	1qo2(A)-1vzw-1thf-2a0n	1003
		1ka9(F)-1ox6(A)	807
		1h5y(A)-1rpx(A)-1hly(A)	711
		1tqj(A)-1tqx(A)-2fli(A)-1dbt(A)-1eix(A)	1161
		1x1z(A)-1dqw(A)-1vqt-2ffc	1085
		1q6o-2q8l-2aqw	871
		1pii-1nsj-1lbf-1i4n-1vc4	1409
		1qop-1ujp-1geq(A)-1rd5	1049
	Aldolase	1plx(A)-1o0y(A)-1mzh(A)-1ub3(A)	965
		1n7k(A)-1vcv(A)-1ojx(A)	723
		1hl2(A)-1f6k(A)-2ats(A)	882
		1xky(A)-2ald-1qo5(A)-1fba(A)	1387
		1a5c(A)-1wbh(A)-1mxs-1wa3(A)-1vhc(A)-1gqn	1488
		1ohl-1pv8(A)-1b4k(A)-1lbs	1332
		1n8f(A)-1hfb-1d9e(A)-1o60(A)-1pe1(A)	1563
	Cobalamin(VB12)-dependent enzymes	7req(A)-1i9c(B)-1iwp(A)-1xrs(A)	2281
	N-(deoxy) ribosyltransferase	1f8y(A)-2f62(A)-1sz1(A)-1t1j(A)-1r1z(A)	868
flavodoxin-like	Class I glutamine amidotransferase-like	1gpm(A)-1c30(A)-1qdl(A)	2020
		1i1q(A)-1k9v(F)-2nv0(A)-1s1m(A)	1462
		1vco-1t3t	1853
		2a9v-1wl8-2abw(A)	628
		1g2i(A)-1p5f-1fye-1t0b(A)-2ghr	1138

The components in a linked protein are selected from the same families or superfamilies.

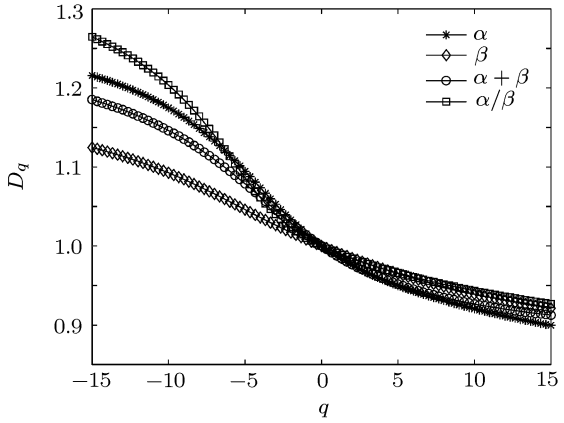


Fig. 1. D_q curves for four linked proteins which are the first linked proteins in each table using the 20-letter model.

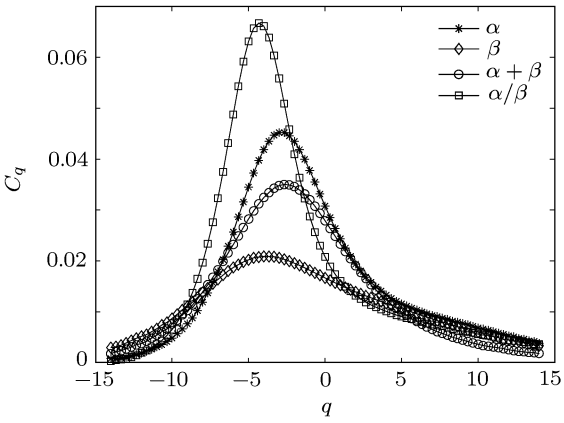


Fig. 2. C_q curves for four linked proteins which are the first linked proteins in each table using the 20-letter model.

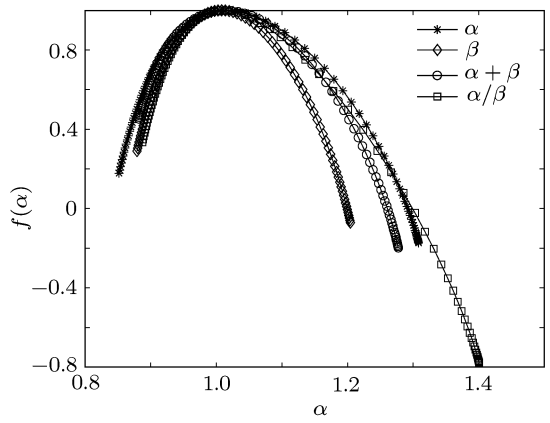


Fig. 3. $f(\alpha)$ curves for four linked proteins which are the first linked proteins in each table using the 20-letter model.

In wavelet analysis, the choice of a wavelet function should be carefully considered. Different wavelet functions each represent a given function with different approximation components. In this work, we chose the commonly used Daubechies wavelet (db2) to solve this problem. With the db2 function, we compute the maximum of the wavelet spectrum and the maximum of the local wavelet spectrum for the six kinds of numerical sequences of all 120 linked proteins. These spectrum parameters are used as candidates to construct parameter spaces.

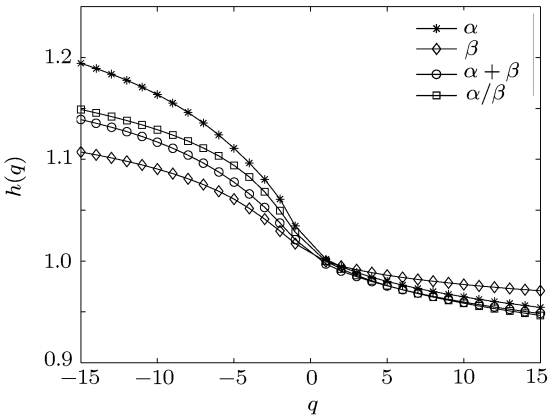


Fig. 4. $h(q)$ curves for four linked proteins which are the first linked proteins in each table using the 20-letter model.

In the prediction of the structure of a protein, it is not enough to have a small number of parameters because the features of a protein cannot be easily captured with a few fixed parameters in our method. That is why we consider different parameter spaces to classify the proteins. To make the methods more feasible, some traditional indexes are chosen to discriminate the linked proteins. These parameters are selected from D_q, ΔD_q, $C_{q\,\max}$, $\alpha_{\max}$, $h(q)$, and local spectrum. The value D_1 is called the information dimension, and D_2 the correlation dimension.

After evaluating many combinations of these parameters, we find the following parameters are useful for the structural classification and family identification of these linked proteins. They are D_{15} from the multifractal analysis of the solvent accessibility, the range of D_q (denoted by $\Delta D = D_{-15} - D_{15}$) of the RSWH, $h(1)$ from analogous multifractal analysis of the 20-letter model, the maximum of C_q (denoted by $C_{q\,\max}$) from the multifractal analysis of the RSWH, $h(2)$ of the 20-letter model, the maximum of local spectrum (denoted by local spectrum$_{\max}$) of the SSSA, the maximum of α (denoted by $\alpha_{\max}$) from the multifractal analysis of the RSWH, local spectrum$_{\max}$ of the SA, and the local spectrum$_{\max}$ of the 12-letter model. We use these parameters to construct parameter spaces in which one linked protein is represented by a point. In the 3D space (D_{15} for SA, ΔD for

RSWH, and $h(1)$ for 20-letter), the proteins from the α class gather together and can be separated from the proteins from the other classes. The results are shown in Fig. 5. In the 3D space ($C_{q\ \max}$ for RSWH, $h(2)$ for 20-letter, and local spectrum$_{\max}$ for SSSA), the proteins from the β class group together and can be separated from the proteins from the $\alpha+\beta$ class and the α/β class (Fig. 6). In the 3D space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, and local spectrum$_{\max}$ for 12-letter), the proteins from the $\alpha+\beta$ class and those from the α/β class can be separated (Fig. 7).

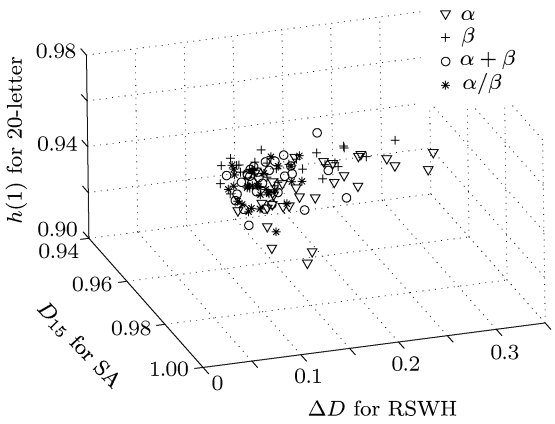


Fig. 5. Space (D_{15} for SA, ΔD for RSWH, $h(1)$ for 20-letter) in which the linked proteins from the α class can be separated from the linked proteins from the other classes.

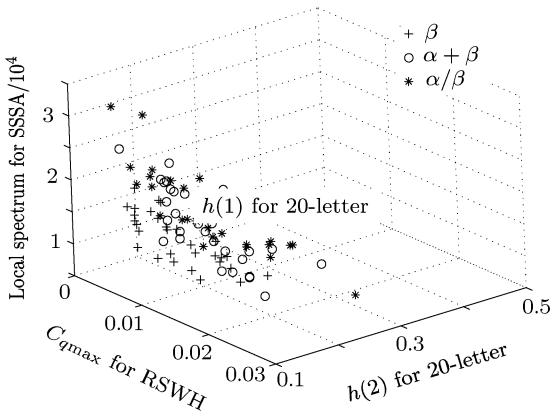


Fig. 6. Space ($C_{q\ \max}$ for RSWH, $h(2)$ for 20-letter, local spectrum$_{\max}$ for SSSA) in which the linked proteins from the β class can be separated from the linked proteins from the $\{\alpha+\beta, \alpha/\beta\}$ classes.

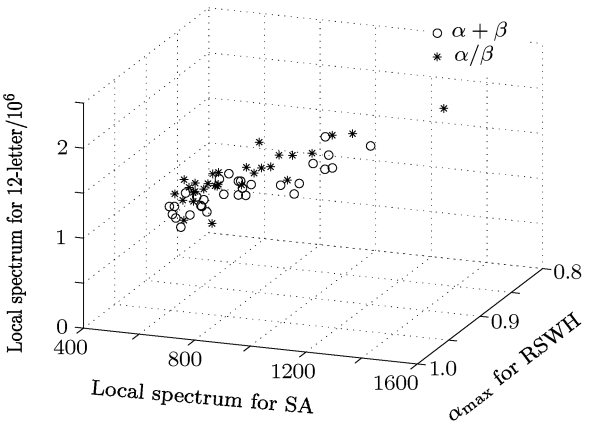


Fig. 7. Space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, local spectrum$_{\max}$ for 12-letter) in which the linked proteins from the $\alpha+\beta$ class can be separated from the linked proteins from the α/β class.

We propose the following three-step procedure to separate proteins:

Step 1 Separating the α proteins from the $\{\beta, \alpha+\beta, \alpha/\beta\}$ proteins in the space (D_{15} for SA, ΔD for RSWH, and $h(1)$ for 20-letter);

Step 2 Separating the β proteins from the $\{\alpha+\beta, \alpha/\beta\}$ proteins in the space ($C_{q\ \max}$ for RSWH, $h(2)$ for 20-letter, and local spectrum$_{\max}$ for SSSA);

Step 3 Separating the $\alpha+\beta$ proteins from the α/β proteins in the space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, and local spectrum$_{\max}$ for 12-letter).

In order to assess the classification of the selected linked proteins, we use Fisher's linear discriminant algorithm[64] to calculate the discriminant accuracies of the three steps.

Fisher's discriminant algorithm is used to find a classifier in the parameter space for a training set $\boldsymbol{H} = \{x_1, \ldots, x_n\}$, which is partitioned into $n_1 \leq n$ training vectors in a subset $\boldsymbol{H}_1$ and $n_2 \leq n$ training vectors in a subset $\boldsymbol{H}_2$, where $n_1 + n_2 = n, \boldsymbol{H} = \boldsymbol{H}_1 \cup \boldsymbol{H}_2$ and each x_i is a point in the parameter space. We should find a parameter vector $\boldsymbol{w} = (w_1, w_2, w_3)$ for a 3D space such that $\{y_i = \boldsymbol{w}^{\mathrm{T}} x_i\}_{i=1}^{n}$ can be classified into two classes in the space. Denoting

$$m_j = \frac{1}{n_j} \sum_{x_i \in \boldsymbol{H}_j} x_i, \quad j = 1, 2,$$

$$S_j = \sum_{x_i \in H_j} (x_i - m_j)(x_i - m_j)^{\mathrm{T}}, \quad j = 1, 2,$$

$$S_w = S_1 + S_2;$$

then the parameter vector w can be estimated as $S_w^{-1}(m_1 - m_2)$.[28] Fisher's discriminant rule becomes assigning x to H_1 if $(m_1-m_2)^{\mathrm{T}} S_w^{-1}(x-\frac{1}{2}(m_1+m_2)) > 0$ and to H_2 otherwise.[64]

We treat the whole data set as the training set because the total number of points is small. The discriminant accuracies are defined as

$$P_{H_1} = n_{H_1}/n_1, \quad P_{H_2} = n_{H_2}/n_2,$$

where n_{H_1} and n_{H_2} are the number of correctly discriminated H_1 elements and that of H_2 elements respectively.

We denote all the α linked proteins as H_1, the remaining $\{\beta, \alpha+\beta, \alpha/\beta\}$ linked proteins as H_2 in the 3D space (D_{15} for SA, ΔD for RSWH, $h(1)$ for 20-letter); all the β linked proteins as H_1, the remaining $\{\alpha+\beta, \alpha/\beta\}$ linked proteins as H_2 in the 3D space ($C_{q\ \max}$ for RSWH, $h(2)$ for 20-letter, local spectrum$_{\max}$ for SSSA); all $\alpha+\beta$ linked proteins as H_1 and all α/β linked proteins as H_2 in the space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, local spectrum$_{\max}$ for 12-letter). The estimated parameters $w = (w_1, w_2, w_3)$ in Fisher's discriminant algorithm and the accuracies are listed in Table 5. It can be seen that our methods work well on the selected linked proteins. The classification accuracies are 86.67% and 84.44% for P_{H_1} and P_{H_2} respectively in Step 1 ; 96.67% and 80% for P_{H_1} and P_{H_2} respectively in Step 2; 80% and 83.33% for P_{H_1} and P_{H_2} respectively in Step 3.

Table 5. The parameters in Fisher's discriminant and the accuracies for the proteins.

H_1	H_2	w_1	w_2	w_3	P_{H_1}	P_{H_2}
α	$\beta, \alpha+\beta, \alpha/\beta$	3.7200	0.4321	–1.3826	86.67%	84.44%
β	$\alpha+\beta, \alpha/\beta$	–4.9992	–0.4726	–0.0000	96.67%	80.00%
$\alpha+\beta$	α/β	1.0574	0.0002	–0.0000	80.00%	83.33%

Prediction accuracies can be evaluated by several methods, and the jack-knife test is considered to be the most objective.[22] It is called leave-one-out test, and each linked protein in the data set is in turn singled out as a test protein while the predictor is trained by the remaining linked proteins. Using our method, the accuracies for 120 linked protein data are listed in Table 6. The average prediction accuracy for the four structural classes of these linked proteins is 86.47%.

Table 6. Predicted results for 120 linked proteins dataset with Jack-knife test and comparison with other methods.

dataset	reference	α	β	$\alpha+\beta$	α/β	mean
linked proteins	**This paper**	**86.2%**	**96.3%**	**80.1%**	**83.3%**	**86.47 %**
linked proteins	GSSP[15]	0%	0%	0 %	100%	25%
linked proteins	PHD[65]	36.7%	50%	93.3%[*]	100%[*]	70%
linked proteins	PREDATOR[69]	33.33%	36.67%	96.67%[*]	100%[*]	66.67%
linked proteins	SOPMA[70]	10%	26.67%	96.67%[*]	100%[*]	58.34%
linked proteins	HNN[71]	20%	20%	96.67%[*]	100%[*]	59.17%

[*] The proteins from the $\alpha+\beta$ class and those from the α/β class are clustered as mixed class in the web server. They cannot be discriminated only with the amounts of helices and strands in these web servers.

Comparison with other methods. There are many web servers in predicting protein secondary structure. In order to compare our method with the existing techniques, we chose five websites to test our data set. The first web server is GSSP (http://218.65.61. 89:8080/bioinfo/GSSP), which is a predictor that uses grey modeling and pseudo amino acid components to predict protein structural classes.[15] The website classifies proteins into four classes. At this website, we test the 120 linked proteins, and find that only the proteins from the α/β class can be successfully predicted while all of the proteins from the other three classes are incorrectly predicted as the α/β class. The mean accuracy is 25%. The second website is the PHD server[65] on the website http://npsa-pbil.ibcp.fr, which is an automatic service that predicts aspects of the secondary structures of proteins. We test the 120 linked pro-

teins with its PHD server. At the server, we cannot immediately obtain the discriminant results because it reports the secondary structure with the amounts of helices, strands and coils. With some structural class definitions, one can define the four classes: α (helix amount $> 15\%$, strand amount $< 10\%$), β (helix amount $< 15\%$, strand amount $> 10\%$), $\alpha + \beta$ (helix amount $> 15\%$, strand amount $> 10\%$), and α/β (helix amount $> 15\%$, strand amount $> 10\%$).[66] In fact, one cannot discriminate between the $\alpha + \beta$ class and the α/β class based only on the amounts of helices and strands. We find the proteins from the $\alpha + \beta$ class and the α/β class can be successfully predicted only when we merge the two classes into a so-called mixed class.[67,68] The mean accuracy is 70%. The third server we test is the PREDATOR[69] on the website http://npsa-pbil.ibcp.fr, and the mean accuracy is 66.67%. The fourth prediction server is the SOPMA,[70] and the overall accuracy is 58.34%. The fifth server is the HNN[71] on the website http://npsa-pbil.ibcp.fr, and the mean accuracy is 59.17%. We also list the prediction accuracies of these methods in Table 6, which confirms that our method works well in clustering proteins.

The family group. Now we want to further look into each structural class. We list some of the linked proteins from the same structural class according to their families in the above parameter spaces and show the results in Figs. 8–11. One can find that in these parameter spaces, the linked proteins from the same family tend to group together and can be separated from the other proteins in different families. Those proteins having similar amino acid sequences cluster together to reveal the information about their structures and functions. This feature can be used in the prediction of an unknown protein from the same family.

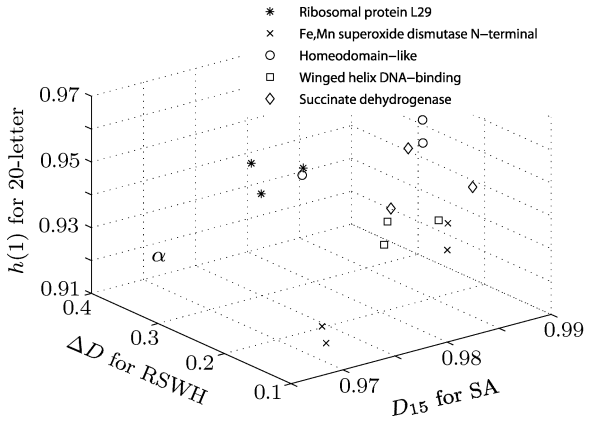


Fig. 8. Space (D_{15} for SA, ΔD for RSWH, and $h(1)$ for 20-letter) in which the linked proteins from the same family of the α class tend to group together.

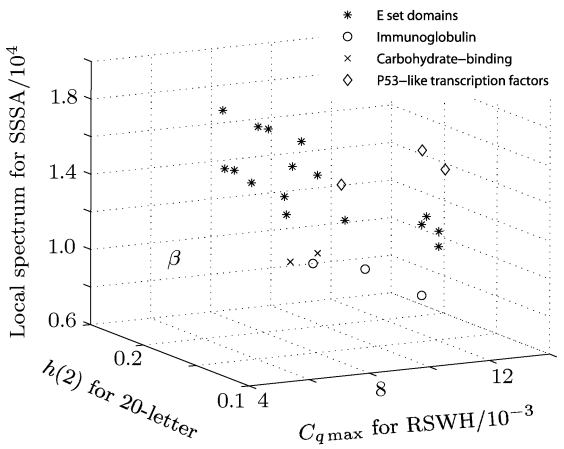


Fig. 9. Sspace ($C_{q\ \max}$ for RSWH, $h(2)$ for 20-letter, and local spectrum$_{\max}$ for SSSA) in which the linked proteins from the same family of the β class tend to group together.

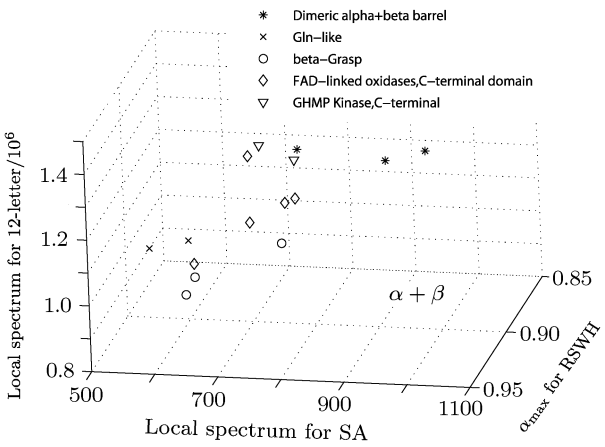


Fig. 10. Space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, and $\alpha_{\min}$ for 12-letter) in which the linked proteins from the same family of the $\alpha + \beta$ class tend to group together.

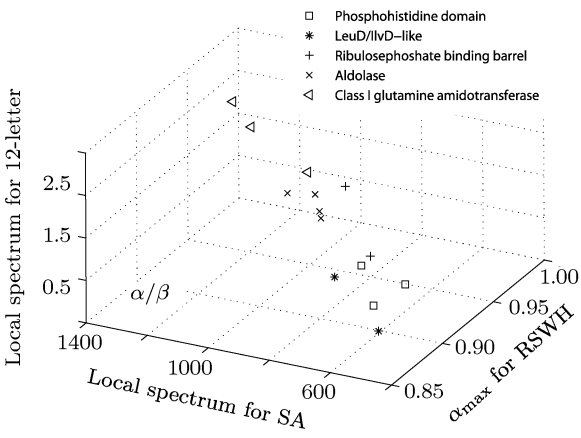


Fig. 11. Space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, and $\alpha_{\min}$ for 12-letter) in which the proteins from the same family of the α/β class tend to group together.

The order effect. The different connection orders of proteins yield different linked protein sequences due to the rearrangement of amino acid sequences. We change the link order of the proteins to see whether the multifractal exponents change. We find the link order does not affect the D_q and $h(q)$ curves (see the D_q curves in the left panel of Fig. 12 and the $h(q)$ curves in the left panel of Fig. 13). We also calculate the discriminant rates by changing the link orders of the data and find that the discriminant rates only vary slightly in those parameter spaces (the results are not shown here, but are available upon request). We also want to know whether the map order in the 6-letter model, or 12-letter model or 20-letter model would affect our results. We test another arbitrary map order in the 20-letter model for a linked protein (map $A = 6$, $R = 2$, $N = 10$, $D = 15$, $C = 3$, $Q = 16$, $E = 9$, $G = 5$, $H = 0$, $I = 11$, $L = 13$, $K = 18$, $M = 1$, $F = 19$, $P = 17$, $S = 7$, $T = 4$, $W = 12$, $Y = 8$, $V = 14$). We find that the D_q and $h(q)$ curves using this order are almost identical to those using the previous order respectively (see the D_q curves in the right panel of Fig. 12 and the $h(q)$ curves in the right panel of Fig. 13). Because C_q and $f(\alpha)$ can be calculated from the D_q curve, they are not dramatically affected by the link order and the map order in the models. The wavelet parameters are affected by the map order and the link order.

The important problems of protein structure classification and family identification have been looked at from different angles by many researchers. In this paper, we focus on capturing the features of the protein families similar to those in Refs. [44] and [51] using multifractal techniques and wavelet analysis.

Once these features are captured and represented, it is possible to use them for identifying new members of the families as noted in Refs. [44] and [51]. Further investigation is needed to apply this method to real proteins.

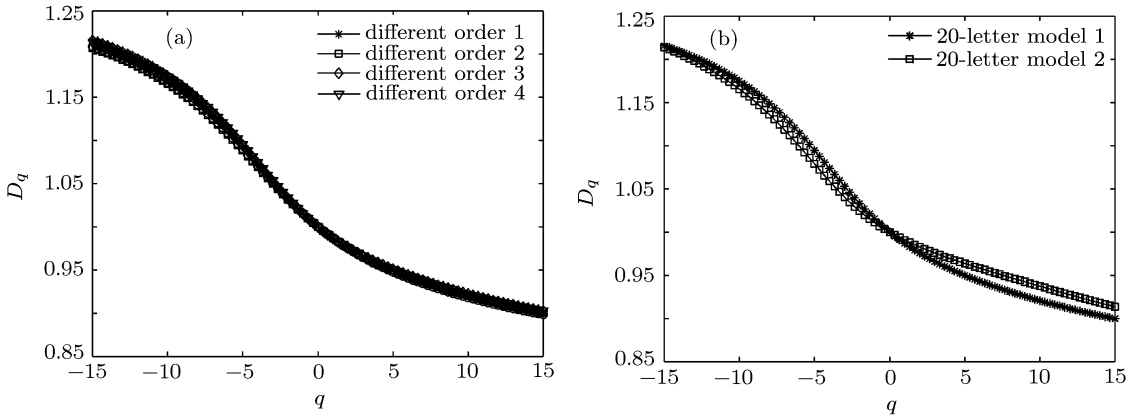


Fig. 12. The D_q curves of the first linked α protein in Table 1 using the 20-letter model with different orders to link (a). The D_q curves for the same linked proteins using different map orders in the 20-letter model (b).

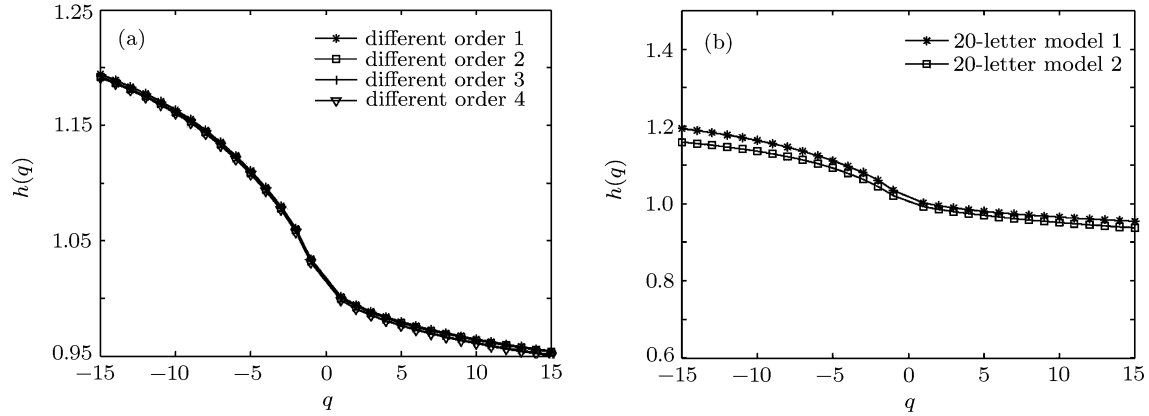


Fig. 13. The $h(q)$ curves of the first linked α proteins in Table 1 using the 20-letter model with different orders to link (a). The $h(q)$ curves for the same linked proteins using different map orders in the 20-letter model (b).

5. Conclusions

The ordinary multifractal analysis, the analogous multifractal analysis and the wavelet spectrum of proteins can provide information about their structures and functions (families).

From the curves of $D_q, C_q, f(\alpha)$ and $h(q)$, it follows that the numerical sequences of linked proteins are multifractal-like and sufficiently smooth. The C_q curves resemble the phase transition at a certain point, while the $f(\alpha)$ and $h(q)$ curves indicate the multifractal scaling features of proteins.

Some parameter spaces are constructed by using the parameters selected from these methods to distinguish proteins. Numerical results show that the α proteins can be separated from the $\{\beta, \alpha+\beta, \alpha/\beta\}$ proteins in the 3D space (D_{15} for SA, ΔD for RSWH, and $h(1)$ for 20-letter). The β proteins can be separated from the $\{\alpha+\beta, \alpha/\beta\}$ proteins in the 3D space ($C_{q\ \max}$ for RSWH, $h(2)$ for 20-letter, and local spectrum$_{\max}$ for SSSA), and the $\alpha+\beta$ proteins from the α/β proteins in the space ($\alpha_{\max}$ for RSWH, local spectrum$_{\max}$ for SA, and local spectrum$_{\max}$ for 12-letter). Fisher's linear discriminant algorithm is applied to the assessment of clustering of the 120 linked proteins.

The 6-letter model, the 12-letter model and the 20-letter model of proteins contain more information than any single model, and the results show that these models are efficient in the clustering of protein structures.

These multifractal analyses and wavelet spectrum are useful to identify the protein family. It is seen that the proteins from the same family tend to group together. It is possible to use the captured features of such protein families to identify new members of the families.

References

[1] Abascal F and Valencia A 2003 *Proteins: Structure, Function, and Genetics* **53** 683
[2] Weisser D and Klein-Seetharaman J 2004 *Proceeding of ACM Symposium on Applied Computing* pp.154-161
[3] Baker D and Sali A 2001 *Science* **294** 93
[4] Gu F, Chen H and Ni J 2008 *BMC Bioinformatics* **9** (Suppl 6) S5
[5] Anfinsen C 1973 *Science* **181** 223
[6] Baxevanis A D and Ouellette B F F (Eds) 2005 *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (3rd edition) (John Wiley & Sons, Inc.)
[7] Wang Z and Yuan Z 2000 *Proteins: Structure, Function, and Genetics* **38** 165
[8] Kurgan L and Homaeian L 2006 *Pattern Recognition* **39** 2323
[9] Zhang Y 2008 *Current Opinion in Structural Biology* **18** 342
[10] Cao Y, Liu S, Zhang L, Qin J, Wang J and Tang K 2006 *BMC Bioinformatics* **7** 20
[11] Chou K C 2000 *Curr. Protein Pept. Sci.* **1** 171
[12] Chou K C and Maggiora G M 1998 *Protein Engineering* **11** 523
[13] Kedarisetti K D, Kurgan L and Dick S 2006 *Biochem. Biophys. Res. Commun.* **348** 981
[14] Chou K C 2001 *Proteins: Structure, Function, and Genetics* **43** 246
[15] Xiao X, Shao S, Huang Z and Chou K C 2006 *J. Comput. Chem.* **27** 478
[16] Chou K C 1995 *Proteins* **21** 319
[17] Kurgan L and Chen K 2007 *Biochem. Biophys. Res. Commun.* **357** 453
[18] Kurgan L, Cios K and Chen K 2008 *BMC Bioinformatics* **9** 226
[19] Jahandideh S, Abdolmaleki P, Jahandideh M, Sadat Hayatshahi S H 2007 *J. Theor. Biol.* **244** 275
[20] Cai Y D and Zhou G P 2000 *Biochimie* **82** 783
[21] Cai Y D, Liu X J, Xu X and Zhou G P 2001 *BMC Bioinformatics* **2** 3
[22] Cai Y D, Liu X J, Xu X B and Chou K C 2003 *J. Theor. Biol.* **221** 115
[23] Feng K Y, Cai Y D and Chou K C 2005 *Biochem. Biophys. Res. Commun.* **334** 213
[24] Chou K C 2005 *Curr. Protein Pept. Sci.* **6** 423
[25] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 *J. Mol. Biol.* **247** 536
[26] Andreeva A, Howorth D, Brenner S E, Hubbard T J, Chothia C and Murzin A G 2004 *Nucleic Acids Res.* **32** D226-9.
[27] Orengo C A, Michie A D, Jones S, Jones D T, Swindells M B and Thornton J M 1997 *Structure* **5** 1093
[28] Yu Z G, Anh V, Lau K S and Zhou L Q 2006 *Phys. Rev.* E **73** 031920
[29] Yang J Y, Yu Z G and Anh V 2009 *Chaos, Solutions and Fractals* **40** 607
[30] Zhou Y, Yu Z G and Anh V 2007 *Phys. Lett.* A **368** 314
[31] Yang J Y, Peng Z L, Yu Z G, Zhang R J, Anh V and Wang D S 2009 *J. Theor. Biol.* **257** 618
[32] Feng J, Liu J H and Zhang H G 2008 *Acta Phys. Sin.* **57** 6868 (in Chinese)
[33] Chen Y P, Fu P P, Shi M H, Wu J F and Zhang C B 2009 *Acta Phys. Sin.* **58** 7050 (in Chinese)
[34] Mandelbrot B 1983 *The Fractal Geometry of Nature* (New York: Academic Press)
[35] Yu Z G, Anh V and Lau K S 2001 *Phys. Rev.* E **64** 031903
[36] Anh V V, Lau K S and Yu Z G 2002 *Phys. Rev.* E **66** 031910
[37] Yu Z G, Anh V and Lau K S 2003 *Phys. Rev.* E **68** 021913
[38] Yu Z G, Anh V and Lau K S 2004 *J. Theor. Biol.* **226** 341
[39] Yang J Y, Zhou Y, Yu Z G and Anh V 2008 *BMC Bioinformatics* **9** 113

[40] Yu Z G, Anh V V, Gong Z M and Long S C 2002 *Chin. Phys.* **11** 1313
[41] Jeffrey H J 1990 *Nucleic Acids Res.* **18** 2163
[42] Han J J and Fu W J 2010 *Chin. Phys.* B **19** 010205
[43] Fiser A, Tusnady G E and Simon I 1994 *J. Mol. Graphics* **12** 302
[44] Basu S, Pan A, Dutta C and Das J 1997 *J. Mol. Graph. Model.* **15** 279
[45] Gao J and Xu Z Y 2009 *Chin. Phys.* B **18** 370
[46] Gao J, Jiang L L and Xu Z Y 2009 *Chin. Phys.* B **18** 4571
[47] Yu Z G, Xiao Q J, Shi L, Yu J W and Anh V 2010 *Chin. Phys.* B **19** 068701
[48] Chen H, Gu F and Liu F 2005 *Proceedings of the 2005 IEEE, Engineering in Medicine and Biology* 27^{th} *Annual Conference*
[49] Marsolo K and Ramamohanarao K 2006 *Proceedings of the* 15^{th} *ACM International Conference on Information and Knowledge Management* pp24–33
[50] Qiu J D, Liang R P, Zou X Y and Mo J Y 2003 *Talanta* **61** 285
[51] Deschavanne P and Tuffery P 2008 *Biochimie* **90** 615
[52] Bebek G and Yang J 2007 *BMC Bioinformatics* **8** 335
[53] Chou P Y and Fasman G D 1974 *Biochemistry* **13** 211
[54] Brown T A 1998 *Genetics* (3rd Edition) (CHAPMAN & HALL, London)
[55] Giuliani A, Benigni R, Sirabella P, Zbilut J P and Colosimo A 2000 *Biophys. J.* **78** 136
[56] Macdonald J R and Johnson W C 2001 *Protein Science* **10** 1172
[57] Bordo D and Argos P 1991 *J. Mol. Biol.* **217** 721
[58] Rose G D, Geselowitz A R, Lesser G J, Lee R H and Zehfus M H 1985 *Science* **229** 834
[59] Hasley T C, Jensen M H, Kadanoff L P, Procaccia I and Shraiman B I 1986 *Phys. Rev.* A **33** 1141
[60] Canessa E 2000 *J. Phys.* A **33** 3637
[61] Balafas J S and Dewey T G 1995 *Phys. Rev.* E **52** 880
[62] Chui C K 1992 *An Introduction to Wavelets* (San Diego: Academic Press Professional)
[63] Arneodo A, Bacry E and Muzy J F 1995 *Physica* A **213** 232
[64] Mardia K V, Kent J T and Billy J M 1979 *Multivariate Analysis* (London: Academic Press)
[65] Rost B and Sander C 1993 *J. Mol. Biol.* **1232** 584
[66] Nakashima H, Nishikawa K and Ooi T 1986 *J. Biochem.* **99** 153
[67] Eisenhaber F, Imperiale F, Argos P and Frommel C 1996 *Proteins* **25** 157
[68] Eisenhaber F, Frommel C and Argos P 1996 *Proteins* **25** 169
[69] Frishman D and Argos P 1996 *Protein Engneering* **9** 133
[70] Geourjon C and Deleage G 1995 *Comput. Appl. Biosci.* **11** 681
[71] Guermeur Y 1997 *Combinaison de Classifieurs Statistiques, Application a la Prediction de Structure Secondaire des Proteines*, PhD Thesis, Universite Paris