Fourth International Conference on Natural Computation

Protein structure classification based on chaos game representation and multifractal analysis

Jian-Yi Yang[1], Zu-Guo Yu[1,2*]
[1]School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China.

Vo Anh[2]
[2]School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia.

Abstract

Classification of protein structures is important in the prediction of the tertiary structures of proteins. In this paper, we propose to decompose the chaos game representation of proteins into two time series, from which the protein sequences can be uniquely reconstructed. Multifractal analysis is applied to measures constructed from these two time series. A total of 26 characteristic parameters are calculated for each protein, which are used to construct a 26-dimensional space. Each protein is represented by one point in this space. A procedure is proposed to classify the structures of 100 large proteins consisting of four structural classes. Fisher's linear discriminant algorithm demonstrates that the average accuracy for our classification can reach 84.67%. Compared with the results for the 46 large proteins reported before, the method proposed here has much better performance.

1. Introduction

The three-dimensional ($3D$) structure of a protein is determined by its amino acid sequence (primary structure) via the process of protein folding [1]. In order to explore the mechanism of the protein folding process, some theoretical works have suggested the designability and other concepts based on lattice models [2-5]. It is possible to predict the $3D$ structures of proteins from their primary structures directly. But this is a challenging problem as there is no simple rule to map the primary structure into the corresponding $3D$ structure of a protein. Four main structural classes of proteins were recognized based on the types and arrangement of their secondary structural elements [6]. They are the α class, the β class and those with a mixture of α and β shapes called the $\alpha+\beta$ class and the α/β class. Information about protein structural classes may shed light on the above problem. For example, the searching scope of conformation will be reduced if the structural class of the protein under study is known [7]. Hence, it is informative to predict the protein structural classes from the primary structures directly.

For the problem on identification of protein structural classes from the amino acid sequences of proteins, there are many methods which have been proposed [8]. Yu *et al.* [9] used the hydrophobic free energy and solvent accessibility of proteins to construct several parameter spaces. They found that some spaces could be used to distinguish and cluster the 43 selected large proteins from the four structural classes. We recently discussed the clustering of 49 large proteins via multifractal analysis and a 6-letter model [10]. In this paper, we classify 100 large proteins structures to structural classes by some new methods, which improve the results in Yu *et al.* [9] and Yang *et al.* [10].

Chaos game representation (CGR) of protein structures was first proposed by Fiser *et al.* [11]. Later Basu *et al.* [12] and Yu *et al.* [13] proposed several other kinds of CGR of proteins. In this paper, we use the CGR method in Fiser *et al.* [11] to convert the primary structures of proteins into two time series, from which the primary structures of proteins can be reconstructed uniquely. Then we apply multifractal analysis to the time series to calculate certain characteristic parameters of proteins. A total of 26 parameters are achieved from the multifractal analysis. These parameters are used to construct a space in order to classify protein structures. Fisher's linear discriminant algorithm demonstrates that our classification method is satisfied.

2. Methods

Chaos game representation of proteins. Chaos game representation was firstly proposed to analyze nucleotide sequences [14]. The technique of CGR has been generalized and applied to analyze both the primary and secondary

*Corresponding author Zu-Guo Yu, e-mail: yuzg1970@yahoo.com or z.yu@qut.edu.au

978-0-7695-3304-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ICNC.2008.295

structures of proteins [11]. CGR is a useful way to visualize protein sequences. We recapture the concept briefly here.

As proteins consist of 20 kinds of amino acids, a 20-sided regular polygon can be used to visualize protein sequences. Firstly, the 20 kinds of amino acids are placed on the vertices of the polygon in a certain manner. In this paper, we use the alphabetical order to arrange the amino acids, which is shown in Figure 1 for a typical protein.

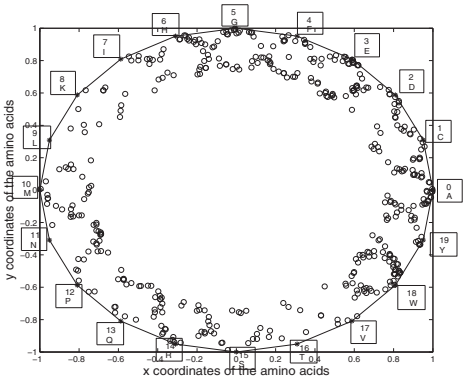

Figure 1. The chaos game representation of protein 1B89 with the arrangement of the amino acid residues in alphabetical order.

If the circle around the main polygon is the unit circle, the (x, y) coordinates of a certain vertex i can be obtained as

$$\begin{cases} v_{i,x}^0 = \cos(2\pi * i/20), \\ v_{i,y}^0 = \sin(2\pi * i/20), \end{cases} \quad (1)$$

where i ranges from 0 to 19.

Given one protein sequence with length L, the coordinates of the 0th point is $(p_{0,x}, p_{0,y}) = (0, 0)$ and the coordinates of the mth point can be determined:

$$\begin{cases} p_{m,x} = (v_{a_m,x}^0 - p_{m-1,x}) * \kappa + p_{m-1,x}, \\ p_{m,y} = (v_{a_m,y}^0 - p_{m-1,y}) * \kappa + p_{m-1,y}, \end{cases} \quad (2)$$

where κ is a constant, which is set as 0.865 by Fiser *et al.* [11] to satisfy the condition that the circles around the inner polygons touch each other but do not overlap; a_m ranges from 0 to 19 and is determined by both the amino acid type of the mth amino acid along the protein sequence and the arrangement of the amino acids on the vertices. It can be seen from Eq.(2) that the coordinates of the mth point are in fact determined by a couple of residues (current residue and nearest preceding residue). We give the CGR of protein **1B89** in Figure 1 as an example.

It can be seen that the protein sequences can be reconstructed uniquely from CGR by the maps

$$\begin{cases} v_{m,x} = [p_{m,x} - (1-\kappa)p_{m-1,x}]/\kappa, \\ v_{m,y} = [p_{m,y} - (1-\kappa)p_{m-1,y}]/\kappa, \end{cases} \quad (3)$$

where m ranges from 1 to L. We can compare $(v_{m,x}, v_{m,y})$ with the (x, y) coordinates of the 20 kinds of amino acids $\{(v_{i,x}^0, v_{i,y}^0)\}_{i=0}^{19}$ to decide the amino acid type corresponding to the mth point. Therefore, we can conclude that all information in the protein sequence is contained in the CGR. Fiser *et al.* [11] concluded that CGR could be used to test protein structure prediction methods. Motivated by this, it seems possible to classify protein structures to protein structural classes using the CGR of proteins. Noticing that the CGR of proteins is determined by the (x, y) coordinates, we decompose the CGR into two time series: $\{p_{m,x}\}_{m=1}^{L}$ and $\{p_{m,y}\}_{m=1}^{L}$. These two kinds of time series also contain all information in the protein sequence and we will analyze them by the multifractal analysis. We denote the former the x time series and the latter the y time series. Figure 2 shows the corresponding time series from CGR of protein **1B89**.

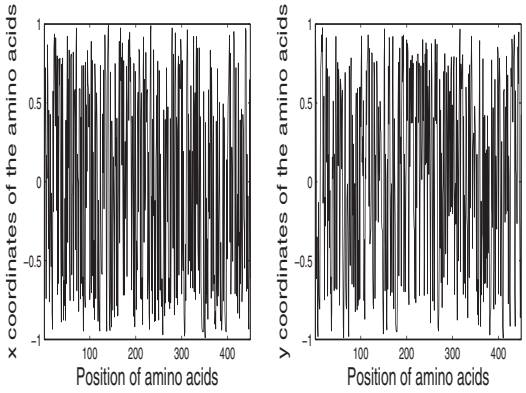

Figure 2. The two corresponding time series from chaos game representation of proteins 1B89.

Measure for the time series and Multifractal analysis (MFA). We construct a measure from the time series with a method same as that of Yu *et al.* [15] and Yang *et al.* [10]. Because of the non-negative condition required to construct a measure, we add 1 (as the smallest value is -1) to the time series of protein proposed above. We [10] have discussed before that such value may affect the values of parameters in multifractal analysis, but it doesn't have significant influences on the final classifying accuracies.

It should be emphasized that the ordering of the time series is rather important in the definition of this measure. By disordering an old times series, one new time series can be got but the the measure for the new time series is different

from the old one. In another words, the measure defined here is sensitive to the ordering of the time series. Therefore, for a protein sequence, the ordering of the amino acids acquires a greater importance than composition in the measure.

The most common algorithms of multifractal analysis are the so called *fixed-size box-counting algorithms*. The definition of the multifractal exponents $\tau(q), D_q, C_q, \alpha(q)$ and $f(\alpha)$ can be see in Refs. [9] and Yang *et al.* [10].

3. Data, results and discussions

The methods introduced in the previous sections can only be used for long protein sequences (corresponding to large proteins) as declared in Yu *et al.* [16]. The amino acid sequences of 100 large proteins are downloaded from the RCSB Protein Data Bank (http://www.rcsb.org/pdb/home/), in which 46 has been studied in Yang *et al.* [10]. In fact, there were 49 large proteins in Yang *et al.* [10]. However, because the class information of three proteins (1F1S, 1OY6 and 1T3T) given in Yang *et al.* [10] and Yu *et al.* [16] is updated, only the remaining 46 proteins are used here. These 100 proteins, which are listed in Table 1, belong to four structural classes.

Given an amino acid sequence of a protein, we use the CGR to covert it into two time series. Then a measure is constructed from each time series as in Ref. [15]. Now we can apply the MFA to the constructed measures. As examples, the D_q, C_q, $\alpha(q)$ curves and the multifractal spectrum $f(\alpha)$ of the x time series of protein **1B89** are shown in Figure 3.

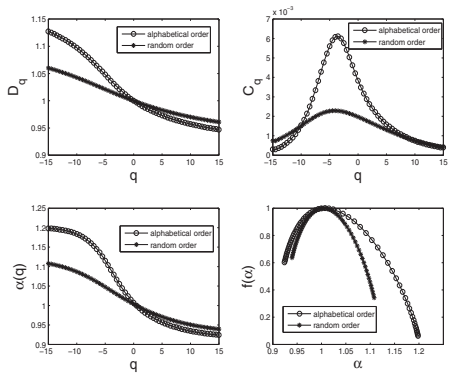

Figure 3. The four kinds of multifractal curves for x time series of protein 1B89 with two kinds of arrangement manners of amino acid residues.

From these calculations, a total of 26 parameters is achieved. They are listed in Table 2.

Table 1. The 100 proteins represented by the PDB ID in the PDB database. The first four letters are the PDB ID and the fifth letter stands for the chain selected. The numbers in the bracket after the proteins are the lengths for the proteins(chains).

Class	Proteins
α (26)	1G6IA(545) 1GKMA(509) 1JFBA(404) 1M5NS(485) 4CTSA(437) 1CPT(428) 1GLM(470) 1MHLC(466) 1MMOD(512) 1OXA(403) 1PHB(414) 1ROM(403) 1VNC(609) 1XSM(390) 2BCT(516) 2BMHA(455) 1B89(449) 1IAL(453) 1HO8(480) 1B8F(509) 1DL2(511) 5EAS(548) 1BKE(581) 1BJ5(585) 1AVC(673) 1ST6(1069)
β (25)	1C9UB(454) 1F8EA(388) 4AAHA(571) 1EUR(365) 1IDK(359) 1PMI(440) 1TSP(559) 2CAS(548) 2SIL(381) 2TBVA(387) 3NN9(388) 4BCL(366) 1A65(504) 1A6C(513) 1B9S(390) 1DAB(539) 1EUT(605) 1FNF(368) 1C8F(548) 1DBG(506) 1DZL(505) 1KCW(1046) 1P2Z(968) 1P30(952) 1W0O(781)
$\alpha+\beta$ (22)	1A2N(419) 1B65A(375) 1GK9B(557) 1R52B(382) 2JDXA(385) 1EPS(427) 1GCB(454) 1LML(478) 1PNKB(557) 1UAE(419) 1DMT(696) 1EWF(456) 1OIE(532) 1W1O(534) 1USH(550) 1AOP(497) 1KA2(499) 1V0R(506) 5JDW(386) 1SIJ(907) 5LDHA(333) 1JMUG(365)
α/β (27)	1LK9A(448) 1LKXD(697) 1LLFA(534) 1M1NA(491) 1PMOC(466) 1UZBA(516) 1BYB(495) 1AG8A(499) 1CBG(490) 1GPB(842) 1MIOA(533) 1TPLA(456) 2DKB(433) 2OLBA(517) 2TS1(419) 1A8I(841) 1AOV(686) 1BFD(528) 1CRL(686) 1AIV(686) 1AK5(503) 1AKN(579) 1AX9(537) 1AXR(842) 1B1X(689) 1FA9(846) 1EJJ(511)

These 26 parameters are then used to construct a 26-dimensional ($26D$) space. In this parameter space, one point represents a protein. We want to find out whether proteins from the four structural classes can be separated in this space.

As in Refs. [9] and Yang *et al.* [10], Fisher's discriminant algorithm is used to find a classifier in the parameter space for a training set. The given training set $H = \{x_1, x_2, \cdots, x_n\}$ is partitioned into $n_1 \leq n$ training vectors in a subset H_1 and $n_2 \leq n$ training vectors in a subset H_2, where $n_1 + n_2 = n$ and each vector x_i is a point in the $26D$ parameter space. Then $H = H_1 \cup H_2$.

We use the whole data set as the training set because the selected protein data set is small. The discriminant accuracies for re-substitution analysis are defined as $P_{H_1} = \frac{n_{cH_1}}{n_1}$ and $P_{H_2} = \frac{n_{cH_2}}{n_2}$, where n_{cH_1} and n_{cH_2} denote the number of correctly discriminating H_1 elements and the number of correctly discriminating H_2 elements in the training set, respectively. The result obtained in this way can be used to check the self-consistency of a predictor, especially for its

Table 2. The 26 parameters from the calculation. x-ts and y-ts represent x time series and y time series, respectively; $C_{\max}$ is the maximum value of C_q with q ranging from -15 to 15 and q_0 is the value of q corresponding to $C_{\max}$; $\Delta f = f(\alpha_{\max}) - f(\alpha_{\min})$, $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$.

Order	Data	Parameter	Order	Data	Parameter
1	x-ts	D_{-1}	14	y-ts	Δf
2	x-ts	D_1	15	y-ts	D_{-1}
3	x-ts	D_2	16	y-ts	D_1
4	x-ts	C_{-1}	17	y-ts	D_2
5	x-ts	C_1	18	y-ts	C_{-1}
6	x-ts	C_0	19	y-ts	C_1
7	x-ts	$C_{\max}$	20	y-ts	C_0
8	x-ts	q_0	21	y-ts	$C_{\max}$
9	x-ts	$\alpha_{\max}$	22	y-ts	q_0
10	x-ts	$\alpha_{\min}$	23	y-ts	$\alpha_{\max}$
11	x-ts	$\Delta\alpha$	24	y-ts	$\alpha_{\min}$
12	x-ts	$f(\alpha_{\max})$	25	y-ts	$\Delta\alpha$
13	x-ts	$f(\alpha_{\min})$	26	y-ts	$f(\alpha_{\max})$

algorithm part. A predictor certainly cannot be deemed to be a good one if its self-consistency rate is poor [17].

We propose the following procedure to classify protein structures in the $26D$ space to structural classes, which consists of three steps: **Step 1:** classify the proteins of the α class from the other proteins in the $\{\beta, \alpha+\beta, \alpha/\beta\}$ classes; **Step 2:** classify the β class proteins from the other proteins in the $\{\alpha+\beta, \alpha/\beta\}$ classes; **Step 3:** classify the $\alpha+\beta$ class proteins from the proteins of the α/β class.

The discriminant accuracies P_{H_1} in step 1 to 3 are **84.62%, 84.00%** and **81.48%** respectively, while the discriminant accuracies P_{H_2} in step 1 to 3 are **83.78%, 87.76%** and **86.36%** respectively. The average accuracy is 84.67%, which is relatively satisfactory.

In order to compare with the methods in Yu *et al.* [9] and Yang *et al.* [10], we calculate the accuracies for the 46 proteins in Yang *et al.* [10] with the method here. Using the present method, the discriminant accuracies P_{H_1} in step 1 to 3 are **97.14%, 100.00%** and **100.00%** respectively, while the discriminant accuracies P_{H_2} in step 1 to 3 are all **100.00%**. If we use the method in Yang *et al.* [10], the discriminant accuracies P_{H_1} in step 1 to 3 are 100.00%, 92.86% and 91.67% respectively, while the discriminant accuracies P_{H_2} in step 1 to 3 are 84.21%, 86.96% and 83.33% respectively. We can see the method proposed here really improves the performance of protein structure classification.

It is importnat to discuss the effect of arrangement of the amino acids on the vertices of the regular 20-polygon. We arranged the amino acids on the vertices of the 20-polygon randomly and then repeat the above calculations. We tested the classification result of proteins to see whether they affect the accuracies. For example, we set the amino acids E, Y, Q, A, K, N, V, H, P, C, S, F, T, D, R, G, W, M, L and I on the first vertex (vertex 0), the second vertex (vertex 1), $\cdots$, the nineteen vertex (vertex 18), and the twentieth vertex (vertex 19), respectively. In this randomly selected order case, the discriminant accuracies P_{H_1} in step 1 to 3 are 84.62%, 68.00% and 96.30% respectively, while the discriminant accuracies P_{H_2} in step 1 to 3 are 77.03%, 85.71% and 90.91% respectively. This suggests that the accuracies in different steps are in fact affected by the arrangement. However, the overall accuracies with different assignments are similar to each other.

It should be stressed that the multifractal analysis here is applied to the measure constructed from time series. As mentioned before, such measure is sensitive to the ordering of the time series. Different arrangements of amino acids on the polygon will lead to different ordering of the time series for a given protein sequence. Such differences will of course lead to different values of the 26 calculated parameters. This can be seen more clearly in Figure 3. As a result, it is expected that the results in each classification step are affected by the arrangement methods.

Yu *et al* [13] proposed the CGR of protein sequences from complete genomes based on the detailed HP model, and the measure they defined was different from that in this paper. As the protein sequences from complete genomes were rather long (more than 10^5) to get enough points in the CGR, the measure could be defined by the number of points lying in the subsets of the CGR [13]. In such a way, the arrangement manner of the amino acid residues has no influence on the measure and furthermore no effect upon the values of D_q, C_q, $\alpha(q)$ and $f(\alpha)$. However, the longest protein sequence considered in this paper is less than 1100, the definition of measure in such a way may be inappropriate as the measure of most subsets of the CGR will be zero when the subsets are too small. As a result, some biological meaning may be lost and we consider using the definition of measure in Yu *et al.* [15] and Yang *et al.* [10].

Basu *et al* [12] proposed another kind of CGR of proteins using a regular 12-polygon by grouping together similar amino acid residues. The 12 groups of amino acids are then {A, G}, {C}, {D, E}, {F, Y}, {H}, {I, L, V, M}, {N}, {P}, {Q}, {R, K}, {S, T}, and {W}. Firstly, we also use an alphabetical order arrangement in the CGR and the discriminant accuracies P_{H_1} in step 1 to 3 are 76.92%, 80.00% and 88.89% respectively, while the discriminant accuracies P_{H_2} in step 1 to 3 are 85.14%, 85.71% and 100.00% respectively. If a certain group contains more that one kind of amino acids, we use the first letter in the group for its representation. For example, for the group {I, L, V, M}, the representing amino acid residue is I. Then we consider a random arrangement of these groups on the vertices of the 12-polygon [12]. In a randomly selected order case, the discriminant accuracies P_{H_1} in step 1 to 3 are 80.77%,

76.00% and 96.30% respectively, while the discriminant accuracies P_{H_2} in step 1 to 3 are 77.03%, 93.88% and 90.91% respectively. The discussion here is analogous to that in the 20-polygon CGR. It can be seen that even though there are some differences between different classification steps, the overall accuracies don't have too much differences. Furthermore, the overall accuracy with the 12-polygon CGR is similar to that with the 20-polygon CGR.

4. Conclusions

Identification of protein structural classes is important in the prediction of the $3D$ structures. Chaos game representation of proteins is a useful way to analyze proteins as it provides a visualization of protein sequences. Most important of all, the protein sequence can be reconstructed uniquely from the CGR. In order to analyze CGR of proteins more conveniently, we decompose the CGR of proteins into two time series.

Multifractal analysis is a useful tool in many different fields. Multifractal analysis of the measure for the constructed time series of proteins provides useful information to classify protein structures. For each protein, a total of 26 parameters is calculated through multifractal analysis. In order to classify protein structures, the 26 parameters are used to construct a $26D$ parameter space. Then each protein is represented by a point in this space. With Fisher's linear discriminant algorithm, we can classify protein structures to structural classes with an average accuracy of 84.67% for the 100 large proteins. Compared with the results for the 46 large proteins in Yang *et al.* [10], it clearly indicates that the methods proposed here improve the results reported in Yu *et al.* [9] and Yang *et al.* [10].

Acknowledgement

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

References

[1] C. Anfinsen, Principles that govern the folding of protein chains. *Science*, 181: 223-230, 1973.

[2] H. Li, R. Helling, C. Tang and N. Wingreen, Emergence of Preferred Structures in a Simple Model of Protein Folding*Science*, 273: 666-669, 1996.

[3] H. Li, C. Tang and N.S. Wingreen, Are protein folds atypical? *Pro. Natl. Acad. Sci. USA*, 95: 4987-4990, 1998.

[4] B. Wang and Z.G. Yu, A way to characterize the compact structures of lattice protein model, *J. Chem. Phys.*, 112: 6084-6088, 2000.

[5] J.Y. Yang, Z.G. Yu and V. Anh, Some quantities including designability to characterize protein lattice models and their correlation relationship, *J. Chem. Phys.*, 126: 195101, 2007.

[6] M. Levitt and C. Chothia, Structural Patterns in Globular Proteins. *Nature*, 261: 552-558, 1976.

[7] I. Bahar, A.R. Atilgan, R.L. Jernigan and B. Erman, Understanding the recognition of protein structural classes by amino acid composition. *Proteins*, 29: 172-185, 1997.

[8] L.A. Kurgan and L. Homaeian, Prediction of structural classes for protein sequences and domains–Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. *Pattern Recognition*, 39: 2323-2343, 2006.

[9] Z.G. Yu , V. Anh, K.S. Lau and L.Q. Zhou, Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins. *Phys. Rev. E*, 73: 031920, 2006.

[10] J.Y. Yang, Z.G. Yu and V. Anh, Clustering structures of large proteins using multifractal analyses based on a 6-letter model and hydrophobicity scale of amino acids. *Chaos, Solitons and Fractals*, (in press), doi:10.1016/j.chaos.2007.08.014, 2008.

[11] A. Fiser, G.E. Tusnády and I. Simon, Chaos game represenation of protein structures. *J. Mol. Graphics*, 12: 302-304, 1994.

[12] S. Basu, A. Pan, C. Dutta and J. Das, Chaos game representation of proteins. *J. Mol. Graphics*, 15: 279-289, 1997.

[13] Z.G. Yu, V. Anh and K.S. Lau, Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses, *J. Theor. Biol.*, 226: 341-348, 2004.

[14] H.J. Jeffrey, Chaos game representation of gene structure. *Nucleic Acids Res.* , 18: 2163-2170, 1990.

[15] Z.G. Yu , V. Anh and K.S. Lau, Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome. *Physica A*, 301: 351-361, 2001.

[16] Z.G. Yu , V. Anh and K.S. Lau, Fractal analysis of measure representation of large proteins based on the detailed HP model. *Physica A*, 337: 171-184, 2004.

[17] K.C. Chou and H.B. Shen, Recent progress in protein subcellular location prediction. *Analytical Biochemistry*, 370: 1-16, 2007.