Vol 18 No 10, October 2009
1674-1056/2009/18(10)/4571–09

Chinese Physics B

© 2009 Chin. Phys. Soc.
and IOP Publishing Ltd

Chaos game representation walk model for the protein sequences*

Gao Jie(高 洁)[a)b)†], Jiang Li-Li(蒋丽丽)[a)], and Xu Zhen-Yuan(徐振源)[a)]

[a)] *School of Science, Jiangnan University, Wuxi 214122, China*

[b)] *School of Information Technology, Jiangnan University, Wuxi 214122, China*

(Received 9 December 2008; revised manuscript received 9 April 2009)

A new chaos game representation of protein sequences based on the detailed hydrophobic–hydrophilic (HP) model has been proposed by Yu *et al* (Physica A **337** (2004) 171). A CGR-walk model is proposed based on the new CGR coordinates for the protein sequences from complete genomes in the present paper. The new CGR coordinates based on the detailed HP model are converted into a time series, and a long-memory ARFIMA(p, d, q) model is introduced into the protein sequence analysis. This model is applied to simulating real CGR-walk sequence data of twelve protein sequences. Remarkably long-range correlations are uncovered in the data and the results obtained from these models are reasonably consistent with those available from the ARFIMA(p, d, q) model.

Keywords: chaos game representation CGR-walk model, protein sequence, long-memory ARFIMA (p, d, q) model, autocorrelation function

PACC: 8710, 0250

1. Introduction

Modelling the three-dimensional structure of proteins is a complex physical, chemical and mathematical problem of prime importance in molecular biology, medicine, and pharmacology.[1,2] It is believed that the dynamical folding process and stable structure, or native conformation, of a protein is determined by its primary structure, namely its amino acid sequence.[3] Twenty different kinds of amino acids have been found in proteins. The arrangement order of amino acid in the protein polypeptide chain high level structures (secondary and space structures) from the amino acid sequence (protein sequence) is a challenging problem in science. Amino acid sequence is closely related to biological function. Its change often leads to the change of biological function of protein and amino acids of homologous protein at many positions are the same, while others are greatly different. The closer the genetic relationship is, the smaller the difference in amino acid composition between them will be. Over the past thirty years, scientists have been studying the characteristics of protein sequence by modeling it. A well-known model of protein sequences is the hydrophobic–hydrophilic (HP) model.[4,5] In this model, 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and hydrophilic (P) (or polar). In recent years the HP model has been extensively studied,[2,6,7] but the HP model may be simplistic and lacking of sufficient information about the heterogeneity and complexity of the natural set of residues.[8] According to Brown,[9] one can divide the polar class into three subclasses in the HP model: 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. This model is called the detailed HP model,[10] and it provides more information than the HP model.

Although statistical analysis performed directly on DNA sequences has yielded some successful results, there has been an indication that this method is neither powerful enough to amplify the difference between a DNA sequence and a random sequence nor to distinguish DNA sequences among themselves.[11] Thus more useful global and visual methods are needed. Jeffrey (1990) proposed the chaos game representation (CGR) for DNA sequences and found fractal patterns in these representations.[12] The correlation properties of coding and noncoding DNA sequences

*Project supported by the National Natural Science Foundation of China (Grant No 60575038), the Natural Science Foundation of Jiangnan University, China (Grant No 20070365) and the Program for Innovative Research Team of Jiangnan University, China.
†Corresponding author. E-mail: ezhun6669@sina.com

were studied by Peng *et al* (1992) in their fractal landscape or DNA walk model.[13] The long-range correlation (long-range dependence) in DNA sequences was first proposed by Peng *et al*. It is a significant contribution to the study in many fields. Since then, the idea of long-range correlation in DNA sequences has been receiving considerable attention of scientists. Hao *et al*[11] proposed a visualisation method based on counting and coarse-graining the frequency of appearance of substrings with a given length. Yu *et al* viewed DNA sequence as a time series, and used it to reveal more information.[14]

The CGR of DNA sequences has been extended to the representation of protein sequences and protein structures.[15,16] Yu *et al* (2004) proposed a CGR of protein sequences based on the detailed HP model and then performed multifractal analysis on the representation.[10] Through the CGR of the simulated protein sequence, they proposed to use the correlation distance based on the difference between these two measures to discuss the phylogenetic relationships of bacteria.

Gao *et al* proposed a CGR-walk model based on CGR coordinates for the DNA sequences.[17] The CGR coordinates were converted into a time series model, and a long-memory ARFIMA(p, d, q) model was introduced to DNA sequence analysis. Remarkably long-range correlations were uncovered in the data.

In the present paper, we construct a similar CGR-walk model based on the detailed HP model for the linked protein sequences, and a long-memory ARFIMA(p, d, q) model is introduced to the linked protein sequence analysis. This model is applied to simulating real CGR-walk sequence data of twelve linked protein sequences from twelve complete genomes of bacteria. We also uncover in the data remarkably long-range correlations and fit these models reasonably by ARFIMA(p, d, q) models.

2. ARFIMA model

The dependences of current or future values on historical values in stochastic process can be divided into short-range dependence (short memory) and long-range dependence (long memory) according to the length. Lopes *et al*[18] first introduced a long memory ARFIMA(0, d, 0) model into the simulation of random walk in one dimension from DNA sequences.

In this section, we present the ARFIMA(p, d, q) model and some related theoretical results. Models that include fractional differentiation d in the interval $(0, 0.5)$ are able to represent any time series that shows persistence, also known as long memory property.[19] Initial studies of time series with long memory characteristics were given by Hurst.[20] ARFIMA processes first appeared in Refs.[21] and [22] and are a generalisation of the ARMA and ARIMA models. The author of Ref.[23] was the pioneer in the application of long memory in hydrological time series. Persistence or long memory property has been observed in time series from different fields such as meteorology, astronomy, hydrology, and economy. One can characterise the persistence by the following two different forms.

(i) In the time domain, the autocorrelation function $\rho_x(.)$ decays hyperbolically to zero, that is, $\rho_x(k) \backsim k^{2d-1}$, when $k \to \infty$.

(ii) In the frequency domain, the spectral density function $f_x(.)$ is unbounded when the frequency is near zero, that is, $f_x(w) \backsim w^{-2d}$, when $w \to 0$.

One of the models in which the persistence can be described is the so-called ARFIMA(p, d, q) process.

Definition 1 Let $\{\varepsilon_t\}_{t\in Z}$ be a white noise process with zero mean and variance $\sigma_\varepsilon^2 > 0$, and B the backward-shift operator, i.e., $B^k(X_t) = X_{t-k}$. If $\{X_t\}_{t\in Z}$ is a linear process satisfying

$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t, \quad t \in Z, \tag{1}$$

where $d \in (-0.5, 0.5)$, $\Phi(\cdot)$, and $\Theta(\cdot)$ are polynomials of degrees p and q, respectively, given by

$$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,$$

$$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q,$$

where ϕ_i, $1 \le i \le p$, and θ_j, $1 \le j \le q$, are real constants, then $\{X_t\}_{t\in Z}$ is called general fractional difference ARFIMA(p, d, q) process, where d is the degree or fractional differentiation parameter.

If $d \in (-0.5,\ 0.5)$, then $\{X_t\}_{t\in Z}$ is a stationary, and an invertible process. The most important characteristic of an ARFIMA(p, d, q) process is the property of long dependence, when $d \in (0,\ 0.5)$, short dependence, when $d = 0$, and intermediate dependence, when $d \in (-0.5,\ 0)$.

3. CGR-walk model of protein sequences based on the detailed HP model

Each coding sequence in the complete genome of an organism can be translated into a protein sequence by using the genetic code.[9] The secondary and the space structures of a protein are determined by its amino acid sequence.[24] The protein structure prediction methods can be divided into three times: first, they are based on statistical analysis of single amino acid residue; second, they were based on statistical analysis of amino acid fragment. The prediction accuracies of α-helix, β-sheet and coil are all smaller than 70 percent, and the prediction accuracy for β-sheet is even smaller than 28–48 percent for which the main reason is that only the local information is used.[25] So in the prediction methods of the third times the multiple protein sequence is taken into consideration and the long-range dependence and evolution information of protein sequences are used.[26] In order to discover the correlation between amino acids over long distances along the protein chain, we link all translated protein sequences from a complete genome to form a long protein sequence according to the order of the coding sequence in the complete genome. As a result, we obtain a linked protein sequence for each organism. In the present paper, we consider this kind of linked protein sequences for the organisms and view them as symbolic sequences. We can classify and model the symbolic sequences according to their whole sequences characteristics.

The protein sequence is composed of 20 different kinds of amino acids, namely Alanine (A), Arginine (R), Asparagine (N), Aspartic acid (D), Cysteine (C), Glutamic acid (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y) and Valine (V). Some amino acids are non-polar (hydrophobic) because their side chains are made up mostly of hydrocarbon chains; some are polar but uncharged; some are polar and normally have a positive or a negative charge in solution. In the detailed HP model,[10] they can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. The eight residues A, I, L, M, F, P, W and V are designated to the non-polar class; the two residues D and E designated to the negative polar class; the seven residues N, C, Q, G, S, T and Y designated to the uncharged polar class; and the remaining three residues R, H and K designated to the positive polar class.

For a given protein sequence $s = s_1 s_2 \dots s_n$ with length n, where s_i is one of the 20 kinds of amino acids for $i = 1, \dots, n$, we define

$$c_i = \begin{cases} \text{A0}, & \text{if } s_i \text{ is non-polar,} \\ \text{A1}, & \text{if } s_i \text{ is negative polar,} \\ \text{A2}, & \text{if } s_i \text{ is uncharged polar,} \\ \text{A3}, & \text{if } s_i \text{ is positive polar,} \end{cases}$$

and then obtain a sequence $X(s) = c_1 c_2 \dots c_n$, where c_i is a letter of the alphabet {A0, A1, A2, A3}. Next, we define the CGR for a sequence $X(s)$, similar to that of DNA sequences, in a square [0, 1]×[0, 1], where the four vertices correspond to the four letters A0, A1, A2 and A3, A0=(0, 0), A1=(0, 1), A2=(1, 1), A3=(1, 0). The first point of the plot is placed half way between the centre of the square and the vertex corresponding to the first letter of the sequence $X(s)$; the i-th point of the plot is then placed half way between the $(i-1)$-th point and the vertex corresponding to the i-th letter

$$\begin{aligned} &\text{CGR}_i = \text{CGR}_{i-1} - 0.5 \cdot (\text{CGR}_{i-1} - c_i), \\ &i = 1, \cdots, n, \quad \text{CGR}_0 = (0.5,\ 0.5). \end{aligned} \tag{2}$$

For example, the CGR of the first ten amino acids of the linked protein sequence of Sulfolobus solfataricus P2 is given in Fig.1.

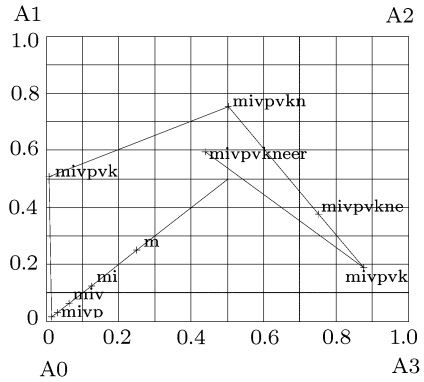


Fig.1. CGR of the first 10 amino acids of the linked protein sequence of Sulfolobus solfataricus P2: mivpvkneer.

For a linked protein sequence, we construct a CGR-walk model $\{t_k = y_k/x_k : k = 1, 2, \dots, n\}$, where y_k is the y-coordinate of CGR_k, x_k is the x-coordinate of CGR_k.

4. Analysis and discussion

4.1. Data analysis for the linked protein sequence of Sulfolobus solfataricus P2

We analyse a CGR-walk model for the linked protein sequence of Sulfolobus solfataricus P2 (Ssol, one of Archae Crenarchaeota). The linked protein sequence has more than one hundred thousand amino acids. Figure 2(a) displays a CGR-walk sequence of the linked protein sequence of Sulfolobus solfataricus P2 (positions 1615–2831) with a total of 1217 observations, i.e. $n = 1217$. Due to increasing variability and trends in the data, the first difference between logarithms of CGR-walk data is considered. The resulting series is plotted in Fig.2(b). The differenced series seems to be stationary, even though a small heteroscedasticity is observed.

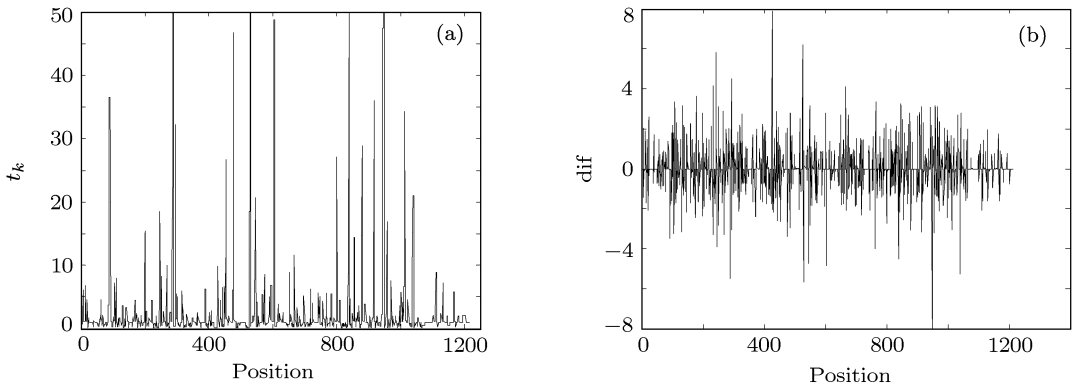


Fig.2. (a) CGR-walk sequence of the linked protein sequence of Sulfolobus solfataricus P2, with 1217 amino acids in total; (b) first differenced log data.

The sample autocorrelation function (ACF) of the CGR-walk data is shown in Fig.3(a), and the partial autocorrelation function (PACF) of the CGR-walk data is shown in Fig.3(b). The ACF of the differenced log data is shown in Fig.4(a), and the PACF of the differenced log data is shown in Fig.4(b). The ACF of the differenced log data decays rapidly, while the PACF decays slowly. This seems to indicate the presence of long-memory component in the initial data.

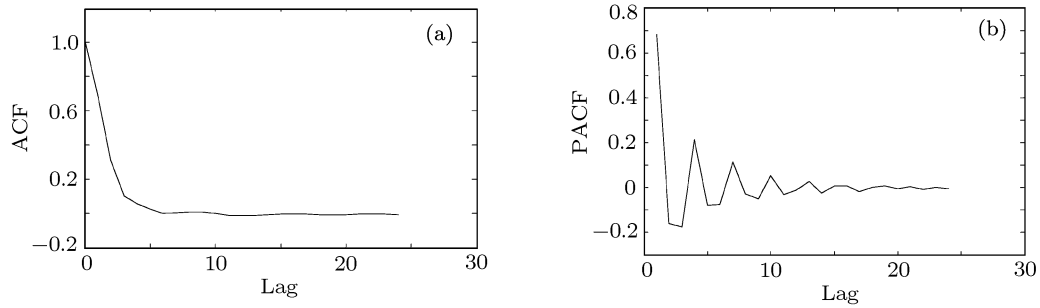


Fig.3. (a) Sample ACF of the CGR-walk data; and (b) sample PACF of the CGR-walk data.

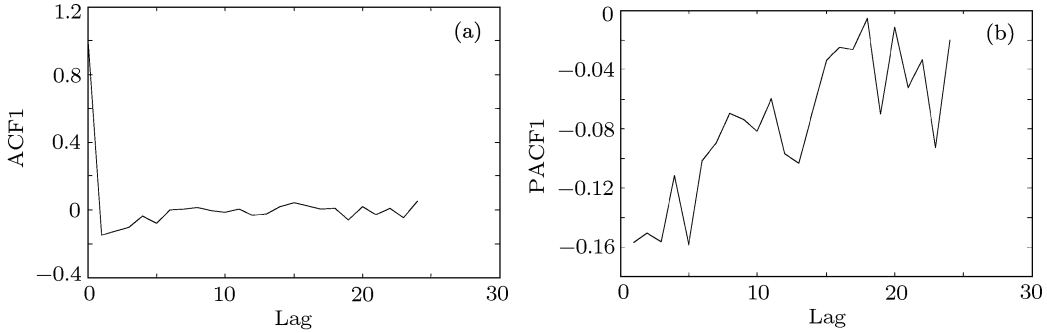


Fig.4. (a) Sample ACF of the differenced log data; and (b) sample PACF of the differenced log data.

The variance plot in Fig.5 is a useful tool to detect the presence of long-memory behaviour in the data. As discussed in Ref.[19] by Beran, for a long-range dependent time series $\{x_t\}$, the variance of its mean values $\bar{x}_k$ satisfies

$$\mathrm{Var}(\bar{x}_k) \sim k^{2d-1},$$

$$\frac{\log[\mathrm{Var}(\bar{x}_k)]}{\log(k)} \sim 2d-1.$$

Therefore by plotting the curves for $\log[\mathrm{Var}(\bar{x}_k)]$ versus $\log(k)$ for different values of k, a straight line with slope $2d-1$ should be found. Since $d=0$ for a short-memory process, the slope would be –1. Plots with slopes greater than –1 would indicate the presence of long-memory behaviour of $d \in (0,\ 0.5)$. For the CGR-walk data, the estimated slope is –0.6171 through the least squares estimation, suggesting a crude estimate of the long-memory parameter $\hat{d} = 0.1915$.

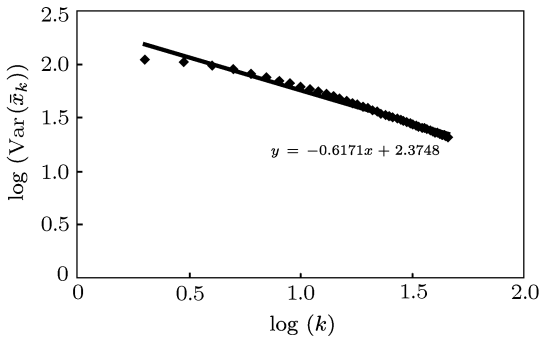


Fig.5. Variance plot, where heavy line is fitted straight line with slope –0.6171.

According to the above reasons, the CGR-walk sequence shows long memory and the goal here is to use these characteristics to construct an adequate model for the CGR-walk sequence. In order to do this, we consider a popular class of models for time series with long-memory behaviour, that is ARFIMA(p, d, q) models, where the fractional parameter d measures the long memory property when $d \in (0, 0.5)$.

Accordingly, a class of ARFIMA(p, d, q) models, with p and q less than or equal to 5 is considered. Based on the Akaike's Information Criterion (AIC),[27,28] ARFIMA(0, 0.1915, 5) model is selected.

4.2. Model test and parameter estimate for the linked protein sequence of Sulfolobus solfataricus P2

To test the selected model, we choose suitable test statistics: the modified portmanteau test (LB test)[29,30]

$$\mathrm{LB} = n(n+2)\sum_{k=1}^{M}\frac{r_k^2}{n-k} \overset{\text{appr.}}{\sim} \chi^2(M-p-q-1),$$

where r_k is the sample autocorrelation at lag k, n is the sample size, and M is a prespecified integer depending on n. In the present context, p-values of the LB test statistics at every lag k are significantly larger than 0.1 (see Table 1). This indicates that the residuals of the fitted model seem to be white noise, and it is reasonable to accept the ARFIMA(0, 0.1915, 5) model.

Table 1. Autocorrelation check of residuals for protein sequence of Sulfolobus solfataricus P2.

to lag	χ^2	$\Pr > \chi^2$
6	0.02	0.8840
12	1.58	0.9793
18	5.86	0.9513
24	14.37	0.7615
30	18.53	0.8190
36	20.27	0.9298
42	23.01	0.9651
48	34.97	0.8031

Table 2 gives the parameter estimates of the selected ARFIMA(0, 0.1915, 5) model. The p-values of the T test statistics for five parameters are significantly smaller than 0.01 (see Table 2). This indicates that the ARFIMA(0, 0.1915, 5) model can fit the CGR-walk model of the linked protein sequence of Sulfolobus solfataricus P2 effectively.

Table 2. Maximum likelihood estimation for protein sequence of Sulfolobus solfataricus P2.

parameter	estimate	standard error	t value	$\Pr > \lvert t \rvert$	lag
θ_1	0.36152	0.0287	12.60	< 0.0001	1
θ_2	0.2662	0.03034	8.77	< 0.0001	2
θ_3	0.18733	0.03079	6.08	< 0.0001	3
θ_4	0.08374	0.03034	2.76	0.0058	4
θ_5	0.09758	0.02864	3.41	0.0007	5

4.3. Data analysis for other linked protein sequences

Currently, there are more than 50 complete genomes of Archaea and Eubacteria available in public databases (for example in Genbank at the web site ftp:/ ncbi.nlm.nih.gov/genbank/genomes/). These include eight Archae Euryarchaeota, two Archae Crenarchaeota, three Gram-positive Eubacteria (high G+C) and 12 Gram-positive Eubacteria (low G+C). The others are Gram-negative Eubacteria, which consist of two hyperthermophilic bacteria, four Chlamydia, two Cyanobacterium, two Spirochaete and 16 Proteobacteria.

Except Sulfolobus solfataricus P2 (Proteobacteria) analysed in detail in the last sections, we download eleven long protein sequences from the complete genomes of the above bacteria. Two Archae Euryarchaeota: Archaeoglobus fulgidus DSM4304 (Aful), Methanocaldococcus jannaschii DSM2661 (Mjan); Two Gram-positive Eubacteria (high G+C): Mycobacterium tuberculosis H37Rv (MtubH), Mycobacterium leprae TN (Mlep); One Gram-positive Eubacteria (low G+C): Mycoplasma pneumoniae M129 (Mpne); Six Gram-negative Eubacteria: Aquifex aeolicus (Aqua) VF5, Chlamydia pneumoniae CWL029 (Cpne), Synechocystis sp.PCC6803 (Syne), Borrelia burgdorferi B31 (Bbur), Pasteurella multocida PM70 (Pmul) and Buchnera sp. APS (Buch).

Table 3. Data information and selected ARFIMA models for all 11 protein sequences selected.

name[a]	positions	sample size	selected model
1	1213–2467	1255	ARFIMA(0, 0.1373, 4)
2	1–1220	1220	ARFIMA(0, 0.2237, 4)
3	1–1194	1194	ARFIMA(1, 0.0535, 2)
4	1196–2393	1198	ARFIMA(0, 0.083, 4)
5	1193–2396	1204	ARFIMA(0, 0.0144, 4)
6	398–1575	1178	ARFIMA(1, 0.1722, 1)
7	1218–2443	1226	ARFIMA(1, 0.0189, 1)
8	820–2021	1202	ARFIMA(0, 0.1344, 5)
9	2001–3190	1190	ARFIMA(1, 0.1759, 1)
10	1–1036	1036	ARFIMA(1, 0.1782, 1)
11	1–1046	1046	ARFIMA(0, 0.1031, 4)

[a] 1. Archaeoglobus fulgidus DSM4304 (Aful); 2. Methanocaldococcus jannaschii DSM2661 (Mjan); 3. Mycobacterium tuberculosis H37Rv (MtubH); 4. Mycobacterium leprae TN (Mlep); 5. Mycoplasma pneumoniae M129 (Mpne); 6. Aquifex aeolicus (Aqua) VF5; 7. Chlamydia pneumoniae CWL029 (Cpne); 8. Synechocystis sp.PCC6803 (Syne); 9. Borrelia burgdorferi B31 (Bbur); 10. Pasteurella multocida PM70 (Pmul); and 11. Buchnera sp. APS (Buch).

Data information and selected ARFIMA(p, d, q) models are all listed in Table 3 for the aboved eleven protein sequences. The values of long-memory parameter d are all lying in (0, 0.5). The p-values of the LB test statistics at every lag k for every selected model are almost all significantly larger than 0.1 (see Table 4). Also, the p-values of the T test statistics for parameters of every selected model are significantly smaller than 0.01 except 0.0673 (< 0.1) (see Table 5). All of these indicate that the ARFIMA(p, d, q) models can fit the CGR-walk models of different protein sequences well.

Table 4. Autocorrelation check of residuals for all 11 protein sequences selected. The name is the same as that in Table 3.

name	to lag	$\Pr > \chi^2$
1	6	0.0640
	12	0.5769
	18	0.5092
	24	0.6968
	30	0.6637
	36	0.7444
	42	0.7736
	48	0.6298
2	6	0.6783
	12	0.5729
	18	0.7461
	24	0.9060
	30	0.7484
	36	0.7561
	42	0.6422
	48	0.6897
3	6	0.7916
	12	0.1717
	18	0.4179
	24	0.4611
	30	0.6640
	36	0.7308
	42	0.7791
	48	0.7428
4	6	0.0300
	12	0.0841
	18	0.0676
	24	0.1781
	30	0.1884
	36	0.4729
	42	0.3186
	48	0.4325

Table 4. (*Continued*).

name[a]	to lag	$\Pr > \chi^2$
5	6	0.3409
	12	0.1321
	18	0.3800
	24	0.4412
	30	0.2443
	36	0.1144
	42	0.1040
	48	0.1059
6	6	0.6131
	12	0.4791
	18	0.4802
	24	0.4794
	30	0.5995
	36	0.3215
	42	0.0819
	48	0.0881
7	6	0.2880
	12	0.2785
	18	0.0239
	24	0.0935
	30	0.1910
	36	0.3472
	42	0.3426
	48	0.0372
8	6	0.1962
	12	0.7052
	18	0.9582
	24	0.9129
	30	0.9141
	36	0.9452
	42	0.9700
	48	0.9370
9	6	0.9720
	12	0.7853
	18	0.6610
	24	0.6410
	30	0.6347
	36	0.7940
	42	0.7823
	48	0.8412
10	6	0.0900
	12	0.1416
	18	0.4827
	24	0.8167
	30	0.7945
	36	0.6524
	42	0.8723
	48	0.1054
11	6	0.0496
	12	0.0372
	18	0.0820
	24	0.3006
	30	0.4918
	36	0.6736
	42	0.7349
	48	0.8175

Table 5. Maximum likelihood estimation for all 11 protein sequences selected.

	parameter	estimate	standard error	t value	$\Pr > \lvert t \rvert$
1	θ_1	0.36247	0.02816	12.87	< 0.0001
	θ_2	0.28111	0.02905	9.68	< 0.0001
	θ_3	0.26266	0.02907	9.03	< 0.0001
	θ_4	0.08152	0.02827	2.88	0.0039
2	θ_1	0.45105	0.02851	15.82	< 0.0001
	θ_2	0.23479	0.03078	7.63	< 0.0001
	θ_3	0.20055	0.03082	6.51	< 0.0001
	θ_4	0.10956	0.02857	3.83	0.0001
3	Φ_1	0.60376	0.03602	16.76	< 0.0001
	θ_1	0.91843	0.04686	19.60	< 0.0001
	θ_2	0.08100	0.04427	1.83	0.0673
4	θ_1	0.31706	0.02820	11.24	< 0.0001
	θ_2	0.22493	0.02909	7.73	< 0.0001
	θ_3	0.20558	0.02906	7.08	< 0.0001
	θ_4	0.22300	0.02823	7.90	< 0.0001
5	θ_1	0.42910	0.02867	14.96	< 0.0001
	θ_2	0.26417	0.03079	8.58	< 0.0001
	θ_3	0.17751	0.03083	5.76	< 0.0001
	θ_4	0.12572	0.02874	4.37	< 0.0001
6	Φ_1	0.65189	0.02259	28.86	< 0.0001
	θ_1	0.99627	0.00340	293.37	< 0.0001
7	Φ_1	0.61421	0.02288	26.85	< 0.0001
	θ_1	0.99625	0.00278	358.77	< 0.0001
8	θ_1	0.36064	0.02880	12.52	< 0.0001
	θ_2	0.22808	0.03053	7.47	< 0.0001
	θ_3	0.23487	0.03046	7.71	< 0.0001
	θ_4	0.08257	0.03054	2.70	0.0069
	θ_5	0.08131	0.02882	2.82	0.0048
9	Φ_1	0.62932	0.02384	26.40	< 0.0001
	θ_1	0.98864	0.00538	183.68	< 0.0001
10	Φ_1	0.60368	0.02540	23.77	< 0.0001
	θ_1	0.99694	0.00394	253.11	< 0.0001
11	θ_1	0.38969	0.03081	12.65	< 0.0001
	θ_2	0.20364	0.03213	6.34	< 0.0001
	θ_3	0.23965	0.03211	7.46	< 0.0001
	θ_4	0.16445	0.03064	5.37	< 0.0001

5. Discussion and conclusion

The chaos game representation of sequences is a method to ordinate the entire domain of possibilities in a continuous two-dimensional space. The CGR transformation makes a certain sequence become an entire new set of statistical analysis tools. Therefore, CGR is a formalism that bridges between sequences of discrete units and numeric coordinates in a continuous space.

The CGR based on the detailed HP model of protein sequences provides a simple and powerful visualisation method to distinguish protein sequences themselves in more detail.

Although some research has been carried out by taking into consideration the long-range correlations in protein sequences, the methods are rough and the results are not satisfactory. In this paper, we convert the CGR coordinates into a time series model (CGR-walk model) and introduce a long-memory ARFIMA (p,d,q) model into protein sequence analysis. Consequently, basic statistic measures and time series analysis techniques can now be applied to the CGR-walk model for protein sequences.

We first analyse the real protein sequence data of Sulfolobus solfataricus P2. From Figs.2–5, we detect the presence of long-memory behaviour in the data. Based on AIC, ARFIMA(0, 0.1915, 5) model is selected to fit the CGR-walk sequence data of Sulfolobus solfataricus P2. From Table 1, one can see that the residuals of the fitted model seem to be white noise, and it is reasonable to accept the ARFIMA(0, 0.1915, 5) model. Table 2 gives the parameter estimates and their T test statistics of the selected ARFIMA(0, 0.1915, 5) model. The p-values, which are significantly smaller than 0.01, can also show that the ARFIMA(0, 0.1915, 5) model can fit the CGR-walk model of Sulfolobus solfataricus P2 effectively.

Then we analyse other eleven long protein sequences from the complete genomes of Archaea and Eubacteria. Data information and selected ARFIMA(p, d, q) models are all listed in Table 3 for the eleven protein sequences. The long-memory parameters d are all lying in (0, 0.5). It can be found that the ARFIMA(p, d, q) models of the protein sequences from the complete genomes of all Archaea (Aful, Mjan, Ssol) are very close; the ARFIMA(p, d, q) models of the protein sequences from the complete genomes of all Gram-positive Eubacteria (Mlep, Mpne, MtubH) except MtubH are very close; the ARFIMA(p, d, q) models of the protein sequences from the complete genomes of all Gram-negative Eubacteria (Aqua, Cpne, Bbur, Pmul, Syne and Buch) except Syne and Buch are very close.

Although the existence of the archaebacterial kingdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy.[31] The evolutionary relationship among the three primary kingdoms, namely, archeabacteria, eubacteria, and eukaryote, is another crucial problem that remains unresolved.[31] Our method shows some patterns which are useful for the classification problem, namely, ARFIMA(p, d, q) models from the same category are closer to each other. As a result, our method is sufficient to give a satisfactory relationship for the organisms selected.

The p-values of the LB test statistics at every lag k for every selected model are almost all significantly larger than 0.1 (see Table 4). And the p-values of the T test statistics for parameters of every selected model are significantly smaller than 0.01 except 0.0673 (< 0.1) (see Table 5). All of these indicate that the ARFIMA (p, d, q) models can fit the CGR-walk models of different protein sequences well. The CGR-walk model can generate a time series easily, and can be fitted highly significantly by a long-memory ARFIMA(p, d, q) model. From the above tables, we can see that the credibility and the accuracy of the model are really good, that is, we can effectively model the protein sequences through our method. As a classical time series model with perfect algorithm, ARFIMA model can help us find out more characteristics of protein sequences, recognie more major similarities among different protein sequences than other models, such as hidden Markov chain model. Our CGR-walk model provides direct approaches to both computational scientists and molecular biologists to analyse protein structures efficiently.

References

[1] Chothia C 1992 *Nature* (London) **357** 543

[2] Shih C T, Su Z Y, Gwan J F, Hao B L, Hsieh C H and Lee H C 2000 *Phys. Rev. Lett.* **84** 386

[3] Shih C T, Su Z Y, Gwan J F, Hao B L, Hsieh C H and Lee H C 2002 *Phys. Rev.* E **65** 41923

[4] Dill K A 1985 *Biochemistry* **24** 1501

[5] Chan H S and Dill K A 1989 *Macromolecules* **22** 4559

[6] Li M, Chen X, Badger J H, Kwong S, Kearney P and Zhang H Y 2001 *Bioinformatics* **17** 149

[7] Wang B and Yu Z G 2000 *J. Chem. Phys.* **112** 6084

[8] Wang J and Wang W 2000 *Phys. Rev.* E **61** 6981

[9] Brown T A 1998 *Genetics* 3rd ed (London: Chapman & Hall)

[10] Yu Z G, Anh V and Lau K S 2004 *Physica* A **337** 171

[11] Hao B L, Lee H C and Zhang S Y 2000 *Chaos, Solitons and Fractals* **11** 825

[12] Jeffrey H J 1990 *Nucleic Acids Res.* **18** 2163

[13] Peng C K, Buldyrev S V, Goldberg A L, Havlin S, Sciortino F, Simons M and Stanley H E 1992 *Nature* **356** 168
[14] Yu Z G, Anh V V and Lau K S 2001 *Phys. Rev.* E **64** 31903
[15] Fiser A, Tusnady G E and Simon I 1994 *J. Mol. Graphics* **12** 302
[16] Basu S, Pan A, Dutta C and Das J 1997 *J. Mol. Graphics Model* **15** 279
[17] Gao J and Xu Z Y 2009 *Chin. Phys.* B **18** 370
[18] Lopes S R C and Nunes M A 2006 *Physica* A **361** 569
[19] Beran J 1994 *Statistics for Long-memory Processes* (New York: Chapman Hall)
[20] Hurst H E 1951 *Trans. Am. Soc. Civil Eng.* **116** 770
[21] Granger C W J and Joyeux R 1980 *J. Time Ser. Anal.* **1** 15
[22] Hosking J R M 1981 *Biometrika* **68** 165
[23] Hosking J R M 1984 *Water Resour. Res.* **20** 1898
[24] Anfinsen C B 1973 *Science* **181** 223
[25] Sun X D, Liu Y J and Huang B X 2008 *Protein Structure Prediction — Supporting the Application of the Vector Machine* (Beijing: Science Press) (in Chinese)
[26] Yu Z G, Anh V V and Lau K S 2003 *Phys. Rev.* E **68** 021913
[27] Akaike H 1974 *IEEE Transaction on Automatics Control* **19** 716
[28] Akaike H 1978 *Ann. Inst. Math. Stat.* **30**A 9
[29] Ljung G M and Box G E P 1978 *Biometrika* **65** 297
[30] Li W K and Mcleod A I 1986 *Biometrika* **73** 217
[31] Iwabe N, Kuma K, Hasegawa M, Qsawa S and Miyata T 1989 *Proc. Natl. Acad. Sci.* (U.S.A.) **86** 9355