Vol 18 No 1, January 2009
1674-1056/2009/18(01)/0370-07

Chinese Physics B

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

Chaos game representation (CGR)-walk model for DNA sequences*

Gao Jie(高 洁)[a)b)†] 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 24 April 2008; revised manuscript received 27 August 2008)

Chaos game representation (CGR) is an iterative mapping technique that processes sequences of units, such as nucleotides in a DNA sequence or amino acids in a protein, in order to determine the coordinates of their positions in a continuous space. This distribution of positions has two features: one is unique, and the other is source sequence that can be recovered from the coordinates so that the distance between positions may serve as a measure of similarity between the corresponding sequences. A CGR-walk model is proposed based on CGR coordinates for the DNA sequences. The CGR coordinates are converted into a time series, and a long-memory ARFIMA (p, d, q) model, where ARFIMA stands for autoregressive fractionally integrated moving average, is introduced into the DNA sequence analysis. This model is applied to simulating real CGR-walk sequence data of ten genomic sequences. Remarkably long-range correlations are uncovered in the data, and the results from these models are reasonably fitted with those from the ARFIMA (p, d, q) model.

Keywords: CGR-walk model, DNA sequence, long-memory, ARFIMA(p, d, q) model
PACC: 8710, 0250

1. Introduction

DNA sequences are of fundamental importance in understanding living organisms, since all the information of the hereditary and species evolution is contained in these macromolecules. The DNA sequence comprises four different nucleotides, namely adenine (A), cytosine (C), guanine (G), and thymine (T). A large number of these DNA sequences are widely available nowadays. One of the challenges of analysing the DNA sequences is to determine the patterns in these sequences. It is useful to distinguish coding from noncoding sequences.

After DNA is sequenced, we need to find their different functional regions. How to gain more bioinformation from these DNA sequences is a challenging problem. The nucleotides stored in GenBank have exceeded hundreds of millions of bases and they increase by ten times every five years. It has become important to improve new theoretical methods of conducting DNA sequences analysis. Many biologists, physicists, mathematicians and computer specialists are attracted to this interesting research field.

A great number of studies of the long-range correlation in DNA sequences have been carried out. One-dimensional (1D) DNA walk was first proposed by Peng *et al.*[1] They defined the walk 'up' step as $u(i) = +1$ when a pyrimidine (C or T) occurred at position i along the DNA chain, while walk 'down' step as $u(i) = -1$ when a purine (A or G) emerged at position i. Using the exponent α from the fluctuation analysis, they found that for coding sequences, the exponent was close to 0.5 corresponding to the case where the walk was random or displayed a local correlation; for noncoding sequences, $\alpha \approx 0.67 > 0.5$ explicitly, which corresponded to the case where the walk displayed a long-range correlation. But their result has not been fully accepted by other researchers because in their DNA walk mode A could not be distinguished from G in purine and C from T in pyrimidine. Then they gave some explanations and improvements in Refs.[2, 3]. By making more detailed analysis, Chatzidimitriou and Larhammar[4] concluded that both coding and noncoding sequences exhibited a long-range correlation. In their subsequent work, Prabhu and Claverie[5] also substantially corroborated these results. On the other hand, Buldyrev *et al*[6,7] developed a generalized Lévy-walk model to generate

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

a model sequence which is in many ways similar to the statistics obtained from the empirical sequence data, and showed that the long-range correlation appeared mainly in noncoding DNA by using all the DNA sequences available. According to this model, Tai *et al*[8] proposed a two-dimensional modified Lévy-walk model and found the value of power (α) to range from 0.64 to 0.68. If one considers more details by distinguishing C from T in pyrimidine, and A from G in purine such as two- or three-dimensional DNA walk models[9] and maps,[10−12] then the base correlation can be found to be present even in coding sequences. Yu *et al*[10,11] viewed the sequence as a time series and used it to reveal more information.

In this paper, we construct a chaos game representation (CGR)-walk model based on CGR coordinates for DNA sequences. The CGR coordinates are converted into a time series, and a long-memory ARFIMA (p, d, q) model, where ARFIMA stands for autoregressive fractionally integrated moving average, is introduced to DNA sequence analysis. This model is applied to simulating the real CGR-walk sequence data of ten genomic sequences. We uncover in the data remarkably long-range correlations and find that the results from these models can reasonably be fitted with those from the ARFIMA (p, d, q) model.

2. ARFIMA model

In this section, we present the ARFIMA(p, d, q) model (also called fractional autoregressive integrated moving average (ARIMA) model) and some relevant 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 (see Ref.[13] for more details of these models). Initial studies of time series with long memory characteristics were given by Hurst.[14] ARFIMA processes first appeared in Refs.[15, 16] and are the extension of the autoregressive moving average (ARMA) and ARIMA models. The author of Ref.[17] was a pioneer in the application of long memory to 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 characterize the persistence in two different forms, i.e.

- in the time domain, where the autocorrelation function $\rho_x(\cdot)$ decays hyperbolically to zero, that is, $\rho_x(k) \sim k^{2d-1}$ when $k \to \infty$.
- in the frequency domain, where the spectral density function $f_x(\cdot)$ is unbounded when the frequency is near zero, that is, $f_x(w) \sim w^{-2d}$ when $w \to 0$.

One of the models that can describe the persistence is the so-called ARFIMA (p, d, q) process.

Definition 1 A stochastic process $\{X_t\}_{t\in Z}$ is Gaussian if, for any set of $t_1, t_2, \ldots, t_n \in Z$, the random variables $X_{t1}, X_{t2}, \ldots, X_{tn}$ have an n-dimensional normal distribution.

We observe that weakly stationary process $\{X_t\}_{t\in Z}$ need not be strongly stationary. However, any weakly stationary Gaussian process will be also strongly stationary.[18]

Definition 2 The process $\{\varepsilon_t\}_{t\in Z}$ is said to be a white noise process with zero mean and variance σ_ε^2, denoted by $\varepsilon_t \sim \mathrm{WN}(0, \sigma_\varepsilon^2)$, if

$$E(\varepsilon_t) = 0, \mathrm{Var}(\varepsilon_t) = E(\varepsilon_t^2) = \sigma_\varepsilon^2,$$

and

$$\gamma_\varepsilon(k) = \begin{cases} \sigma_\varepsilon^2, & k = 0, \\ 0, & k \neq 0. \end{cases} \tag{1}$$

Definition 3 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

$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t\ , \quad t \in Z, \tag{2}$$

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

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

and

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

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.

The term $(1-B)^d$, for $d \in R$, is determined through the binomial expansion

$$\begin{aligned}(1-B)^d &= \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \\ &= 1 - dB - \frac{d}{2!}(1-d)B^2 \ldots .\end{aligned}$$

If $d \in (-0.5,\ 0.5)$, then $\{X_t\}_{t\in Z}$ is a stationary and an invertible process. The most important

characteristics of an ARFIMA(p, d, q) process are 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

CGR was proposed as a scale-independent representation for genomic sequences by Jeffrey[19] in 1990. The technique, formally an iterative mapping, can be traced further back to the foundation of statistical mechanics, in particular, to Chaos theory.[20] The original proposition has been considerably expanded and generalized to sequences of arbitrary symbols,[21] and therefore they have included other biological sequences such as proteins.[22,23] However, the possibility that the CGR format can be used for representing the nucleotide sequence as well as identifying the resulting sequence scheme has never been fully explored. The CGR space is a continuous reference system where all possible sequences of any length have a unique position. Consequently, all possible nucleotide succession schemes will be encoded in the continuous space.

The CGR space generated by genomic sequences is planar, and it is confined by the four possible nucleotides as vertices of a binary square (Fig.1). The CGR coordinates are calculated iteratively by moving a pointer to half the distance between the previous position and the current binary representation (Eq.(3)). The binary CGR vertices are assigned to the four nucleotides as $A = (0,0), C = (0,1), G = (1,1)$, and $T = (1,0)$ and (0.5, 0.5) as an arbitrary starting position. The procedure is illustrated by analysing the sequence of NC_005336 orf virus in Fig.1.

$$\mathrm{CGR}_i = \mathrm{CGR}_{i-1} - 0.5 \cdot (\mathrm{CGR}_{i-1} - g_i),$$
$$i = 1, \cdots, n_G, \quad \mathrm{CGR}_0 = (0.5,\ 0.5). \tag{3}$$

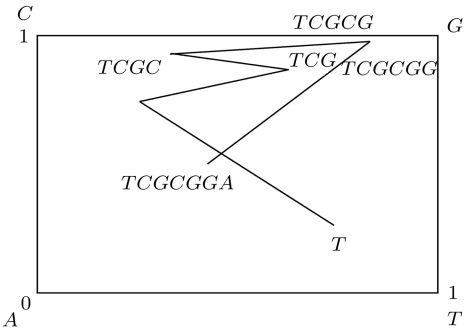


Fig.1. CGR of the first 7 nucleotides of NC_005336 orf virus: *TCGCGGA*.

For a DNA sequence, we define an equation as follows: $t_k = y_k/x_k$, where y_k is the y-coordinate of CGR_k, x_k is the x-coordinate of CGR_k, then we obtain a data sequence $\{t_k : k = 1, 2, \ldots, N\}$, which we term a 'CGR-walk model'.

4. Analysis and discussion

4.1. Data analysis for the DNA sequence of NC_005336 orf virus

In order to illustrate the long-range correlation in DNA sequences, we analyse a CGR-walk model for a DNA sequence of NC_005336 orf virus.

Figure 2(a) displays a CGR-walk sequence plot of NC_005336 orf virus (positions 2745–3745) with a total of 1001 observations, i.e. $n = 1001$. Owing to increasing variability and trends in the data, the first difference of the log of the 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 degree of heteroscedasticity is observed.

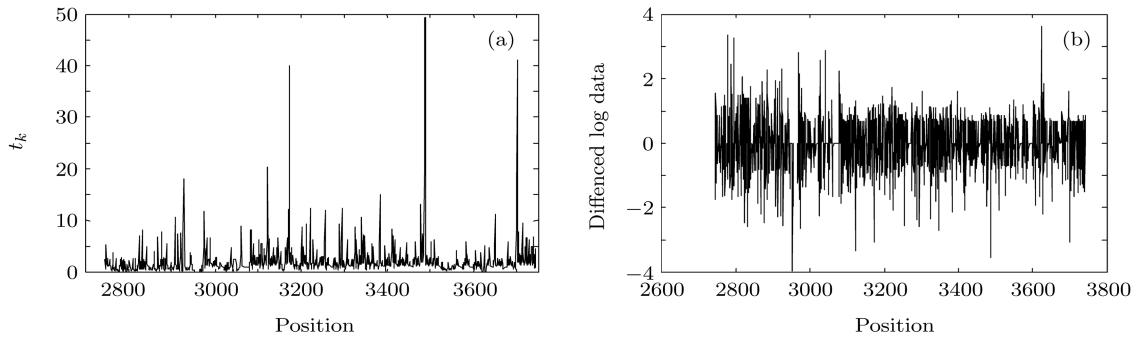


Fig.2. CGR-walk sequence of NC_005336 orf virus with a total of 1001 pairs of bases (a) and first differenced log data (b).

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 indicated in Fig.3(b).

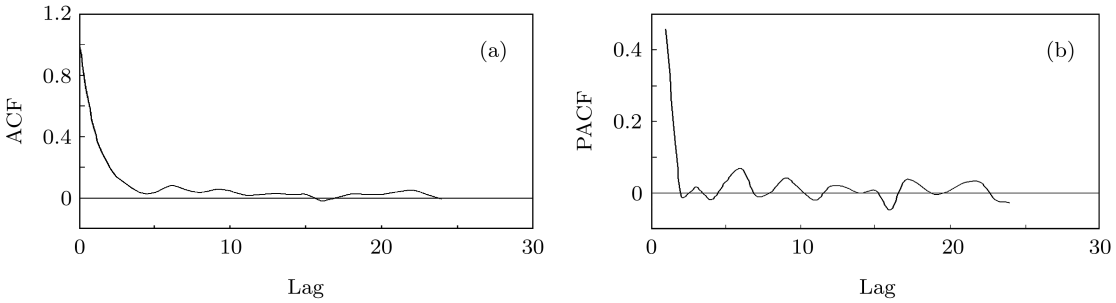


Fig.3. Sample ACF of the CGR-walk data (a) and sample PACF of the CGR-walk data (b).

The ACF of the differenced log data is given in Fig.4(a), and the PACF of the differenced log data is presented in Fig.4(b). The ACF of the differenced log data decays rapidly, while the PACF decay slowly, which seems to indicate the presence of long-memory component in the initial data.

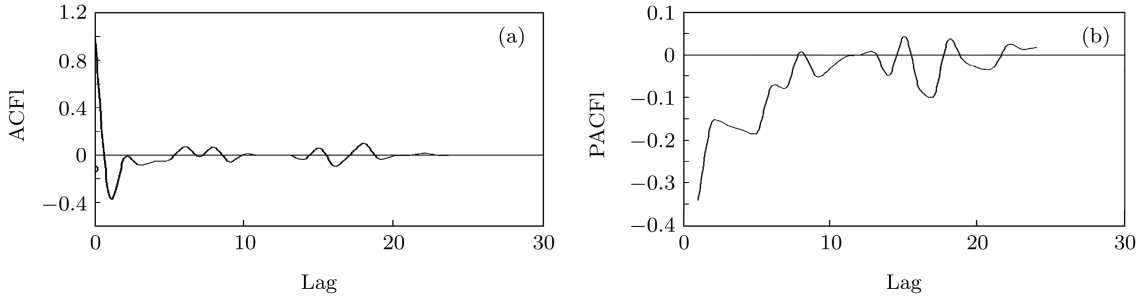


Fig.4. Sample ACF of the differenced log data (a), and sample PACF of the differenced log data (b).

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.[13], for a long-range dependent time series $\{x_t\}$ the variance of its mean values $\bar{x}_k$ satisfies

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

and

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

Therefore by plotting $\log[\mathrm{Var}(\bar{x}_k)]$ versus $\log(k)$ for different values of k, a straight line with a slope of $2d-1$ 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 -0.6383 through the least squares estimation, suggesting that a crude estimate of the long-memory parameter ($\hat{d}$) is 0.18, i.e. $\hat{d}=0.18$.

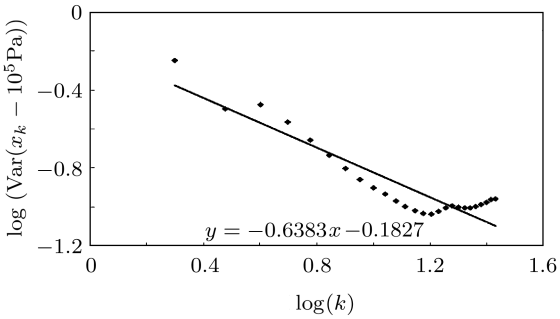


Fig.5. Variance plot, where solid line is the fitted straight line with a slope of -0.6383.

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

Accordingly, a class of ARFIMA (p, d, q) models, with the values of p and q both taken to be less than

or equal to 5. Based on the Akaike's information criterion (AIC),[24,25] the ARFIMA (0, 0.18, 3) model is selected.

4.2. Model test and parameter estimate for the DNA sequence of NC_005336 orf virus

To test the selected model, we choose a suitable test statistics, i.e. the modified portmanteau test (LB test)[26,27]

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

where r_k is the sample autocorrelation at lag k, n is the sample size, and M is a presetting 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 well reasonable to accept the ARFIMA (0, 0.18, 3) model.

Table 1. Autocorrelation check of residuals.

To Lag	Chi-Square	Pr > ChiSq
6	5.47	0.1402
12	8.78	0.4579
18	13.02	0.6011
24	15.44	0.8003
30	16.34	0.9462
36	19.89	0.9650
42	25.27	0.9563
48	25.76	0.9906

Table 2 gives the parameter estimates of the selected ARFIMA (0, 0.18, 3) model. The p-values of the T test statistics for four parameters are significantly smaller than 0.005 (see Table 2). This indicates that the ARFIMA (0, 0.18, 3) model can fit the CGR-walk model of NC_005336 orf virus effectively.

Table 2. Conditional least squares estimation.

parameter	estimate	standard error	t value	Pr> $\lvert t\rvert$	Lag
MU	2.23284	0.17625	12.67	< .0001	0
θ_1	–0.46072	0.03153	–14.61	< .0001	1
θ_2	–0.19716	0.03417	–5.77	< .0001	2
θ_3	–0.09812	0.03154	–3.11	0.0019	3

4.3. Data analysis for other DNA sequences

In order to examine whether this CGR-walk model works for other sequences, we analyse the genomic sequences of Acanthamoeba polyphaga mimivirus, Acheta domesticus densovirus, Acyrthosiphon pisum virus, Aconitum latent virus, Acute bee paralysis virus, Amsacta moorei entomopoxvirus, Mice minute virus, Human adenovirus C and Homo sapiens dystrophin. All data are available from the NCBI (National Center for Biotechnology Information) website. Its homepage address is http://www.ncbi.nlm.nih.gov/.

Data information, selected ARFIMA (p, d, q) models and parameter estimates are all listed in Table 3 for the above nine genomic sequences. The values of long-memory parameter (d) lie in an interval (0, 0.5).

Table 3. Data information, selected ARFIMA models and parameter estimates for nine genomic sequences.

name[a]	positions	sample size	selected model	parameter estimate
A	859–686	828	ARFIMA(1,0.312,1)	$\Phi_1 = 0.69099$, $\theta_1 = 0.99999$
B	639–485	847	ARFIMA(1,0.34,1)	$\Phi_1 = 0.75730$, $\theta_1 = 0.99997$
C	683–605	923	ARFIMA(1,0.338,1)	$\Phi_1 = 0.50959$, $\theta_1 = 0.99044$
D	727–690	964	ARFIMA(1,0.284,1)	$\Phi_1 = 0.61788$, $\theta_1 = 0.99999$
E	2399–38	982	ARFIMA(1,0.349,2)	$\Phi_1 = -0.99994$, $\theta_1 = -0.73218$, θ_2=0.26731
F	1057–04	987	ARFIMA(0,0.479,4)	$\theta_1 = 0.29106, \theta_2 = 0.28157, \theta_3 = 0.20963, \theta_4 = 0.20595$
G	705–824	1120	ARFIMA(0,0.496,4)	$\theta_1 = 0.31078, \theta_2 = 0.25452$, $\theta_3 = 0.17224, \theta_4 = 0.16299$
H	441–561	1121	ARFIMA(0,0.16,0)	
I	485–628	1144	ARFIMA(1,0.348,1)	$\Phi_1 = 0.5412$, $\theta_1 = 0.9969$

[a]Capital letters A, B, C, D, E, F, G, H, and I respectively denote Mice minute virus (NC_001510), Acute bee paralysis virus (NC_002548), Acheta domesticus densovirus (NC_004290), Acanthamoeba polyphaga mimivirus (NC_006450), Aconitum latent virus, (NC_002795), Acyrthosiphon pisum virus (NC_003780), Human adenovirus C (NC_001405), Amsacta moorei entomopoxvirus (NC_002520), and Homo sapiens dystrophin, (NM_004023).

The p-values of the LB test statistics at each value of lag k for each selected model are significantly larger than 0.1. And the p-values of the T test statistics for parameters of each selected model are all significantly smaller than 0.01. All of these indicate that the ARFIMA (p, d, q) models can fit the CGR-walk models of different DNA sequences well.

5. Conclusion

The CGR of sequences is a method to coordinate the entire domain of possibilities in a continuous two-dimensional space. The CGR transformation makes DNA sequences amenable to an entirely new set of statistical analysis tools. Therefore, the CGR is a formalism that bridges between sequences of discrete units and numeric coordinates in a continuous space.

Although quite a lot of studies have been carried out by taking into consideration the long-range correlations in DNA sequences, the models and methods are somewhat rough and the results obtained from these models are not satisfactory. In this paper, we convert the CGR coordinates into a time series (CGR-walk model) and introduce a long-memory ARFIMA (p, d, q) model into the DNA sequence analysis. Consequently, basic statistic method and time series analysis technique can now be applied to the CGR-walk model.

We first analyse the real DNA sequence data of the NC_005336 orf virus genome. From Figs.2–5 we detect the presence of long-memory behaviour in the data. Based on AIC, the ARFIMA (0, 0.18, 3) model is selected to fit the CGR-walk sequence data of the NC_005336 orf virus genome. 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, 0.18, 3) model. Table 2 gives the parameter estimates and their T test statistics of the selected ARFIMA (0, 0.18, 3) model. The p-values, which are significantly smaller than 0.005, can also tell us whether the ARFIMA (0, 0.18, 3) model can fit the CGR-walk model of NC_005336 orf virus effectively.

Then we analyse the genomic sequences of Acanthamoeba polyphaga mimivirus, Acheta domesticus densovirus, Acyrthosiphon pisum virus, Aconitum latent virus, Acute bee paralysis virus, Amsacta moorei entomopoxvirus, Mice minute virus, Human adenovirus C and Homo sapiens dystrophin. Data information, selected ARFIMA (p,d,q) models and parameter estimates are all listed in Table 3 for the nine genomic sequences. The values of long-memory parameter (d) lie in an interval (0, 0.5). The p-values of the LB test statistics are significantly larger than 0.1. And the p-values of the T test statistics for parameters are all significantly smaller than 0.01. All of these indicate that the ARFIMA (p, d, q) models can well fit the CGR-walk models of different DNA sequences.

In the 'DNA-walk' analysis for the long-range correlations, the 1D-walk model proposed by Peng *et al*[1] and the generalized Lévy-walk model[6] both are apparently rough. The time series model proposed by Yu and Anh[28] obtained only Hurst exponent H, and the two-dimensional modified Lévy-walk model[8] obtained only the value of power α as well. They distinguished the long-range correlations in DNA sequences only by the value of power α or Hurst exponent H. Furthermore, they did not test their models, and they did not provide the credibility and the accuracy for their models either. In this paper, we can see that the CGR-walk model can generate a model sequence easily, and they can be fitted with a long-memory ARFIMA (p, d, q) model well. From the above tables, we can see that the credibility and the accuracy of the model are very good indeed. As a classical time series model with a perfect algorithm, the ARFIMA model can help us predict DNA sequences and solve many other problems.

References

[1] Peng C K, Buldyrev S V, Goldberg A L, Havlin S, Sciortino F, Simons M and Stanley H E 1992 *Nature* **356** 168

[2] Buldyrev S V, Dokholyan N V, Goldberg A L, Havlin S, Peng C K, Stanley H E and Visvanathan G M 1998 *Physica* A **249** 430

[3] Buldyrev S V, Goldberg A L, Havlin S, Peng C K and Stanley H E in: 1994 Bunde A and Havlin S (eds) *Fractals in Science* (Berlin: Springer) pp49–87

[4] Chatzidimitriou-Dreismann C A and Larhammar D 1993 *Nature* **361** 212

[5] Prabhu V V and Claverie J M 1992 *Nature* **359** 782

[6] Buldyrev S V, Goldberger A L, Havlin S, Peng C K, Simons M and Stanley H E 1993 *Phys. Rev.* E **47** 4514

[7] Buldyrev S V, Goldberger A L, Havlin S, Mantegna R N, Matsa M E, Peng C K, Simons M and Stanley H E 1995 *Phys. Rev.* E **51** 5084

[8] Tai Y Y, Li P C and Tseng H C 2006 *Physica* A **369** 688

[9] Luo L F, Lee W J, Jia L J, Ji F M and Tsai L 1998 *Phys. Rev.* E **58** 861
[10] Yu Z G and Chen G Y 2000 *Commun. Theor. Phys.* **33** 673
[11] Yu Z G, Anh V, Gong Z M and Long S C 2002 *Chin. Phys.* **11** 1313
[12] Liu T, Wang Y and Wang K L 2007 *Chin. Phys.* **16** 272
[13] Beran J 1994 *Statistics for Long-Memory Processes* (New York: Chapman Hall)
[14] Hurst H E 1951 *Trans. Amer. Soc. Civil Eng.* **116** 770
[15] Granger C W J and Joyeux R 1980 *J. Time Ser. Anal.* **1** 15
[16] Hosking J R M 1981 *Biometrika* **68** 165
[17] Hosking J R M 1984 *Water Resour. Res.* **20** 1898
[18] Brockwell P J and Davis R A 1991 *Time Series: Theory and Methods* (New York: Springer)
[19] Jeffrey H J 1990 *Nucleic Acids Res.* **18** 2163
[20] Bar-Yam Y 1997 *Dynamics of Complex Systems* (Cambridge, MA: Rersens)
[21] Tino P 1999 *IEEE Trans. Syst. Man Cybernet.* **29** 386
[22] Basu S, Pam A, Dutta C and Das J 1997 *J. Mol. Graph. Model* **15** 279
[23] PleiSner K P, Wernisch L, Osvald H and Fleck E 1997 *Electrophoresis* **18** 1709
[24] Hosking J R M 1984 *Water Resources Research* **20** 1898
[25] Crato N and Ray B K 1996 *J. Forecasting* **15** 107
[26] Ljung G M and Box G E P 1978 *Biometrika* **65** 297
[27] Li W K and Mcleod A I 1986 *Biometrika* **73** 217
[28] Yu Z G and Anh V 2001 *Chaos, Solitons and Fractals* **12** 1827