Low dimensional chaos in the AT and GC skew profiles of DNA sequences*

Zhou Qian(周 茜)† and Chen Zeng-Qiang(陈增强)

Department of Automation, Nankai University, Tianjin 300071, *China*

(Received 27 January 2010; revised manuscript received 16 February 2010)

This paper investigates the existence of low-dimensional deterministic chaos in the AT and GC skew profiles of DNA sequences. It has taken DNA sequences from eight organisms as samples. The skew profiles are analysed using continuous wavelet transform and then nonlinear time series methods. The invariant measures of correlation dimension and the largest Lyapunov exponent are calculated. It is demonstrated that the AT and GC skew profiles of these DNA sequences all exhibit low dimensional chaotic behaviour. It suggests that chaotic properties may be ubiquitous in the DNA sequences of all organisms.

Keywords: chaos, phase space reconstruction, DNA sequences, AT and GC skew profiles
PACC: 0545, 8715

1. Introduction

Chaos is a ubiquitous phenomenon in nature. It is applied to deterministic dynamical systems which are aperiodic and exhibit sensitive dependence on initial conditions. The researches of chaos[1−6] on one hand advance the development of nonlinear science, and on the other hand make chaos become a tool and method to solve problems in other disciplines. Biological systems are complex nonlinear systems in which chaotic phenomenon have been reported.[7−11] Therefore it is possible that chaotic theory may be used to explain or understand the complexity of life activity.[12]

The DNA is a nucleic acid which contains the genetic instructions used in the development and functioning of all known living organisms and some viruses.[13] In living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules which are held tightly together.[14] These two long strands entwine like vines, in the shape of a double helix as shown in Fig.1. A DNA sequence is a succession of letters representing the primary structure of a DNA molecule strand. It is made up of four nucleotides: guanine G, adenine A, cytosine C, thymine T, which carry the genetic information as well as the evolutionary information of living organisms. Consequently, manipulation, searching, and mining useful information from these DNA sequences have recently become a real challenge and hot subject of research.[15−17] Nicolay *et al.*[18,19] use the continuous wavelet transform to perform a space-scale analysis of the AT and GC skews[20−22] (strand asymmetries) in human genomic sequences, which have been shown to correlate with gene transcription. It is shown that the observed nonlinear oscillations are likely to display all the characteristic features of chaotic strange attractors. They discussed the possibility that replication and gene regulation processes are governed by a low-dimensional dynamical system that displays deterministic chaos.

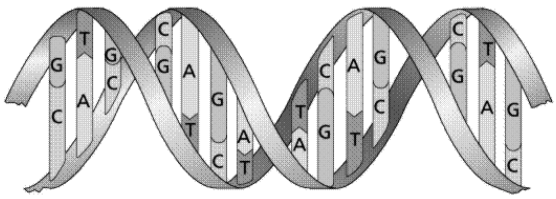


Fig. 1. The structure of part of a DNA double helix.

We wonder if the chaotic features also exist in the skew profiles of other organisms. In this paper, we investigate the AT and GC skew profiles of Archaea, Bacteria, Plants and Animals, each kind with two typical samples. The skew profiles are analysed using continuous wavelet transform and then nonlinear time series methods. The invariant measures of

*Project supported in part by the National Natural Science Foundation of China (Grant No. 60774088), the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20090031110029), and the Foundation of the Application Base and Frontier Technology Research Project of Tianjin (Grant No. 08JCZDJC21900).
†Corresponding author. E-mail: zhouqian@mail.nankai.edu.cn
© 2010 Chinese Physical Society and IOP Publishing Ltd

correlation dimension and the largest Lyapunov exponent are calculated. It is demonstrated that the AT and GC skew profiles of these DNA sequences exhibit similar low-dimensional chaotic behaviour. We suggest that chaotic properties may be ubiquitous in the DNA sequences of all organisms.

2. Data

In our work the eight DNA sequences samples are Methanosarcina barkeri str. fusaro (NC_007355.1) and Methanosarcina acetivorans C2A (NC_003552.1) from Archaea, Sorangium cellulosum 'So ce 56' (NC_010162.1) and Chitinophaga pinensis DSM 2588 (NC_013132.1) from Bacteria, Oryza sativa chromosome 1 (NC_008394.1) and Arabidopsis thaliana chromosome 2 (NC_003071.7) from Plants, contigs of Pan troglodytes chromosome 21 (NT_106996.1) and Mus musculus chromosome 6 (NT_039340.7) from Animals. These DNA sequences were all downloaded from GenBank (http://www.ncbi.nlm.nih.gov/). The AT and GC skew profiles, namely $S_{\rm AT} = (A-T)/(A+T)$ and $S_{\rm GC} = (G-C)/(G+C)$[20] in adjacent 1 kb windows, are computed for each chromosome or contig. We consider $S = S_{\rm AT} - S_{\rm GC}$, in order to obtain larger fluctuation amplitudes than those displayed by each individual skew.

3. Methods

3.1. Continuous wavelet transformation

Continuous wavelet transformation possesses the ability to construct a time-frequency representation of a signal that offers very good time and frequency localization. In mathematics, the continuous wavelet transform of a continuous, square-integrable function $x(t)$ at a scale $a > 0$ and translational value b is expressed by the following integral

$$X(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) {\rm d}t, \qquad (1)$$

where $\psi(t)$ is a continuous function in both the time domain and the frequency domain. It is called the mother wavelet. In our work, we use the first derivative of the Gaussian function $g^{(0)}(x) = {\rm e}^{-x^2/2}/\sqrt{2\pi}$, namely $g^{(1)}(x) = {\rm d}g^{(0)}(x)/{\rm d}x$.

3.2. Phase space reconstruction

Phase space reconstruction is a powerful technique to recover the topological structure of a dynamical system given by a scalar time series. The reconstructed attractor preserves the invariant characteristics of the original one. In our study, time delay embedding[16] is used to reconstruct the attractor of skew profiles.

Suppose a single system variable time series x_1, x_2, $\ldots x_N$ of a dynamical system is observable, according to Taken's theorem[23] the original phase space geometry of the system may be reconstructed in the m-dimensional phase space. The reconstructed attractor of the original system is given by the vector sequence:

$$P(i) = (x_i, x_{i+\tau}, x_{i+2\tau} ..., x_{i+(m-1)\tau}), \qquad (2)$$

where τ and m are the embedding delay and the embedding dimension, respectively. Takens states that for an infinite, noise free data series and a large enough m, this procedure provides a one-to-one image of the original system. A sufficient condition is $m \geq 2d+1$, where d is the dimension of the original attractor.

3.3. Time delay

A reasonable time delay τ is very important for efficient phase space reconstruction because usually we have to deal with a finite amount of noisy data. A small delay will yield strongly correlated vector elements, while a large delay leads to vectors whose components are almost uncorrelated.[24] The time delayed mutual information method suggested by Fraser and Swinney[25] is commonly used to determine a reasonable delay. Unlike the autocorrelation function, the mutual information takes into account also the nonlinear correlations. A measure of mutual information between data of a time series $X(i)$, $i = 1, 2, \ldots$ is

$$S(\tau) = \sum_{ij} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j}, \qquad (3)$$

where p_i is the probability to find a time series value in the i-th interval, and p_{ij} is the joint probability that an observation falls into the i-th interval and the observation time τ later falls into the j-th interval. The first minimum of $S(\tau)$ is chosen as a delay time for the reconstruction of phase space.

3.4. Embedding dimension

The false nearest neighbour (FNN) algorithm suggested by Kennel *et al.*[26] is the most commonly employed method to determine the minimal

sufficient embedding dimension. The idea of the algorithm is the following. For each point X_n in the time series we look for its nearest neighbour X_l in an m-dimensional space. Calculate the separation between these two points $R_n(m) = \sqrt{(X_l - X_n)^2 + (X_{l-1} - X_{n-1})^2 + \ldots}$. Then compute the separation $R_n(m+1)$ in an $m+1$-dimensional space. If $R_n(m+1)$ significantly exceeds $R_n(m)$, this point is marked as having an FNN. The criterion that the embedding dimension is high enough is that the fraction of FNNs is zero, or at least sufficiently small.

3.5. Correlation dimension

The correlation dimension method is used for detecting the possible presence of chaos. An algorithm suggested by Grassberger and Procaccia[27] is the most commonly employed method. Let $C(\varepsilon)$ be the number of points within all the sphere of radius ε,

$$C(\varepsilon) = \frac{2}{N(N-1)} \sum_{j=1}^{N} \sum_{i=j+1}^{N} \Theta(\varepsilon - \varepsilon_{ij}), \qquad (4)$$

where Θ is the Heaviside function defined as $\Theta(x) = \begin{cases} 0 \text{ for } x < 0 \\ 1 \text{ for } x \geq 0 \end{cases}$ and ε_{ij} is the spatial separation between two points labeled i and j. The $C(r)$ is called correlation sum. A plot of $\log C(r)$ versus $\log r$ should give an approximately straight line whose slope in the limit of small r and large N is the correlation dimension

$$D_2 = \lim_{r \to 0} \lim_{N \to \infty} \frac{\mathrm{d} \log C(r)}{\mathrm{d} \log r}. \qquad (5)$$

If the system is chaotic, the slope of $\log C(r)$ vs $\log r$ converges to m over an appropriate interval as m increases.

3.6. Largest Lyapunov exponent (LLE)

The Lyapunov exponent measures how predictable or unpredictable the system is. For a chaotic system, there is at least one positive Lyapunov exponent. According to the methodology of Kantz,[28] we consider the representation of the time series as a trajectory in the embedding space. Assume that we observe a close return $s_{n'}$ to a previously visited point s_n. Then one can consider the distance $d_0 = \|s_n - s_{n'}\|$ as a small perturbation, which should grow exponentially in time. At time t the distance is $d_t = \|s_{n+t} - s_{n'+t}\|$. For a chaotic system one obtains $|d_t| \approx d_0 \mathrm{e}^{\lambda t}$, where λ is the LLE. In practice, one computes

$$S(\varepsilon, m, t) = \left\langle \ln \left(\frac{1}{|u_n|} \sum_{s_{n'} \in u_n} |s_{n+t} - s_{n'+t}| \right) \right\rangle_n . \qquad (6)$$

If $S(\varepsilon, m, t)$ exhibits a linear increase with identical slope for all m larger than some m_0 and for a reasonable range of ε, then this slope can be taken as an estimate of the LLE.

4. Analysis

According to Ref. [18], for each sample first we use the continuous wavelet transform to perform a space-scale analysis of the AT and GC skews of the selected genomic sequences. The analysing wavelet is first derivative of Gaussian function $g^{(1)}(x)$. A reasonable filtering scale has been chosen to smooth out the high frequency noisy fluctuations of the profile. The denoised skew profiles are then used to phase space reconstruction.

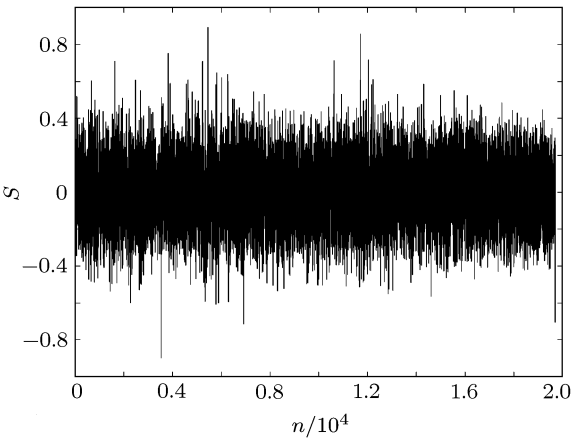


Fig. 2. Skew profile S of Arabidopsis thaliana chromosome 2 (NC_003071.7).

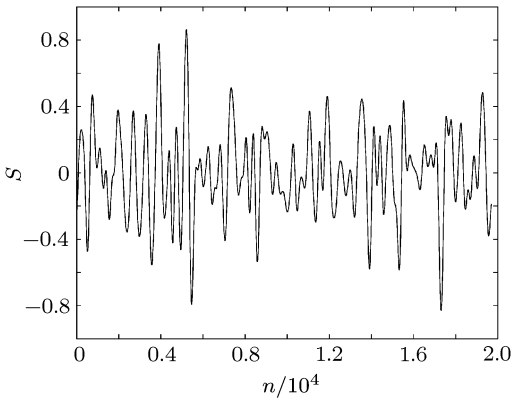


Fig. 3. Denoised skew profile S' of Arabidopsis thaliana chromosome 2 after low-pass filtering at scale $s = 160$ kb.

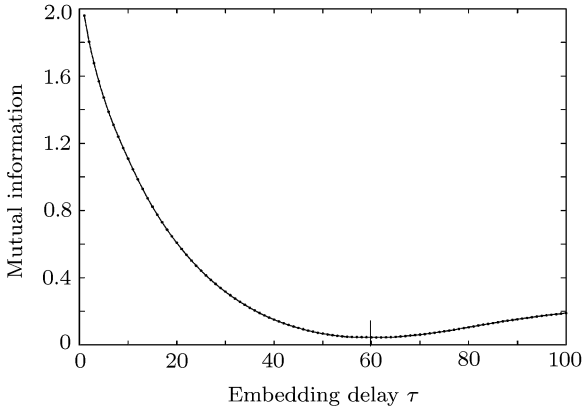


Fig. 4. Result of mutual information method for the denoised skew profile S'.

Take the analysis of Arabidopsis thaliana as example. Figure 2 shows the skew profile S of Arabidopsis thaliana chromosome 2. Using $g^{(1)}(x)$ as smoothing filter, we show the skew profile obtained at the scale $s = 160$ kb in Fig. 3. Mutual information method presents its first minimum at $\tau = 60kb$, which is chosen as the appropriate time delay for phase space reconstruction as shown in Fig. 4. The results obtained with the FNN method are presented in Fig. 5. It can be well observed that the fraction of FNN convincingly drops to zero for $m = 4$. So we take the minimal sufficient embedding dimension as 4. With $\tau = 60$ kb and $m = 4$, the reconstructed attractor is shown in Fig. 6(e).

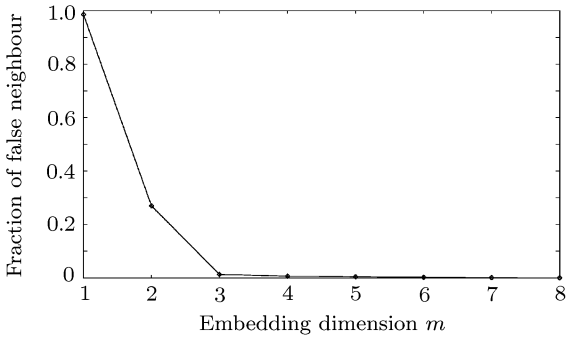
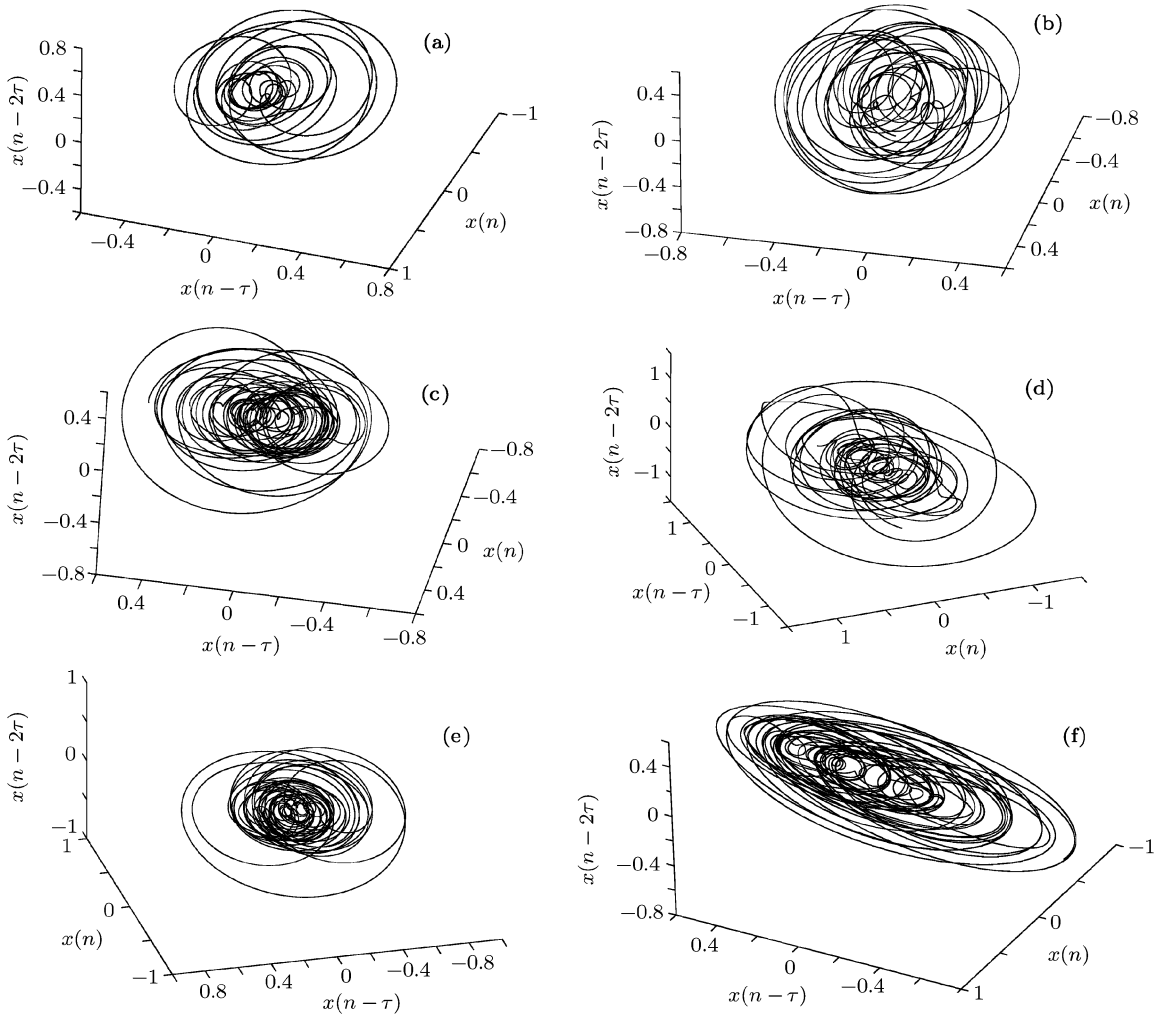


Fig. 5. Result of FNN method for the denoised skew profile S'.



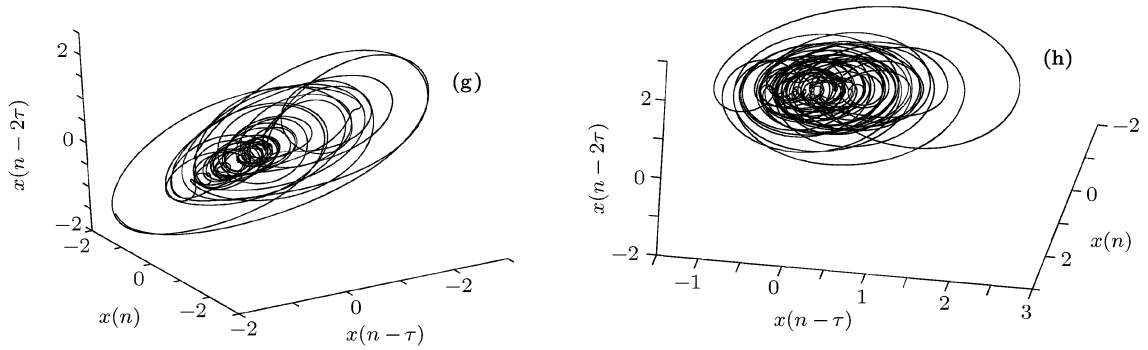


Fig. 6. Reconstructed attractors of the eight samples with $m = 4$: (a) Methanosarcina barkeri str. fusaro (NC_007355.1) with $s = 60$ kb, $\tau = 40$ kb; (b) Methanosarcina acetivorans C2A (NC_003552.1) with $s = 65$ kb, $\tau = 40$ kb; (c) Sorangium cellulosum 'So ce 56' (NC_010162.1) with $s = 100$ kb, $\tau = 40$ kb; (d) Chitinophaga pinensis DSM 2588 (NC_013132.1) with $s = 80$ kb, $\tau = 40$ kb; (e) Oryza sativa chromosome 1 (NC_008394.1) with $s = 160$ kb, $\tau = 60$ kb; (f) Arabidopsis thaliana chromosome 2 (NC_003071.7) with $s = 160$ kb, $\tau = 60$ kb; (g) Pan troglodytes chromosome 21 (NT_106996.1) with $s = 160$ kb, $\tau = 60$ kb; (h) Mus musculus chromosome 6 (NT_039340.7) with $s = 160$ kb, $\tau = 60$ kb.

Next, using program $d2$ in TISEAN package,[28] we obtain the correlation dimension of the denoised skew profile S' of Arabidopsis thaliana. The $D_2 \approx 1.25$ as shown in Fig. 7. This value indicates the possible existence of low-dimensional attractor. Further, using the TISEAN package we compute the LLE for the denoised skew profile. In Fig. 8, the computed LLE is the mean value of local slopes gathered from the linear regime when $m = 4, 5, 6$ and converges to 0.00725. This positive value indicates the existence of chaotic behaviour and is consistent with the estimation from the correlation dimension method and the FNN method as discussed above.

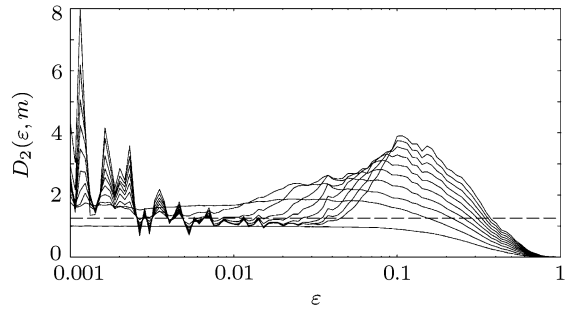


Fig. 7. Estimation the correlation dimension D_2 of the denoised skew profile S', ε is the radius of a sphere in the phase space.

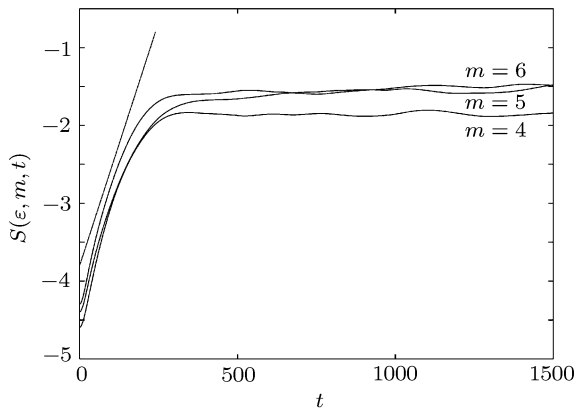


Fig. 8. Estimation the LLE of the denoised skew profile S'.

Finally, we report our results of all these eight samples. In fact, just like what Nicolay *et al.* find in human DNA sequences, we detect low-dimensional chaotic behaviour in the skew profiles of all these DNA sequences. Figure 5 shows the reconstructed attractors of these samples. We see that the attractors share similar shape. They all look like two strands of intersecting rings. The values of the LLE and correlation dimension of these DNA sequences are shown in Table 1. It is shown that the LLE of these samples are very close to each other and are in the range of [0.00675, 0.0095]. The values of correlation dimension are in the range of [1.15, 1.50]. Thus these DNA sequences exhibit similar low-dimensional chaotic behaviour.

Table 1. Computation of the largest Lyapunov exponents and D_2 of the eight samples.

sample name		$\lambda_{\max}$	D_2
Archaea	Methanosarcina acetivorans (NC_003552.1)	0.00775	1.15
	Methanosarcina barkeri (NC_007355.1)	0.00675	1.20
Bacteria	Chitinophaga pinensis (NC_013132.1)	0.0095	1.50
	Sorangium cellulosum (NC_010162.1)	0.00925	1.35
Plants	Arabidopsis thaliana (NC_003071.7)	0.00725	1.25
	Oryza sativa (NC_008394.1)	0.00675	1.30
Animals	Pan troglodytes (NT_106996.1)	0.00705	1.30
	Mus musculus (NT_039340.7)	0.00715	1.31

5. Conclusion

In summary, we have investigated the DNA sequences of eight samples from archaea, bacteria, plants and animals. After denoising with continuous wavelet transformation, the skew profiles of these DNA sequences are analysed by phase space reconstruction method. Correlation dimension and the LLE indicate low dimensional chaotic behaviour in all these skew profiles. Furthermore, the skew profiles have similar attractor shape. We suggest that chaotic properties may be ubiquitous in the DNA sequences of all organisms.

References

[1] Song Y Z 2007 *Chin. Phys.* **16** 1918

[2] Song Y Z, Zhao G Z and Qi D L 2006 *Chin. Phys.* **15** 2266

[3] Liu C X 2007 *Chin. Phys.* **16** 3279

[4] Zhang Y H, Qi G Y, Liu W L and Yan Y 2006 *Acta Phys. Sin.* **55** 3307 (in Chinese)

[5] Xu Z, Liu C X and Yang T 2010 *Acta Phys. Sin.* **59** 131 (in Chinese)

[6] Wang F Z, Qi G Y, Chen Z Q and Yuan Z Z 2007 *Acta Phys. Sin.* **56** 3137 (in Chinese)

[7] Innocenti G, Morelli A, Genesio R and Torcini A 2007 *Chaos* **17** 043128

[8] Deng B and Hines G 2002 *Chaos* **12** 533

[9] Benincá E, Huisman J, Heerkloss R, Jöhnk K D, Branco P, Van Nes E H, Scheffer M and Ellner S P 2008 *Nature* **451** 822

[10] Stone L and He D 2007 *Journal of Theoretical Biology* **248** 382

[11] Elbert T, Ray W J, Kowalik Z J, Skinner J E, Graf K E and Birbaumer N 1994 *Physiological Reviews* **74** 1

[12] Robertson R and Combs A 1995 *Chaos Theory in Psychology and the Life Sciences* (Mahwah, NJ: Lawrence Erlbaum Associates)

[13] http://en.wikipedia.org/wiki/DNA

[14] Watson J D and Crick F H C 1953 *Nature* **171** 737

[15] Segal E, Fondufe-Mittendorf Y, Chen L Y, Thastrom A, Field Y, Moore I K, Wang J P Z and Widom J 2006 *Nature* **442** 772

[16] Yu Z G, Vo A, Gong Z M and Long S C 2002 *Chin. Phys. Letters* **13** 1313

[17] Redon R, Ishikawa S, Fitch K R, Feuk L, Perry G H, Andrews T D, Fiegler H, Shapero M H, Carson A R, Chen W, Cho E K, Dallaire S, Freeman J L, Gonzalez J R, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald J R, Marshall C R, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville M J, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad D F, Estivill X, Tyler-Smith C, Carter N P, Aburatani H, Lee C, Jones K W, Scherer S W and Hurles M E 2006 *Nature* **444** 444

[18] Nicolaya S, Brodie of Brodie E B, Touchonb M, d'Aubenton-Carafab Y, Thermesb C and Arneodoa A 2004 *Physica A* **342** 270

[19] Nicolaya S, Argoul F, Touchon M, d'Aubenton-Carafa Y, Thermes C and Arneodo1 A 2004 *Phys. Rev. Lett.* **93** 108101

[20] Lobry J R 1996 *Mol. Biol. Evol.* **13** 660

[21] Grigoriev A 1998 *Nucleic Acids Research* **26** 22860

[22] Gao F and Zhang C T 2008 *BMC Bioinformatics* **9** 79

[23] Takens F 1981 *Detecting Strange Attractors in Turbulence* (Berlin: Springer) p366

[24] Casdagli M, Eubank S, Farmer J D and Gibson J 1991 *Physica* D **51** 52

[25] Frazer A M and Swinney H L 1986 *Phys. Rev.* A **33** 1134

[26] Kennel M, Brown R and Abarbanel H 1992 *Phys. Rev.* A **45** 3403

[27] Grassberger P and Procaccia I 1983 *Physica* D **9** 189

[28] Schreiber T, Kantz H and Hegger R 1999 *Chaos* **9** 413