Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Scaling behavior of nucleotide cluster in DNA sequences*

CHENG Jun†1, TONG Zi-shuang[1], ZHANG Lin-xi[2]
([1]*Department of Physics, Jinhua College of Profession and Technology, Jinhua 321017, China*)
([2]*Department of Physics, Wenzhou University, Wenzhou 325027, China*)
†E-mail: jh_chengjun@163.com

Received June 27, 2006; revision accepted Feb. 10, 2007

Abstract: In this paper we study the scaling behavior of nucleotide cluster in 11 chromosomes of Encephalitozoon cuniculi Genome. The statistical distribution of nucleotide clusters for 11 chromosomes is characterized by the scaling behavior of $P(S) \propto e^{-\alpha S}$, where S represents nucleotide cluster size. The cluster-size distribution $P(S_1+S_2)$ with the total size of sequential C-G cluster and A-T cluster S_1+S_2 were also studied. $P(S_1+S_2)$ follows exponential decay. There does not exist the case of large C-G cluster following large A-T cluster or large A-T cluster following large C-G cluster. We also discuss the relatively random walk length function $L(n)$ and the local compositional complexity of nucleotide sequences based on a new model. These investigations may provide some insight into nucleotide cluster of DNA sequence.

Key words: Nucleotide cluster, DNA sequences, Scaling behavior
doi:10.1631/jzus.2007.B0359 **Document code:** A **CLC number:** Q615

INTRODUCTION

The sequence dependency of DNA sequences is important for protein-DNA recognition (Gromiha, 2005; Gromiha *et al*., 2004; Olson *et al*., 2004). Statistical analysis of DNA chain bending profiles for complete genome sequences revealed that long-range correlations in the 10~200 bp range are the signatures of the nucleosomal in structure and that over larger distances (>200 bp) are likely to play a role in the hierarchical packaging of DNA. To which extent sequence-dependent DNA mechanical properties help to regulate the structure and dynamics of chromatin is an issue of fundamental importance. We recently explored long-range correlations in the base composition of DNA and adopted a new method to study the scaling behaviors of C-G clusters in different organism chromosomes (Cheng and Zhang, 2005a; 2005b). The term cluster was first defined by Provata and Almirantis (1997; 2002). They did not differentiate between adenine and guanine, both considered as Pu, cytosine and thymine, both considered as Py, and they defined the Pu-cluster as an ensemble of consecutive Pu bound by at least one Py on the left and right respectively, equivalent to Py-clusters. Here the Pu-cluster and Py-cluster are both called "nucleotide cluster", which may be considered to constitute an entire DNA sequence. Some statistical properties of nucleotide clusters are linked to a higher level of organization, and statistical dynamics of clustering in the genome structure were investigated (Provata and Almirantis, 2002; Zhang and Jiang, 2004; Sun *et al*., 2004; Chen *et al*., 2004; Gromiha *et al*., 1997; Hogan and Austin, 1987). Here we change the definition of the cluster (Chen *et al*., 2004). It is well known that the nucleotide A is paired with T, and C is paired with G. The CG content of DNA sequences was used to gain understanding of several properties of DNA, such as curvature, bending, etc. (Harrington and Winicov, 1994; Sugiarto *et al*., 2006; Vaillant *et al*., 2003; Arnéodo, 1998). In particular, in possible relation to the isochore structure of the human genome, it

* Project supported by the National Natural Science Foundation of China (No. 20574052), Program for New Century Excellent Talents in University, and the Natural Science Foundation of Zhejiang Province (Nos. R404047 and Y405011), China

had been clearly shown that the long-range correlation properties of human DNA sequences are dependent upon their CG content (Vaillant *et al*., 2003; Arnéodo, 1998). If we can know the statistical properties and size distributions of sequential CG and AT clusters, we can then understand the DNA sequences in more detail.

In this work, we investigate the size distribution of C-G clusters and A-T clusters in the complete nucleotide sequences for Encephalitozoon cuniculi Genome which has 11 chromosomes. We picked this organism as example since the size of the genome is manageable and evidence for a power law is particularly impressive in this species. Our aim is to investigate the scaling behaviors of nucleotide clusters in DNA sequences.

METHOD OF CALCULATION

In this paper we change the definition of the Py- and Pu-cluster proposed by Provata and Almirantis (1997; 2002), and adopt hydrogen bond energy rule (Poland, 2004; Azbel *et al*., 1982; Azbel, 1973). We refer 1's to strongly bonded pairs (C or G) and 0's to weakly bonded pairs (A or T) and define the average number $\overline{n}$ of C-G clusters in each block (Cheng and Zhang, 2005a; 2005b; Zhang and Chen, 2005) as follows:

$$\overline{n} = \sum S_i \Big/ \sum i = \sum_{i=1}^{N} S_i \Big/ N, \tag{1}$$

here N refers to the number of CG clusters, and S_i the size of CG clusters of ith cluster in the consecutive, non-overlapping block, where the size of non-overlapping block is m. If the value of m is small to a certain extent, the average number $\overline{n}$ can show all characteristics of a whole sequence per block including the C-G content and A-T content. For example, there are two blocks with the sequence: 0101001111 and 1010101011 respectively. We can obtain the average size of CG cluster $\overline{n}(m)$=6/3 and $\overline{n}(m)$=6/5 respectively, and get to know that the distribution of CG cluster is different in those two sequences although the number of strong pairs is the same. Therefore, the total number of CG cluster n in the consecutive, non-overlapping block is unilateral because it neglects the array sequence and covers the feature of different sequences with the same number of CG cluster n, and $\overline{n}$ can include more information of the sequence perfectly (Zhang and Jiang, 2004; Poland, 2004).

First, we collect statistics for boxes in the complete nucleotide sequence containing n C-G units, thus giving the distribution function for the C-G content in m-blocks. The function $P(S)$ which represents the fraction of clusters corresponding to a certain cluster size S is called cluster-size distribution. We define the cluster-size distribution of CG (or AT) clusters $P(S)$ as

$$P(S) = \frac{N(S)}{N(S=1)}, \tag{2}$$

where $N(S$=1) is the number of clusters with cluster size S=1, and it is apparently the largest one in the DNA sequence. Therefore, the cluster-size distribution of CG (or AT) clusters $P(S)$ is always smaller than 1.0.

Second, we can gain some insight into the correlations in DNA sequences by interpreting the sequence as a random walk (Mandelbrot, 1982; Feder, 1989), and make the following assignments to each base in the sequences:

$$\alpha_i = -1, \text{ if C or G; } \alpha_i = +1, \text{ if A or T.} \tag{3}$$

The two equations can be regarded as a random walk where α_i=+1 indicates a step to the right and α_i=−1 indicates a step to the left. The distance of the walk from the origin after n steps (n bases) is then defined as

$$L(n)' = \sum_{i=1}^{n} \alpha_i. \tag{4}$$

On the average, this function will have the value

$$<L(n)'> = n\Delta f, \tag{5}$$

where

$$\Delta f = f_{\mathrm{at}} - f_{\mathrm{cg}}. \tag{6}$$

It is convenient to define the length of the walk compared to the average walk given above, i.e.,

$$L(n)=\sum_{i=1}^{n}\alpha_i - n\Delta f. \quad (7)$$

The average value of this function $L(n)$ is zero, i.e., $<L(n)>=0$.

Third, we study the local compositional complexity of nucleotide sequences. Given a window of length L, we define the local complexity state by the vector for the nucleotide sequence as $\boldsymbol{n}=[n_1, n_2, n_3, n_4]$, such that $n_1 \geq n_2 \geq n_3 \geq n_4$ are all nonnegative integers that satisfy $n_1+n_2+n_3+n_4=L$. The number n_1 represents the number of occurrences of the most frequent nucleotide α_1, n_2 the number of the second most frequent nucleotide α_2 and so on. Here $\alpha_i \in$ {A, C, G, T}. The multinomial coefficient is given by

$$\Omega=\frac{L!}{n_1!\,n_2!\,n_3!\,n_4!}, \quad (8)$$

and it is the total number of distinguishable arrangements of n_1 outcomes of α_1, n_2 outcomes of α_2, and so forth. Over a sequence window length L, we adopt the measure of complexity (Salamon and Konopka, 1992; Salamon *et al*., 1993; Wootton and Federhen, 1993),

$$K=(\ln\Omega)/L. \quad (9)$$

At last we introduce a new way to study the local compositional complexity of nucleotide sequences. Given a window of length L, we define the local complexity state by the vector for the nucleotide sequence as $\boldsymbol{n}=[n_1, n_2, n_3, \ldots, n_{16}]$, $\sum_{i=1}^{16} n_i = L-1$. The number n_1 represents the number of occurrences of the most frequent nucleotide $\alpha_1\beta_1$, n_2 the number of the second most frequent nucleotide $\alpha_2\beta_2$ and so on, where $\alpha_i\beta_i \in$ {AA, AC, AG, AT, …, TG, TT}. The multinomial coefficient is given by

$$\Omega'=(L-1)!\Big/\prod_{i=1}^{16} n_i!, \quad (10)$$

and it is the total number of distinguishable arrangements of n_1 outcomes of $\alpha_1\beta_1$, n_2 outcomes of $\alpha_2\beta_2$, and so forth, over a sequence window length L. Then we adopt the measure of complexity,

$$K'=(\ln\Omega)/(L-1). \quad (11)$$

The species we have chosen to be treated as examples are Encephalitozoon cuniculi Genome, which has 11 chromosomes with the range of lengths from 194439 bp to 267509 bp. The complete genome sequences data used in this paper were all taken from the www at the National Center for Biotechnology Information (USA) (http://www.ncbi.nih.gov/genbank/genomes/) in GenBank format.

RESULTS AND DISCUSSION

Cluster-size distribution $P(S)$

The cluster-size distribution $P(S)$ of A-T cluster for Encephalitozoon cuniculi Genome serving cluster size S is shown in Fig.1a. In the upper right of Fig.1a, the straight line represents an exact power law with the exponents as 0.647 and the correlation coefficient as 0.991 for Chromosome I sequence. Similar results are also obtained for other chromosomes from Fig.1a. In the meantime, the cluster-size distribution $P(S)$ of C-G cluster for Encephalitozoon cuniculi Genome serving cluster size S is given in Fig.1b. In the upper right of Fig.1b, the straight line represents an exact power law with the exponents as 0.812 and the correlation coefficient as 0.988 for Chromosome I sequence. Similar results are obtained from Fig.1. Apparently, we can find that the results conform to the power law from Fig.1:

$$P(S) \propto \mathrm{e}^{-\alpha S}. \quad (12)$$

It can be found that the value of the scaling exponent α for C-G and A-T clusters ranges from 0.771 to 0.887 ($\alpha_{\text{C-G}}$) and 0.470 to 0.662 ($\alpha_{\text{A-T}}$) respectively, and that the average of the scaling exponent $\overline{\alpha}_{\text{C-G}}$ for CG cluster (0.825) is greater than that for AT cluster (0.582). The correlation coefficient for those scaling exponents ranges from 0.949 to 0.999 with most of them being double numbers 9 after the radix point. It is nearly equal to 1, and this fit is perfectly good.

In Fig.1, we find that the maximum value of A-T and C-G cluster size is different for these chromosomes. The maximum value of A-T cluster is 27 (Chromosome XI). However, the maximum value of C-G cluster of is only 16 (Chromosome V).

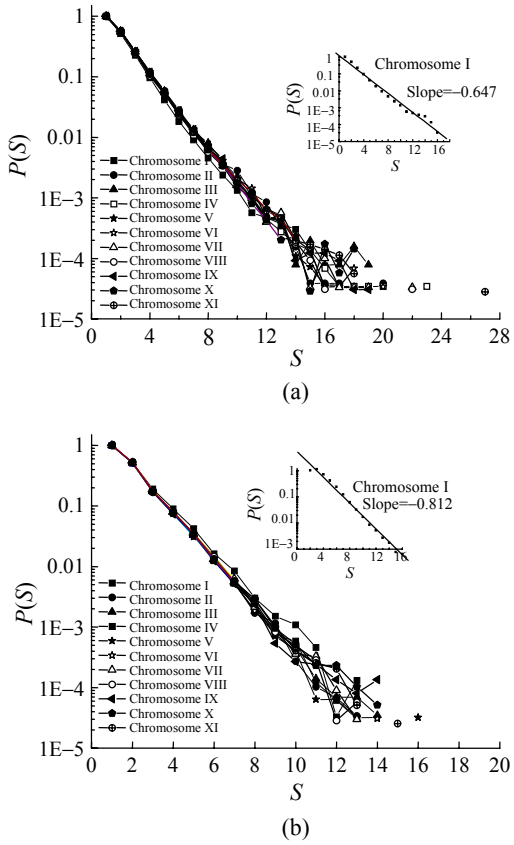

Fig.1 (a) The cluster-size distribution $P(S)$ of A-T cluster as a function of cluster size S for Encephalitozoon cuniculi Genome. The straight line in upper right represents an exact power law with slope=−0.647 for Chromosome I; (b) The cluster-size distribution $P(S)$ of C-G cluster as a function of cluster size S for Encephalitozoon cuniculi Genome. The straight line in the upper right represents an exact power law with slope=−0.812 for Chromosome I

Cluster-size distribution $P(S_1+S_2)$ with the total size of sequential C-G and A-T S_1+S_2

In DNA sequence, the C-G and A-T clusters appear alternately. Here we discuss the cluster-size distribution $P(S_1+S_2)$ with the total size of sequential C-G and A-T S_1+S_2. In order to explain our calculation method in more details, we express the sequence 000101111110101011011111110 0… as $B_3A_1B_1A_6B_1A_1B_1A_1B_1A_2B_1A_7B_2$… Here A_i represents C-G clusters of size i and B_j represents A-T clusters of size j. For example, B_3 represents A-T clusters of size 3. Here S_1, S_2 stand for the size of sequential C-G cluster and A-T cluster respectively. The cluster-size distribution $P(S_1+S_2)$ versus S_1+S_2 for Encephalitozoon cuniculi Genome is shown in Fig.2. In the upper right of Fig.2, the straight line represents an exact power law with the exponents as 0.566 and the correlation coefficient as 0.996 for Chromosome I. Similar results are obtained for other chromosomes in Fig.2. Apparently, we can find that the results conform to the power law in Fig.2:

$$P(S_1+S_2) \propto \mathrm{e}^{-\alpha_{S_1+S_2}S}. \tag{13}$$

The scaling exponent $\alpha_{S_1+S_2}$ ranges from 0.494 to 0.596, and the correlation coefficient for those scaling exponents ranges from 0.971 to 0.996 with the average 0.989.

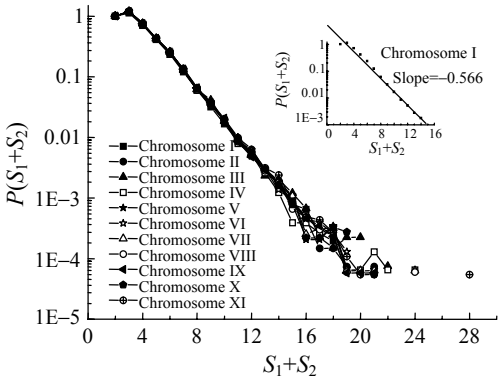

Fig.2 The cluster-size distribution $P(S_1+S_2)$ versus the total size of sequential C-G and A-T S_1+S_2 for Encephalitozoon cuniculi Genome. The straight line in the upper right represents an exact power law with slope=−0.566 for Chromosome I sequence

In Fig.2, we find that the maximum value of S_1+S_2 is different for these chromosomes. The maximum value of S_1+S_2 for these chromosomes is only 28 (Chromosome XI). This means that there does not exist the case of large C-G cluster following large A-T cluster or large A-T cluster following large C-G cluster. At last, the values of the scaling exponents for 11 chromosomes of Encephalitozoon cuniculi Genome are all shown in Fig.3. The solid squares, dots, and triangles refer to the α_{CG}, α_{CAT}, and $\alpha_{S_1+S_2}$, respectively. The dashed straight lines give the average of them respectively.

Fractional Brownian walk

In Fig.4, the upper solid curve gives the random walk function $L(n)'$ defined in Eq.(4) for Chromo-

some I of the Encephalitozoon cuniculi Genome. The upper dashed straight line gives the locus of the average length of the walk as a function of the number of steps given in Eq.(5). The lower solid curves give the random walk as a function of the average length of the walk as defined in Eq.(7) and for a random sequence on the same scale respectively. The average of this function is zero as indicated by the lower solid straight line.

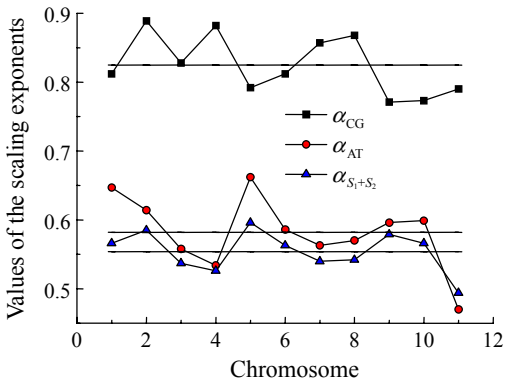

Fig.3 The values of the scaling exponents for 11 chromosomes of Encephalitozoon cuniculi Genome. The dashed straight lines give the average of them respectively

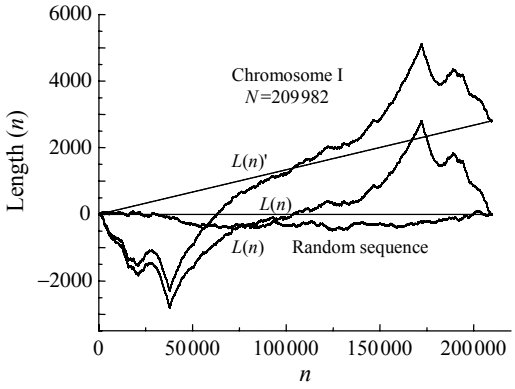

Fig.4 The upper solid curve gives the random walk function $L(n)'$ defined in Eq.(4) for the Chromosome I of the Encephalitozoon cuniculi Genome. The upper dashed straight line gives the locus of the average length of the walk as a function of the number of steps as given in Eq.(5). The lower solid curves give the random walk function relative to the average length of the walk as difined in Eq.(7) and for a random sequence on the same scale respectively. The average of this function is zero as indicated by the lower solid straight line

In the meantime, we study the relative random walk length function $L(n)$ defined in Eq.(7) for the other 10 chromosomes of Encephalitozoon cuniculi Genome. We find that the small amplitude curves give $L(n)$ a random sequence for comparison in each case. Apparently, $L(n)$ is larger for a given actual sequence as compared to a corresponding random sequence, which means the information for the DNA of chromosomes is greatly different in a corresponding random sequence. This investigation can provide some insight into nucleotide clusters properties, and help us understand the long-range correlation and other properties of DNA sequences.

DNA sequence complexity analysis

Fig.5 is about complexity versus bp position (L=70 bp). The upper solid curve gives the local compositional complexity of nucleotide sequences K' defined in Eq.(11) and the lower solid curve gives K difined in Eq.(9) for Chromosome I of Encephalitozoon cuniculi Genome. We find that the value of K' is almost twice that of K. The changes of K' versus bp position are extremely similar to that of K. However, the amplitude modulation of K' is twice larger than that of K. Evidently, K' can demonstrate better the information involved in DNA, and the value is much bigger compared to K. Meanwhile, we study other chromosomes of the Encephalitozoon cuniculi Genome and find the same conclusion. Determining the cause of such behavior is a difficult task to accomplish, which is due to many factors that need to be taken into account. Further research is required to develop comparisons and search methods appropriate for the local compositional complexity of nucleotide sequences.

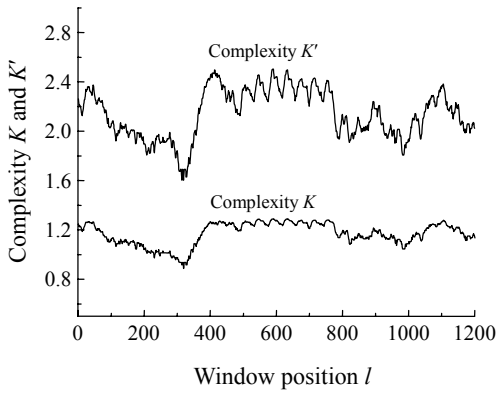

Fig.5 Complexity versus window position. A sliding-window starts at the beginning of the DNA sequence and is computed along its position l (L=70 bp)

References

Arnéodo, A., 1998. Nucleotide composition effects on the long-range correlation in human genes. *Eur. Phys. J. B*, **1**(2):259-263. [doi:10.1007/s100510050180]

Azbel, M., 1973. Random two-component one-dimensional Ising model for heteropolymer melting. *Phys. Rev. Lett.*, **31**(9):589-593. [doi:10.1103/PhysRevLett.31.589]

Azbel, M.Y., Kantor, Y., Verkh, L., Vilenkin, A., 1982. Statistical analysis of DNA sequences. *Biopolymers*, **21**(8):1687-1690. [doi:10.1002/bip.360210816]

Chen, J., Zhang, L.X., Cheng, J., 2004. Elastic behavior of adsorbed polymer chains. *J. Chem. Phys.*, **121**(22): 11481-11488. [doi:10.1063/1.1818673]

Cheng, J., Zhang, L.X., 2005a. Scaling behaviors of CG clusters for chromosomes. *Chaos, Solitons & Fractals*, **25**(2): 339-346. [doi:10.1016/j.chaos.2004.12.004]

Cheng, J., Zhang, L.X., 2005b. Statistical properties of nucleotide clusters in DNA sequences. *Journal of Zhejiang University SCIENCE*, **6B**(5):408-412. [doi:10.1631/jzus.2005.B0408]

Feder, J., 1989. Fractals. Plenum Press, New York.

Gromiha, M.M., 2005. Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnology*, **117**(2):137-145. [doi:10.1016/j.jbiotec.2004.12.016]

Gromiha, M.M., Munteanu, M.G., Simon, I., Pongor, S., 1997. The role of DNA bending in Cro protein-DNA interactions. *Biophys. Chem.*, **69**(2-3):153-160. [doi:10.1016/S0301-4622(97)00088-4]

Gromiha, M.M., Siebers, J.G., Selvaraj, S., Kono, H., Sarai, A., 2004. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **337**(2): 285-294. [doi:10.1016/j.jmb.2004.01.033]

Harrington, R.E., Winicov, I., 1994. New concepts in protein-DNA recognition: sequence-directed DNA bending and flexibility. *Prog. Nucleic. Acid. Res. Mol. Biol.*, **47**: 195-270.

Hogan, M.E., Austin, R.H., 1987. Importance of DNA stiffness in protein-DNA binding specificity. *Nature*, **329**(6136):263-266. [doi:10.1038/329263a0]

Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. W.H. Freeman and Company, New York.

Olson, W.K., Swigon, D., Coleman, B.D., 2004. Implications of the dependence of the elastic properties of DNA on nucleotide sequence. *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences*, **362**(1820):1403-1422. [doi:10.1098/rsta.2004.1380]

Poland, D., 2004. The persistence exponent of DNA. *Biophys. Chem.*, **110**(1-2):59-72. [doi:10.1016/j.bpc.2004.01.003]

Provata, A., Almirantis, Y., 1997. Scaling properties of coding and noncoding DNA sequences. *Physica A Statistical and Theoretical Physics*, **247**(1-4):482-487. [doi:10.1016/S0378-4371(97)00424-X]

Provata, A., Almirantis, Y., 2002. Statistical dynamics of DNA clustering in the genome structure. *J. Stat. Phys.*, **106**(1/2):23-56. [doi:10.1023/A:1013115911328]

Salamon, P., Konopka, A.K., 1992. A maximum entropy principle for the distribution of local complexity in naturally occurring nucleotide sequences. *Comput. Chem.*, **16**(2):117-124. [doi:10.1016/0097-8485(92)80038-2]

Salamon, P., Wooten, J.C., Konopka, A.K., Hansen, L.K., 1993. On the robustness of maximum entropy relationships for complexity distributions of nucleotide sequences. *Comput. Chem.*, **17**(2):135-148. [doi:10.1016/0097-8485(93)85005-W]

Sugiarto, R., Han, L.Y., Wang, J.S., Chen, Y.Z., 2006. Super paramagnetic clustering of DNA sequences. *J. Biol. Phys.*, **32**(1):11-25. [doi:10.1007/s10867-006-2120-0]

Sun, T.T., Zhang, L.X., Chen, J., Jiang, Z.T., 2004. Statistical properties and fractals of nucleotide clusters in DNA sequences. *Chaos, Solitons & Fractals*, **20**(5):1075-1084. [doi:10.1016/j.chaos.2003.09.012]

Vaillant, C., Audit, B., Thermes, C., Arnéodo, A., 2003. Influence of the sequence on elastic properties of long DNA chains. *Phys. Rev. E*, **67**(3):032901-032904. [doi:10.1103/PhysRevE.67.032901]

Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**(2):149-163. [doi:10.1016/0097-8485(93)85006-X]

Zhang, L.X., Jiang, Z.T., 2004. Long-range correlations in DNA sequences using 2D DNA walk based on pairs of sequential nucleotides. *Chaos, Solitons & Fractals*, **22**(4): 947-955. [doi:10.1016/j.chaos.2004.03.012]

Zhang, L.X., Chen, J., 2005. Scaling behaviors of CG cluster for coding and non-coding DNA sequence. *Chaos, Solitons & Fractals*, **24**(1):115-123. [doi:10.1016/j.chaos.2004.07.013]