Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Statistical properties of nucleotide clusters in DNA sequences*

CHENG Jun (成 军)†1, ZHANG Lin-xi (章林溪)2

(1Department of Physics, Jinhua University, Jinhua 321017, China)

(2Department of Physics, Zhejiang University, Hangzhou 310027, China)

†E-mail: Jh_Chengjun@163.com

Received Oct. 29, 2004; revision accepted Jan. 27, 2005

Abstract: Using the complete genome of *Plasmodium falciparum* 3D7 which has 14 chromosomes as an example, we have examined the distribution functions for the amount of C or G and A or T consecutively and non-overlapping blocks of m bases in this system. The function $P(S)$ about the number of the consecutive C-G or A-T content cluster conforms to the relation $P(S) \propto e^{-\alpha s}$; values of the scaling exponent α_{CG} are much larger than α_{AT}; and α_{AT} of 14 chromosomes are hardly changed, whereas α_{CG} of 14 chromosomes have a number of fluctuations. We found maximum value of A-T cluster size is much larger than C-G, which implies the existence of large A-T cluster. Our study of the width function $\xi(m)$ of cluster C-G content showed that follows good power law $\xi(m) \propto m^{-\gamma}$. The average $\overline{\gamma}$ for 14 chromosomes is 0.931. These investigations provide some insight into the nucleotide clusters of DNA sequences, and help us understand other properties of DNA sequences.

Key words: DNA sequence, *Plasmodium falciparum* 3D7, Nucleotide clusters, Power law
doi:10.1631/jzus.2005.B0408 **Document code:** A **CLC number:** Q615

INTRODUCTION

The genetic information on organisms is stored in DNA sequences, represented by a string of just four letters, each corresponding to a definite type of nucleotides: adenine A (purines), guanine G (purines), cytosine C (pyrimidines), and thymine T (pyrimidines) (Albert *et al*., 1994). These letters can form different combinations. Each DNA sequence is packed in a chromosome, varying in length from 10^5 base pairs (bp) in yeast to 10^9 bp in human. During the past few years, many biologists, chemists and physicists have researched intensively the intriguing statistical behaviour of DNA sequences. The most common research fields include mutual information functions (Herzel *et al*., 1998; Herzel and Grobe, 1997; Li, 1992), the power spectra (Li, 1997; Voss, 1992; Viswanathan *et al*., 1998; Li and Kaneko, 1992), auto-correlation analysis (Herzel *et al*., 1998; Azbel, 1995; De Sousa Vieira, 1999), detrended fluctuation analysis (Sun *et al*., 2004; Peng *et al*., 1994; Buldyrev *et al*., 1995). However, some statistical aspects of the DNA sequences are still obscure.

The clustering of similar nucleotides can be clarified by studying the properties of the cluster size distribution on various real DNA sequences, ranging from viral to higher eukaryotic sequences. The term "cluster of similar (homologous) nucleotides" means a string in a sequence containing only one type of nucleotides, e.g., only A's or only C's, etc. The term "cluster" was defined by Provata and Almirantis (1997; 2002). They did not differentiate between adenine and guanine, both considered by them as Pu; and between cytosine and thymine, both considered by them as Py. They defined the Pu-cluster as an ensemble of consecutive Pus bound by at least one Py respectively on the left and on the right, and to be equivalent to Py-clusters. Here the Pu-cluster and Py-cluster are both called "nucleotide cluster". Therefore, an entire DNA sequence may be consid-

* Project supported by the National Natural Science Foundation of China (Nos. 20174036, 20274040), and the Natural Science Foundation of Zhejiang Province (Nos. R404047, 10102), China

ered as a collection of nucleotide clusters. Some statistical properties of nucleotide clusters are linked to a higher level of organization. The statistical dynamics of clustering in the genome structure were also investigated (Provata and Almirantis, 2002). In fact, we found that Pu and Py clusters appear alternately in those clusters. If we can know the statistical properties and size distributions of sequential Pu-clusters and Py-clusters, we will have more detailed understanding of the DNA sequences.

In this work, we investigated the size distribution of purine and pyrimidine clusters in the complete nucleotide sequences of *Plasmodium falciparum* 3D7 which has 14 chromosomes. We found that the distribution values for every chromosome are similar and obtained some information on the statistical properties of nucleotide clusters in DNA sequences.

METHOD OF CALCULATION

Here we modify the definition of the Py-cluster and Pu-cluster proposed by Provata and Almirantis (1997; 2002), using hydrogen bond energy rule (Azbel, 1973; Azbel *et al.*, 1982). We regard 1's as strongly bonded pairs (C or G) and 0's as weakly bonded pairs (A or T). We define the function of the size S, the number $N(S)$ of continuous C-G cluster and the average number $\overline{n}$ of C-G units in each block (Chen and zhang, 2005; Sun *et al*., 2004) as follows:

$$\overline{n}=\frac{n_{\text{C-G}}}{N_{\text{C-G}}} \tag{1}$$

where $n_{\text{C–G}}$ represents the number of 1's (C or G) per block and $N_{\text{C–G}}$ is the total number of C-G units (consecutive C and G are treated as one unit; Fig.1 for illustration) in the block.

The division of the genome into consecutive non-overlapping blocks containing m bases is illustrated in Fig.1. We selected the snippet with 45 nucleotide letters as the example to show this change. The upper row illustrates the assignment of bases to blocks in the case of m=10. In the lower row, we indicate how we assign 0's to A or T bases and 1's to C or G bases and count the number of 1's (C-G units) in each box (Sun *et al*., 2004). We calculate $\overline{n}$ using Eq.(1). If the value of m is small to a certain extent, the average number $\overline{n}$ can show all the characteristics of the whole sequence per block including the C-G content and A-T content. The parameter number n is unilateral as it neglects the array sequence and covers the feature of different sequences with the same value m, while $\overline{n}$ can include more information on the sequence perfectly.

Then we collect the statistics of the boxes in the complete nucleotide sequence which contain n C-G units or n A-T units, thus obtaining the distribution function for the C-G content or the A-T content in m-blocks. The function $P(S)$ which represents the fraction of clusters corresponding to a certain cluster size S is called the "cluster-size distribution". To simplify the parameter $N(S)$, we rewrite the function as

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

where $N(S=1)$ is the number of units on cluster size S=1, which is apparently the largest in one DNA sequence. So the values of this parameter are always smaller than 1.

On the other hand, we can get the distribution function for a block of m bases if we assume independent units. In terms of the definition of the average number $\overline{n}$, we obtain the following relations giving the first two moments of the distribution

Blocksize: m=10

AATCACCTAG	AATTCGCCTA	TGCCCGGCAA	GCCACTCTCG	ACACC---------
0 0 0 1 0 1 1 0 0 1	0 0 0 0 0 1 1 1 1 0 0	0 1 1 1 1 1 1 1 0 0	1 1 1 0 1 0 1 0 1 1	0 1 0 1 1 ---------
n=4	n=4	n=7	n=7	
$\overline{n}=\frac{4}{3}$	$\overline{n}=\frac{4}{1}$	$\overline{n}=\frac{7}{1}$	$\overline{n}=\frac{7}{4}$	

Fig.1 Illustration of the use of blocks to obtain C-G distributions. The upper row illustrates the division of the genome into consecutive non-overlapping blocks each containing m bases. For the example shown m=10. The lower row indicates the translation of the base composition into 1's (C or G) and 0's (A or T). $\overline{n}$ and n are the average number and number of the 1's (C-G units) with m=10 respectively

$$\begin{cases} \mu_2(m)=<\overline{n}^2> \\ \mu_1(m)=<\overline{n}> \end{cases} \quad (3)$$

We take as a standard measure of the breadth of the distribution the root-mean-square width (Poland, 2004)

$$\sigma_m = \sqrt{\mu_2(m) - \mu_1(m)^2} \quad (4)$$

This is a useful method for defining a parameter in terms of the differences between the average of the square and the square of the average.

We are interested in how the actual C-G cluster distributions obtained from 14 chromosomes of *Plasmodium falciparum* 3D7 genome deviate from the dependence on m given in Eq.(4). To this end, we define the following function to investigate the relations between the width of the empirical distribution σ_m and the m-dependence.

$$\xi(m) = \sigma_m / \sqrt{m} \quad (5)$$

From Eq.(5), we can note that if the distribution of the C-G content in m-blocks is random, $\xi(m)$ should be equal to a constant value or close to a constant value as the length of the random sequence is increased while the slope of the line fitting Eq.(5) should equal or approach 1 too. To test this conclusion we computed the random sequences 1000000 bp and get the slope 0.999 and the correlation coefficient 0.999 that we presupposed (Chen and Zhang, 2005).

The species we chose as example was the *Plasmodium falciparum* 3D7, which has 14 chromosomes with length of 643292 bp to 13515526 bp. The complete genome sequences data we use 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. We chose this organism as an example since the size of the genome is manageable and the evidence for power law is particularly impressive in this species.

RESULTS AND DISCUSSIONS

In terms of Eq.(2), we discuss the cluster-size distribution $P(S)$ of the *Plasmodium falciparum* 3D7 genome. The A-T clusters distribution of chromosome XII is shown legibly in Fig.2. In this single-logarithm, we can link these squares to a straight line with slope of −0.087. The correlation coefficient is 0.972. We found that maximum value of A-T Cluster size is 100. We can note from the graph that the result complies with the power law:

$$P(S) \propto e^{-\alpha S} \quad (6)$$

where α is the absolute value of the line slope. The other 13 chromosomes were also similar to chromosome XII in the statistical property of the A-T cluster. The results are shown in Table 1. The values of α_{AT} range from 0.085 to 0.090 and the average is 0.087. Apparently, α_{AT} of 14 chromosomes are hardly changed. The values of the correlation coefficient are 0.960 to 0.974. The maximum values of A-T Cluster size are 98 to 100. This means that there exists large A-T cluster in the DNA sequences. Likewise, the C-G clusters distribution of chromosome XII is shown in Fig.3 legibly. The straight line represents a power law with exponent α_{CG}=1.114 and correlation coefficient 0.982. The other 13 chromosomes are also similar to chromosome XII in the statistical property of the C-G cluster. The results are also shown in Table 1. The values of α_{CG} are 0.877 to 1.287 and the average is 1.103. The α_{CG} values of 14 chromosomes have a number of fluctuations. The values of correlation coefficient are 0.961 to 0.996. The maximum values of C-G Cluster size are 10 to 15, which means that large C-G cluster does not exist in sequences different from those of the A-T cluster.

The width function $\xi(m)$ is defined in Eq.(5). Fig.4 shows the results of $\xi(m)$ based Eq.(5) with

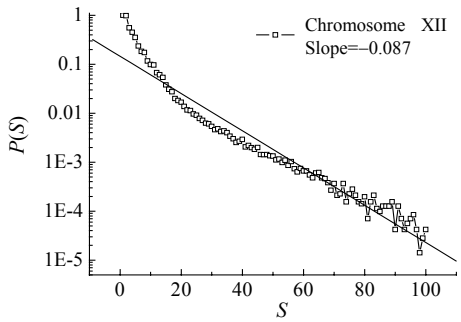

Fig.2 The cluster-size distribution $P(S)$ of A-T cluster as a function of cluster size for *Plasmodium falciparum* 3D7 Chromosome XII sequences

Table 1 Values of the scaling exponents α_{CG}, α_{AT}, γ and maximum cluster for 14 chromosomes of *Plasmodium falciparum* 3D7

No.	Sequence	$P(S)\propto e^{-\alpha s}$					$\xi(m)\propto m^{-\gamma}$		Maximum value of cluster	
		Length	α_{AT}	Correlation coefficient	α_{CG}	Correlation coefficient	γ	Correlation coefficient	A-T cluster	C-G cluster
1	Chromosome I	643292	0.085	0.960	1.011	0.968	0.918	0.999	99	10
2	Chromosome II	947102	0.087	0.969	1.158	0.984	0.920	0.998	100	11
3	Chromosome III	1060087	0.087	0.968	1.121	0.985	0.938	0.998	100	12
4	Chromosome IV	1204112	0.087	0.968	0.974	0.978	0.856	0.998	99	13
5	Chromosome V	1343552	0.087	0.969	1.296	0.996	0.948	0.998	100	14
6	Chromosome VI	1378756	0.090	0.969	1.094	0.989	0.927	0.999	98	15
7	Chromosome VII	13515526	0.087	0.969	1.045	0.973	0.917	0.998	100	12
8	Chromosome VIII	1325595	0.085	0.967	1.234	0.994	0.907	0.999	100	10
9	Chromosome IX	1541723	0.087	0.972	1.119	0.989	0.951	0.998	100	12
10	Chromosome X	1694445	0.090	0.974	1.287	0.995	0.948	0.998	99	10
11	Chromosome XI	2035250	0.090	0.974	1.002	0.967	0.949	0.998	100	14
12	Chromosome XII	2271477	0.087	0.972	1.114	0.982	0.945	0.998	100	12
13	Chromosome XIII	2732359	0.090	0.974	0.877	0.961	0.932	0.999	100	15
14	Chromosome XIV	3291006	0.087	0.974	1.094	0.964	0.976	0.999	100	14
	Average		0.087		1.103		0.931		99.6	12.4

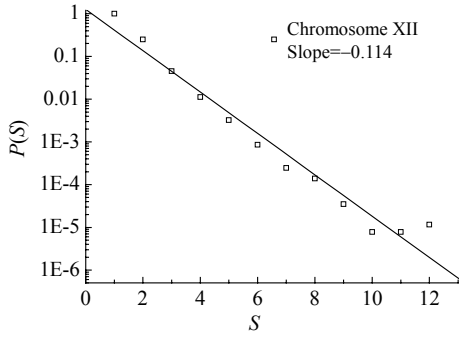

Fig.3 The cluster-size distribution $P(S)$ of C-G cluster as a function of cluster size for *Plasmodium falciparum* 3D7 Chromosome XII sequences

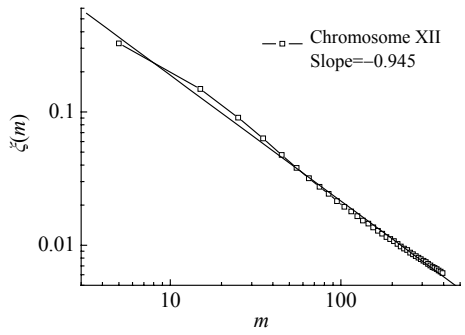

Fig.4 The width function $\xi(m)$ as a function of m for *Plasmodium falciparum* 3D7 Chromosome XII sequences

m=400 on chromosome XII. The straight line represents an exact power law with exponent equal to 0.945 correlation coefficient of 0.998. We can get the direct conclusion when the results are expressed in natural logarithms for both coordinates. $\xi(m)$ in Fig.4 strongly suggests a power-law behavior.

$$\xi(m) \propto m^{-\gamma} \tag{7}$$

Here γ is the absolute value of the slope of the fit line in Fig.4. This linear fit agrees nicely with the results obtained by Poland (2004).

All of the other 13 chromosomes exhibit power-law behavior for the width of the C-G distributions similar to that shown in Fig.4. The values of γ are 0.856 to 0.976 and the average is 0.931. The γ values of 14 chromosomes have a number of fluctuations. The correlation coefficient values are from 0.998 to 0.999. At last, the values of the scaling exponents for 14 chromosomes of *Plasmodium falciparum* 3D7 are shown in Fig.5. The solid dots, squares and pentagons refer to the α_{AT}, α_{CG} and γ respectively.

From the above discussion, we conclude that

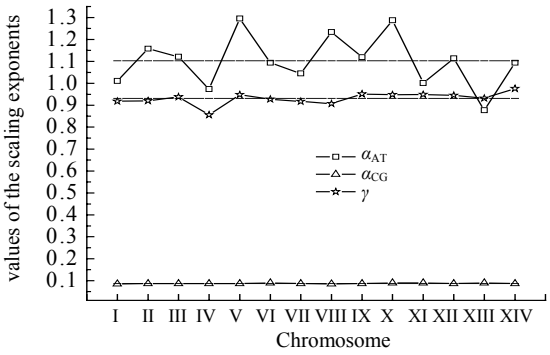

Fig.5 The values of the scaling exponents for 14 chromosomes of *Plasmodium falciparum* 3D7

large C-G and A-T clusters are formed as a result of the correlations in the unabridged DNA sequences. In this process the cluster-size distribution $P(S)$ follows exponential decay law and the widths of the C-G cluster distribution $\xi(m)$ follow a power law. In addition to the behaviour of the genome for the *Plasmodium falciparum* 3D described here, we studied the genomes of yeast, nematode, erenothecium gossypii and human (Cheng and Zhang, 2005). The question of course arises as to whether the occurrence of a power law with a characteristic exponent γ describing the width of the C-G distributions and α_{AT} or α_{CG} describing the cluster-size distributions is a general feature of the DNA of all organisms.

Our result is straightforward, but neither exhaustively nor completely resolve the problem of the statistical properties of nucleotide clusters in DNA sequences.

References

Albert, B., Bray, D., Lewis, J., Raff, M., Robert, K., Watson, J.D., 1994. Molecular Biology of the Cell. Garland Publishing, New York.

Azbel, M., 1973. Random two-component one-dimensional ising model for heteropolymer melting. *Physical Review Letter*, **31**:589-593.

Azbel, M., 1995. Universality in a DNA statistical structure. *Physical Review Letters*, **75**:168-171.

Azbel, M., Kantor, Y., Verkh, L., Vilenkin, A., 1982. Statistical analysis of DNA sequences. *Biopolymers*, **21**:1687-1690.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simons, M., Stanley, H.E., 1995. Long-range correlation properties of coding and noncoding DNA Sequences: GenBank analysis. *Physical Review E*, **51**:5084-5091.

Chen, J., Zhang, L.X., 2005. Scaling behaviors of DNA cluster for conding and non-conding sequence. *Chaos Solitons & Fractals*, **24**:115-121.

Cheng, J., Zhang, L.X., 2005. Scaling behaviours of C-G cluster for Chromosomes. *Chaos Solitons & Fractals*, **25**:339-346.

De Sousa Vieira, M., 1999. Statistics of DNA sequences: A low-frequency analysis. *Physical Review E*, **60**:5932-5937.

Herzel, H., Grobe, I., 1997. Correlations in DNA sequences: the role of protein coding segments. *Physical Review E*, **55**:800-810.

Herzel, H., Trifonov, E.N., Weiss, O., Grobe, I., 1998. Interpreting correlations in biosequences. *Physica A*, **248**: 449-459.

Li, W., 1992. Generating nontrivial long-range correlations and 1/*f* spectra by replication and mutation. *Int J Bif & Chaos*, **2**:137-154.

Li, W., 1997. The study of correlation strctures of DNA sequences: a critical review. *Journal of Computer Chemistry*, **21**:257-271.

Li, W., Kaneko, K., 1992. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA squence. *Europhysics Letter*, **17**:655-660.

Peng, C.K., Buldyrev, S.V., Havlin, S., Simonis, M., Stanley, H.E., Goldberger, A.L., 1994. Mosaic organization of DNA nucleotides. *Physical Review E*, **49**:1685-1689.

Poland, D., 2004. The persistence exponent of DNA. *Biophysical Chemistry*, **110**:59-72.

Provata, A., Almirantis, Y., 1997. Mosaic organization of DNA sequences. *Physica A*, **247**:482-487.

Provata, A., Almirantis, Y., 2002. Statistical dynamics of DNA clustering. *Journal Statistial Physics*, **106**:23-56.

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**:1075-1084.

Viswanathan, G.M., Buldyrev, S.V., Havlin, S., Stanley, H.E., 1998. Long-rang correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA. *Physica A*, **249**:581-586.

Voss, R.F., 1992. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. *Physical Review Letters*, **68**:3805-3808.