Scaling behaviors of CG clusters in coding and noncoding DNA sequences

Linxi Zhang *, Jin Chen
Department of Physics, Wenzhou Normal College, Wenzhou 325027, PR China
Accepted 6 July 2004

Abstract

In this paper the statistical properties of CG clusters in coding and non-coding DNA sequences are investigated through calculating the cluster-size distribution of CG clusters \( P(S) \) and the breadth of the distribution of the root-mean-square size of CG clusters \( \sigma_m \) in consecutive, non-overlapping blocks of \( m \) bases. There do exist some differences between coding and non-coding sequences. The cluster-size distribution of CG clusters \( P(S) \) for both coding and non-coding sequences follows an exponential decay of \( P(S) \propto e^{-\alpha S} \), and the value of \( \alpha \) depends on the percentage of C–G content for coding sequences. It can fit into a linear line regularly but the case is contrary for noncoding sequences. We find that \( n(m) \propto m^{-\gamma} \) of CG clusters all obeys the good power-law decay of \( \zeta(m) \propto m^{-\gamma} \) in both coding and non-coding sequences, and the value of \( \gamma \) is 0.949 ± 0.014 and 0.826 ± 0.011 for coding and noncoding sequences, respectively. Therefore, we can distinguish between coding and non-coding sequences on the basis of the value of \( \gamma \). At the meantime, we also discuss the power-law of \( n(m) \propto m^{-\gamma} \) for random sequence, and find that the value of \( \gamma \) for random sequence is very close to 1.00. So we can know that the value of \( \gamma \) for coding sequences is more close to the random sequence, and obtain the conclusion that the behavior of coding sequence trends to random sequence more similarly. This investigation can provide some insights into DNA sequences.

© 2004 Elsevier Ltd. All rights reserved.

1. Introduction

The genetic information of organisms is stored in DNA, which is a sequence of four different bases: adenine A, guanine G (purines), cytosine C, and thymine T (pyrimidines) [1]. Each DNA molecule is packed in a chromosome, which varies in length from \( 10^5 \) base pairs (bp) in yeast to \( 10^8 \) bp in human. During the past few years, there has been intense discussion about the existence, the nature and the origin of long-rang correlations in DNA sequences. Different techniques including mutual information functions [2], auto-correlation functions [3], power spectra [4], “DNA walk” representation [5,6], and Zipf analysis [7] were used for statistical analysis of DNA sequences. Moreover it is still a challenge how to distinguish coding and noncoding regions. Noncoding region is a segment of DNA that does not comprise a gene and thus does not code for a protein. Noncoding regions are interspersed throughout DNA, however, coding region is the sequence of DNA consisting of a series of nucleotide bases (code) giving rise to the mature messenger
RNA that will be translated into the specific amino acids of the protein product. For example, roughly 5% of human DNA is coding, and can be translated into proteins by various combinations of three nucleotides. Although the rest of human DNA is noncoding, some regions are known to be involved in various regulatory processes. The statistical properties of coding sequences are different from noncoding sequences. In recent years scientists have worked out many kinds of gene maps and the continual task is to get the available information from these raw data. The computational recognition of genes is one of the challenges in the analysis of newly sequenced genomes, which is fundamental for modern functional genomic. For protein-coding genes, many biologists and computer scientists have developed computational gene finding tools to predict novel genes in sequence data with reasonable efficiency. The software such as GenScan [8], GraiEXP [9], GeneWise [10], Genie [11], GeneParser [12] etc can elementarily identify the characteristics of DNA sequences. At the same time some articles were published to discuss the characteristics in coding and noncoding DNA sequence using the way of simple sequence repeats (SSR) [13–20], dimeric tandem repeats (DTR) [21,22], entropic segmentation method [23], scale-invariant analysis [5,22], and statistical linguistics method [7]. However these methods do not give specific information for how different regions are characterized and also failed to distinguish one given species from another. Furthermore it is an important and pressing question to rapidly know whether or not the appointed sequence is coding or noncoding for more and more sequences accumulate while the modern computer technique could not get up with the increase in time, even there are so many useful tools. In fact, only a few quantitative and statistical methods are currently available for analyzing such information so we can develop a great deal of work in this field.

In this paper, we will discuss the cluster-size distribution of CG clusters \( P(S) \) and the distribution of the root-mean-square size of CG clusters \( \sigma_m \) in consecutive, non-overlapping blocks of \( m \) bases for coding and noncoding sequences. Some different behaviors are obtained in coding and noncoding sequences.

2. Model

Our aim here is to investigate the difference of statistical properties of DNA sequences between coding and noncoding sequences. The division of the genome into consecutive non-overlapping blocks containing \( m \) bases is shown in Fig. 1. Here we select the snippet with 45 nucleotide letters as an example to show our model. The upper row illustrates the assignment of bases to blocks for the case of \( m = 10 \). The name of “cluster” has been defined as \( P_a \) (A and G) and \( P_c \) (C and T) by Provata and Almirantis on the class of purine and pyrimidine in their early work [24–26]. Here we change the definition of the cluster [27]. It is well known that the nucleotide A is paired with T, and C is paired with G in the complementary respectively. So in the lower row, we assign 0’s to A or T bases and 1’s to C or G bases [27]. Therefore, we can obtain the number of continuous C–G contents, and the distribution of CG clusters. It is interesting that this unexpected parameter can help us to understand the DNA sequences clearly.

On the other hand we define another important parameter of the average size \( \bar{n} \) of CG clusters (1’s) in each block as:

\[
\bar{n} = \frac{1}{N} \sum_{i=1}^{N} \bar{S}_i
\]

Block-size: \( m = 10 \)

<table>
<thead>
<tr>
<th>AAGCTTGTCA</th>
<th>ATTACCGAT</th>
<th>CTCGTTAGTG</th>
<th>CTCGGAAATGA</th>
<th>AAGGG......</th>
</tr>
</thead>
<tbody>
<tr>
<td>0011001010</td>
<td>0000111100</td>
<td>1011000101</td>
<td>1011100010</td>
<td>00111......</td>
</tr>
</tbody>
</table>

\( \bar{S} \) : \( \frac{2+1+1}{3} = \frac{4}{3} \)  \( \frac{4}{4} \)  \( \frac{1+2+1+1}{4} = \frac{5}{4} \)  \( \frac{1+3+1}{5} = \frac{3}{5} \)  

\( n \) : 4  4  5  5  

Fig. 1. Illustration of the use of blocks to obtain CG cluster-size distributions. The upper row illustrates the division of the genome into consecutive non-overlapping blocks each containing bases. Here we list 45 nucleotide letters, and \( m = 10 \). \( \bar{n} \) and \( n \) are the average size of CG clusters and the average number of the 1’s (C–G contents) with, \( m = 10 \) respectively.
here \( N \) is the number of CG clusters, and \( S_i \) is the size of CG clusters of \( i \)-th cluster in the consecutive, non-overlapping block. In fact, Poland have discussed the distribution of C–G content according to the number of C–G contents in block [27]. In order to compare with Poland’s results, we also list the total number of C–G contents \( n(= \sum_{i=1}^{N} S_i) \) in the lowest row with this same sequence for block size \( m = 10 \) [27]. In general the total number of C–G contents \( n \) in the consecutive, non-overlapping block is always larger than the average size of \( \bar{n} \) in the same block. At the meantime, we can obtain more accurate and detailed information from the distribution of average size \( \bar{n} \) of CG clusters instead of the distribution of C–G contents in the consecutive, non-overlapping block. For example there are two blocks with the sequence: 1000101100 and 0000111100 respectively. It is obvious that these two sequences are completely different but we can get the same number of C–G contents with \( n = 4 \) yet. On the other hand, we can obtain the correct results according to the average size \( \bar{n} \) of CG clusters. i.e. \( \bar{n} = 4 \) and \( \bar{n} = 4 \) respectively for these two blocks. Therefore, we can find out that the average size of CG cluster \( \bar{n} \) can exhibit not only the numerical value distribution function for the C–G contents in the consecutive, non-overlapping block but also the position distribution function for CG clusters simultaneously. In Fig. 1 with \( m = 10 \), if \( n = \frac{4}{3} \) in one block, we can obtain that there are four C–G units and six A–T units in one block, these four 1’s divide into three CG clusters and the array must be . . . 1 . . . 1 . . . 1 . . . 1 . . . . If the value of \( m \) is small in a certain extent, the average size of CG clusters \( \bar{n} \) can show all characteristics of whole sequence per block including the C–G and A–T contents. In fact, the number of C–G content \( n \) is only unilateral for considering the total number of C–G contents, and it neglects the array sequence and covers the feature of different sequences with the same number of C–G contents. However, the average size of CG clusters \( \bar{n} \) can include more information of the sequence perfectly.

3. Results and discussion

3.1. Cluster-size distribution of CG clusters \( P(S) \) in whole DNA sequence

We define the cluster-size distribution of CG clusters \( P(S) \) as

\[
P(S) = \frac{N_S}{N_1}
\]

Here \( N_1 \) is the number of CG clusters with cluster size \( S = 1 \) in whole DNA sequence. Apparently, \( N_1 \) is the largest one in DNA sequence so the value of \( P(S) \) is always smaller than 1.0. In Fig. 2, the cluster-size distribution of CG cluster in whole sequence (solid square) is presented for the noncoding DNA sequence DROMHC (Drosophila melanogaster MHC, 22663 bp) and open square is for the coding DNA sequence HUMDYS (Human Dystrophin, 13957 bp). The cluster-size distribution of CG cluster in semi-logarithmic scale can obtain with the slopes of \(-0.360\) and \(-0.340\), respectively. From Fig. 2, we can obtain the power law:

\[
P(S) \sim e^{-2S}
\]
Similar results are obtained for the other 24 different DNA sequences (including 12 coding DNA sequences and 12 non-coding DNA sequences). In Table 1, we give the values of $\alpha$ for these DNA sequences. In the upper table there are 13 coding sequences including different categories human, animal epiphyte and bacteria with the length range from 6008 to 924430 bp. The extrema of 0.233 and 0.627 fall in HSHSPG2B (14327 bp) and CDTOXACD (8407 bp) respectively. The column next to $\alpha$ is the correlation coefficient. If the line can fit all the data completely the value of correlation coefficient should be equal to 1.0. Here our results for the correlation coefficient here range from 0.983 to 0.999, which can illuminate that this fit is perfectly well. In the lower table there are the results for noncoding sequences with the length range from 7272 to 313573 bp. HSAK1 (12229 bp) has the minimum of $\alpha$ ($= 0.236$) and HSALBGC (19002 bp) has the maximum one ($= 0.507$). Similarly, the correlation coefficient is so good for all the fit lines are close to lines successfully with small changes from 0.982 to 0.998. In Table 1, the values of $\alpha$ seem to change arbitrarily, so it is dif-

<table>
<thead>
<tr>
<th>No.</th>
<th>Sequence</th>
<th>Code</th>
<th>Length</th>
<th>$P(S)\propto e^{-\alpha S}$</th>
<th>$\xi(m)\propto m^{-\gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha$</td>
<td>Correlation coefficient</td>
<td>$\gamma$ Correlation coefficient</td>
</tr>
<tr>
<td>1</td>
<td><em>Clostridium difficile</em> toxin A gene</td>
<td>CDTOXACD</td>
<td>8407</td>
<td>0.627 0.999</td>
<td>0.907 0.996</td>
</tr>
<tr>
<td>2</td>
<td>Human mitochondrion, complete genome</td>
<td>MIHSGG</td>
<td>16569</td>
<td>0.351 0.997</td>
<td>1.025 0.998</td>
</tr>
<tr>
<td>3</td>
<td>Human mRNA for dystrophin</td>
<td>HSDMDR</td>
<td>12446</td>
<td>0.451 0.997</td>
<td>0.930 0.999</td>
</tr>
<tr>
<td>4</td>
<td>Bacteriophage sk1</td>
<td>AF011378</td>
<td>28451</td>
<td>0.559 0.990</td>
<td>0.941 0.998</td>
</tr>
<tr>
<td>5</td>
<td>Human hepuran sulfate proteoglycan</td>
<td>HSHSPG2B</td>
<td>14327</td>
<td>0.233 0.993</td>
<td>1.099 0.990</td>
</tr>
<tr>
<td>6</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>NC_001145</td>
<td>924430</td>
<td>0.381 0.991</td>
<td>0.921 0.999</td>
</tr>
<tr>
<td>7</td>
<td><em>Escherichia coli</em></td>
<td>ECO110K</td>
<td>111408</td>
<td>0.278 0.991</td>
<td>0.912 0.999</td>
</tr>
<tr>
<td>8</td>
<td><em>Drosophila melanogaster</em> MHC</td>
<td>DROMYONMA</td>
<td>6338</td>
<td>0.290 0.983</td>
<td>1.058 0.993</td>
</tr>
<tr>
<td>9</td>
<td>Human $\beta$-cardiac MHC</td>
<td>HUMBMYH7CD</td>
<td>6008</td>
<td>0.330 0.984</td>
<td>0.918 0.998</td>
</tr>
<tr>
<td>10</td>
<td>Human Dystrophin</td>
<td>HUMDYD</td>
<td>13957</td>
<td>0.460 0.998</td>
<td>0.925 0.998</td>
</tr>
<tr>
<td>11</td>
<td><em>Dictostelium discoideum</em> MHC</td>
<td>DDIMYHC</td>
<td>6680</td>
<td>0.522 0.985</td>
<td>0.974 0.995</td>
</tr>
<tr>
<td>12</td>
<td><em>Escherichia coli</em></td>
<td>ECOTSF</td>
<td>91430</td>
<td>0.288 0.995</td>
<td>0.898 0.996</td>
</tr>
<tr>
<td>13</td>
<td>Chicken nonmuscle myosin heavy chain gene</td>
<td>CHKMYHN</td>
<td>7003</td>
<td>0.454 0.997</td>
<td>0.916 0.998</td>
</tr>
<tr>
<td></td>
<td>Average value</td>
<td></td>
<td></td>
<td></td>
<td>0.949 ± 0.014</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Sequence</th>
<th>Code</th>
<th>Length</th>
<th>$P(S)\propto e^{-\alpha S}$</th>
<th>$\xi(m)\propto m^{-\gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha$</td>
<td>Correlation coefficient</td>
<td>$\gamma$ Correlation coefficient</td>
</tr>
<tr>
<td>14</td>
<td>Human Ter-C-delta</td>
<td>HUMTCRADCV</td>
<td>97630</td>
<td>0.386 0.994</td>
<td>0.854 0.999</td>
</tr>
<tr>
<td>15</td>
<td><em>Caenorhabditis elegans</em> Myosin</td>
<td>CELMYUNC</td>
<td>9000</td>
<td>0.347 0.997</td>
<td>0.871 0.998</td>
</tr>
<tr>
<td>16</td>
<td><em>Drosophila melanogaster</em> MHC</td>
<td>DROMHC</td>
<td>22663</td>
<td>0.360 0.994</td>
<td>0.813 0.999</td>
</tr>
<tr>
<td>17</td>
<td>Chicken embryonic MHC</td>
<td>CHKMYHE</td>
<td>31111</td>
<td>0.478 0.988</td>
<td>0.831 0.996</td>
</tr>
<tr>
<td>18</td>
<td><em>Caenorhabditis elegans</em></td>
<td>CEY57G11C</td>
<td>313573</td>
<td>0.341 0.986</td>
<td>0.863 0.999</td>
</tr>
<tr>
<td>19</td>
<td><em>Caenorhabditis elegans</em></td>
<td>CEY17G7B</td>
<td>143092</td>
<td>0.362 0.998</td>
<td>0.829 0.999</td>
</tr>
<tr>
<td>20</td>
<td>Homo sapiens chromosome 16 BAC clone</td>
<td>AF_001550</td>
<td>173882</td>
<td>0.335 0.993</td>
<td>0.788 0.999</td>
</tr>
<tr>
<td>21</td>
<td><em>Caenorhabditis elegans</em></td>
<td>CEY41E3</td>
<td>150641</td>
<td>0.374 0.996</td>
<td>0.827 0.998</td>
</tr>
<tr>
<td>22</td>
<td>Human MHC class II</td>
<td>HSHCDBG</td>
<td>7272</td>
<td>0.320 0.985</td>
<td>0.734 0.995</td>
</tr>
<tr>
<td>23</td>
<td>HLA-DC-Beta gene</td>
<td>HSAK1</td>
<td>12229</td>
<td>0.236 0.988</td>
<td>0.832 0.994</td>
</tr>
<tr>
<td>24</td>
<td>Human serum albumin gene</td>
<td>HSALBGC</td>
<td>19002</td>
<td>0.507 0.998</td>
<td>0.858 0.998</td>
</tr>
<tr>
<td>25</td>
<td>Human adenosin deaminase gene</td>
<td>HSADAG</td>
<td>36741</td>
<td>0.351 0.993</td>
<td>0.761 0.995</td>
</tr>
<tr>
<td>26</td>
<td>Human vitamin D-binding protein (GC) gene</td>
<td>HSVITDBP</td>
<td>55136</td>
<td>0.470 0.982</td>
<td>0.873 0.994</td>
</tr>
<tr>
<td></td>
<td>Average value</td>
<td></td>
<td></td>
<td></td>
<td>0.826 0.997 ± 0.011</td>
</tr>
</tbody>
</table>
difficult for us to find the relationship between coding and noncoding sequences even though the fit is so well by the power law. In fact, we can find the value of $\alpha$ depends on the percentage of cluster C–G content in different sequences for coding sequences (see Fig. 3(a)), while the dependence of value of $\alpha$ on the percentage of cluster C–G content for noncoding sequence is not obvious (see Fig. 3(b)).

### 3.2. Cluster-size distribution in consecutive non-overlapping block

Now we begin to discuss the distribution of CG cluster by the parameter of the average size of CG clusters $\bar{n}$ in consecutive non-overlapping blocks for the actual DNA sequences and random sequences. As we all know that the nucleotide A, T, C and G equably can fill with one DNA sequence ideally, and there are must be some differences between the actual and ideal percentage. Therefore we also can get the fixed relation:

$$f_{\text{AT}} + f_{\text{CG}} = 1$$  \hspace{1cm} (4)

Here $f_{\text{AT}}$ and $f_{\text{CG}}$ is the overall fractions of the two states (A or T and C or G).

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 size of CG clusters $\bar{n}$, we list these following two parameters, which give the first two moments of the distribution:

$$\bar{\Pi}(m) = \langle \bar{n} \rangle$$  \hspace{1cm} (5)

$$\bar{\Pi}(m) = \langle \bar{n}^2 \rangle$$  \hspace{1cm} (6)

Here $\langle \rangle$ indicates an average over all $m$ in the DNA sequences. Therefore, we take as a standard measure of the breadth of the distribution the root-mean-square size of CG clusters:
This is a useful method to define a parameter in terms of the difference between the average of the square and the square of the average \[\sigma(m) = \sqrt{\overline{m^2}} - \overline{m^2}\] (7)

Here we are interested in how to distinguish the actual coding and noncoding DNA sequence handily by using the distribution function of CG clusters. To investigate the relation between the width of the empirical distribution \(\sigma(m)\) and the \(n\)-dependence, we give the following function:

\[\zeta(m) = \frac{\sigma(m)}{\sqrt{m}}\] (8)

In Fig. 4, we show the results of \(\zeta(m)\) according to Eq. (8) for three DNA sequences. The curves with solid and open squares represent \textit{Drosophila melanogaster} MHC (DROMHC, 22663 bp, noncoding sequence) and Human Dystrophin (HUMDYS, 13957 bp, coding sequence), respectively. We can get the direct conclusion that the values of \(\zeta(m)\) in a double-logarithmic scale all can be fitted to be linear well and the slopes of HUMDYS and DROMHC sequences are 0.813 and 0.925, respectively. Therefore, the behavior of \(\zeta(m)\) in Fig. 4 strongly suggests a power law of the form:

\[\zeta(m) \propto m^{-\gamma}\] (9)

In order to investigate the scaling behaviors of cluster-size distribution in consecutive non-overlapping blocks in more detail, we also compute the values of \(\zeta(m)\) for the other 24 coding and noncoding sequences (12 coding sequences and 12 noncoding sequences) and give the data comparatively in Table 1. For coding DNA sequences, the minimum of \(\gamma\) is 0.898 for ECOTSF (91430 bp) and the maximum one is 1.058 for DROMYONMA (6338 bp), and the average value of \(\gamma\) is 0.949 ± 0.014. The slope seems to be independent of the length of DNA sequences and fluctuates only a little for different coding DNA sequences. In Table 1, we can also obtain the results that the correlation coefficients for the slopes of the lines range from 0.990 to 0.999 and the average value is 0.997. From the fact that the correlation coefficients are very close to 1.0, we can confidently to say that our results are believable greatly. The lower table also gives the results for 13 noncoding sequences with the length from 7272 to 313573 bp. HSHCDCB (7272 bp) has the minimum value of \(\gamma\) (= 0.734) and HSVITDBP (55136 bp) has the maximum value of \(\gamma\) (= 0.873), and the average value is 0.826 ± 0.011. We can also work out the correlation coefficients for the slopes of the lines from 0.994 to 0.999 with the average of 0.997, which can also illuminate that this fit is very good. Compare with the values of \(\gamma\) for coding and noncoding sequences, we can inspiringly find that there are visible differences for coding and noncoding DNA sequences, and the value of \(\gamma\) for coding sequences is larger than that for noncoding sequences. This means that this method can primitively help us analyze and know whether the unknown sequence is coding sequence or not simply and conveniently.

Now we begin to discuss the behavior of \(\zeta(m)\) for random sequence, and the results are also shown in Fig. 4. In order to compare with DROMHC (22663 bp) and HUMDYS (13957 bp), here we make a random sequence with the length of 20000 bp, not far from these two actual DNA sequences. In the same way the behavior of the random sequence in a double-logarithmic scale is so excellent that the line give a linear fit to the data with the slope of 0.993. To examine our conclusion in more detail, we also compute the value of \(\zeta(m)\) for the random sequence with the length of 1,000,000 bp and get the slope of 0.999 in a double-logarithmic scale with the correlation coefficient of 0.999, which is very close to

Fig. 4. \(\zeta(m)\) as a function of for \textit{Drosophila melanogaster} MHC (DROMHC, 22663 bp, non-coding sequence, ■), Human Dystrophin (HUMDYS, 13957 bp, coding sequence, □), and random sequence (20000 bp, ∆).
1.0 (This line is not given here). Therefore, our results are reliable. Compare with the values of $\gamma$ in Table 1, we can find that the value of $\gamma$ in coding DNA sequences is more close to the random sequence than that in noncoding DNA sequences. Therefore we can get the important conclusion from our calculation that the behavior of coding sequence is more close to the random sequence, and noncoding sequence exhibit deterministic signatures, which agrees with the others’ work entirely [5,28].

![Graph showing $\xi(m)$ as a function of $m$ for Human chromosomes 21 (■) and 22 (□).](image1)

**Fig. 5.** $\xi(m)$ as a function of $m$ for Human chromosomes 21 (■) and 22 (□).

![Graph showing average size of C-G content for different block sizes.](image2)

**Fig. 6.** Average size of CG cluster with different blocksize for ECOTSF (91430 bp). Here (a) $m = 5$, and (b) $m = 495$. 

![Graph showing average size of C-G content for different block sizes.](image3)
Here we also discuss the behavior of $\zeta(m)$ for human chromosomes 21 (33,924,807 bp) and 22 (34,352,073 bp) [29–31], and the results are given in Fig. 5. Our aim here is to know whether the length of DNA sequences can influence the behavior of cluster-size distribution or not. In Fig. 5, the curves with solid and open squares represent human chromosomes 21 and 22 respectively. The lines are fitted to the power law well in the same way with the slope of $-0.659$ and $-0.686$, respectively. Therefore, the same conclusion can be obtained for very long DNA sequences.

Here we will discuss the reason why $\zeta(m)$ decreases with $m$. Fig. 6 gives the average size $\bar{n}$ of cluster C–G as a function of position in DNA sequence with different block size $m = 5$ and $m = 495$ respectively for Escherichia coli genomic DNA sequence (ECOTSF, 91430 bp, coding sequence). In the case of $m = 5$ (see Fig. 6(a)), there is more block than the case of $m = 495$. At the same time, there are seven cases for the average size $\bar{n}$ for Escherichia coli genomic DNA sequence, i.e. $n = 0, 1, 2, 1.5, 3, 4, \text{ and } 5$, respectively. However, the average size of $\bar{n}$ of CG clusters with the blocksize of $m = 495$ is more complicated because when the blocksize becomes large, the number of the blocks reduces and the average size of CG clusters $\bar{n}$ may have more varies accordingly. In Fig. 6(b), all the values of $\bar{n}$ range from 1.5 to 2.8, especially around $\bar{n} = 2.2$. In this region it is densely covered while this distribution trends to disperse more irregularly beyond this range. We can daringly consider that with the increase of the block size this graph, collecting together in middle and dispersing in both sides, can be seen more distinctly. This leads the mean-square fluctuation $\zeta(m)$ to decease with increasing $m$.

The statistics of DNA sequence is an active topic of research nowadays. This investigation can help us to distinguish coding and noncoding sequences, and the cluster-size distribution can provide some insights into DNA sequences.

**Acknowledgments**

This research was financially supported by National Natural Science Foundation of China (Nos. 20174036, 20274040), and Natural Science Foundation of Zhejiang Province (No. 10102).

**References**


