Genome-Wide Search for Local DNA Segments with Anomalous GC-Content

ANDREY ILATOVSKIY and MICHAEL PETUKHOV

ABSTRACT

An anomalous (i.e., significantly different from genome-average) GC-content is often used as one of the markers to reveal the events of horizontal gene transfer (HGT). Unfortunately, results obtained by the traditional fixed-length window analysis strongly depend on an arbitrary selection of DNA window length. Here we present a new method for genome-wide statistical analysis of GC-content without that drawback. The method is based on a set of nonparametric statistical tests and is capable of providing reliable estimations of both a local and global GC-content, and thus can identify small local areas (as short as 30 bp) with anomalous GC-content in a bacterial genome. The tests, applied to a well-studied bacterial genome of Escherichia coli K-12, show that approximately 21% of the genome belongs to the anomalous GC-content areas. Among top 23 anomalous GC-content areas, seven correspond to the annotated prophages, four to Rhs elements, and two to IS elements. A remaining 10 areas contain putative horizontally transferred DNA and genes with still unknown functions.

Software is available at http://mml.spbstu.ru/gcstat.

Key words: GC-content, genome, horizontal gene transfer, sequence analysis.

1. INTRODUCTION

An anomalous GC-content (i.e., significantly different from its genome-average value) is one HGT event marker, which was shown to be an important mechanism of microbial evolution (Koonin et al., 2001). Genes known to be responsible for drug resistance (Mazel and Davies, 1999), pathogenicity (Hacker et al., 1997), capability for symbiosis (Sullivan and Ronson, 1998), and new metabolic pathways (Friedrich, 2002) are likely to be acquired by the HGT mechanism. Since complete data on genome sequences of closely related microorganisms are often unavailable, HGT events can be identified by means of the genome sequence analysis. As has been reported, the GC-content heterogeneity in a single genome is less significant than that in genomes of different microorganisms (Karlin et al., 1998). This fact provides an opportunity to identify HGT events by a local GC-content genome analysis. Such an approach was implemented in a number of methods (Lawrence and Ochman, 1998; Yoon et al., 2005).
To identify areas with anomalous GC-content, the fixed-length window method is traditionally used (Kunst et al., 1997). However, this approach has a serious drawback. Particularly, both a local GC-content and its standard deviation (SD), which is often used as a criterion of GC-content anomaly, depend on an arbitrary window length selection (Bernaola-Galván et al., 2004). This throws into question the reliability of results obtained by this method.

Several alternative methods for GC-content estimation have been reported in the literature. For instance, the method of genome segmentation via recursive dichotomy has been developed to characterize GC-content heterogeneity (Bernaola-Galván et al., 2004). First, initial global GC-content distribution is calculated using the fixed-length DNA window. Then, the distribution is subdivided into two parts by a point, which provides a maximal difference of GC-content of these parts. This recursive process is continued until there is no required point, resulting into segments, each of which has GC-content significantly different from that of the next one. Although the method is able to localize areas with anomalous GC-content, it is hard to interpret the results, for whole genome with relatively heterogeneous GC-content due to each segment is “normalized” to neighbor segments rather than to a genome-average GC-content.

Alternatively, one can calculate GC-content at an arbitrary point of genome by the windowless method based on the Z curve (Zhang et al., 2001). This method is capable of localizing areas of high GC-content heterogeneity. However, similar to a previous case, the deviation of local GC-content from genome-average value is difficult to obtain.

In this work, we present a new method for genome-wide statistical analysis of GC-content. It is based on a set of nonparametric statistical tests and is capable of providing both reliable estimations of local GC-content and localization of DNA segments with anomalous GC-content of minimal length of 30 bp. The method has been applied to a well-studied genome of *Escherichia coli* K-12.

Software is available at http://mml.spbstu.ru/gcstat.

2. METHODS

2.1. The genome

Complete DNA sequence of *E. coli* K-12 MG1655 genome and its annotation (Blattner et al., 1997) were taken from GenBank (accession number U00096) (Benson et al., 2006). *E. coli* K-12 genome is a circular chromosome of 4.64 Mbp. As of April 24, 2007, 4267 protein-coding genes were annotated, approximately 25% of which have no attributed functions. The coding density (i.e., relative length of protein-coding regions) is ~87%. *E. coli* K-12 genome-average GC-content is ~50.8%.

2.2. Shapiro-Wilk W-test

The Shapiro-Wilk W-test is a widely used test for normality of a distribution. Its null hypothesis is that the sample follows the normal distribution. The test statistic W is the ratio of two estimates of variance of a normal distribution based on a random sample of n observations \( \{y_i\}_{i=1}^n \) with mean \( \bar{y} \). The numerator of W is the square of linear estimator of SD (based on weight coefficients \( \{a_i\}_{i=1}^n \) ), and the denominator is the sum of squares of the observations about the sample mean: 

\[
W = \frac{\left[ \sum_{i=1}^{n} a_i y_i \right]^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.
\]

Quantity \( Z = \frac{W^{1/2} - \mu}{\sigma} \) follows normal distribution, and the p-value can be determined by integration of standard normal probability function

\[
p = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \exp(-\frac{x^2}{2}) \, dx.
\]

Coefficients \( \{a_i\}_{i=1}^n \), \( \lambda, \mu \) and \( \sigma \) were calculated by a special algorithm (Royston, 1995), which is applicable to any sample of n observations in the range of 3 ≤ n ≤ 5000.

2.3. Mann-Whitney U-test

The Mann-Whitney U-test is similar to the two-sample T-test but does not assume normal distribution for samples. Its null hypothesis is that the samples have the same median. The U-test (as well as the majority of nonparametric tests) is based on sample ranks. Using the rank sums \( R_1 \) and \( R_2 \), the test statistic \( U \) for two samples of \( N_1 \) and \( N_2 \) observations can be calculated as

\[
U_i = N_1 N_2 + N_1 (N_1 + 1)/2 - R_i, \quad i = 1, 2;
\]

\[
U = \min(U_1, U_2).
\]

Moreover, the test statistic \( U \) can be calculated directly by its definition (Mann and Whitney, 1947). Quantity \( Z = U - \frac{1}{2} (N_1 N_2 - 1) \) follows normal distribution for \( N_{1,2} \geq 8 \), and the p-value (two-sided) can be determined by integration of standard normal probability function

\[
p = \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \exp(-\frac{x^2}{2}) \, dx.
\]

In this work, \( N_2 \approx 5000 \), while \( N_1 \) can vary on wide range of values. In order to
find minimal $N_1$ value, we reproduced calculations of the test statistic $U$ and its probability as described in Mann and Whitney, (1947). It was found that minimal $N_1 = 5$ provides required approximation level of $U$ distribution by normal one (data not shown).

2.4. The binomial test and the segmentation algorithm

As will be shown below, the results by conventional fixed-length window method are affected by an arbitrary window length selection. To overcome this drawback, we developed a new approach based on the binomial test, capable of estimating the significance of GC-content calculations.

Any given position of DNA sequence is occupied by one of two base pairs, GC or AT. For a random sequence, nucleotide occurrence is independent, and the probability of GC base pair to be at a given position is equal to the average GC-content, $C_{GC}$. For a DNA segment, the probability to have certain GC-content is calculated as follows. Let a DNA segment of $N$ bp have $N_{GC}$ GC base pairs. Obviously, $N_{GC}$ ranges from 0 to $N$. It is known from statistics that the probability of a definite number of occurrences ($N_{GC}$) of a random event (GC base pair), having constant probability of occurrence ($C_{GC}$) in finite ($N$) sample of independent observations (the DNA segment), follows the binomial distribution. Thus, the probability of a DNA segment of $N$ bp to have $N_{GC}$ GC base pairs by a chance is equal:

$$P(N, N_{GC}, C_{GC}) = \frac{N!}{N_{GC}!(N - N_{GC})!} C_{GC}^{N_{GC}} (1 - C_{GC})^{N - N_{GC}} \quad (1)$$

Estimation of binomial probability by equation (1) will be named “binomial test” in further discussions. If binomial probability of a DNA segment is less than predefined significance level of the binomial test (for instance 5%), then such segment will be considered as a nonrandom one.

An increase of DNA segment length results in two effects. Since a number of discrete GC-content values increases, the binomial probability distribution significantly narrows and the value of maximal probability decreases. The latter indicates that above a critical segment length $N_C$, binomial probability of a DNA segment to have a random GC-content becomes less than any predefined significance level. For example, at a conventional significance level of 5% and genome-average GC-content of 50.0%, the critical length $N_C$ is 255 bp. GC-content of sequenced bacterial genomes ranges from 16% to 72% for *Carsonella ruddii* (Nakabachi et al., 2006) and *Streptomyces coelicolor* (Bentley et al., 2002), respectively, and therefore $N_C$ does not exceed 500 bp at the 5% significance level for the bacterial genomes.

Based on the binomial test, the guaranteed convergent iterative nonrecursive nonparametric algorithm of genome segmentation was developed for calculations of nonrandom global GC-content distribution. First in the algorithm, a DNA segment of small length (with binomial probability certainly exceeds the significance level of 5%, e.g., 2 bp) is selected. Second, GC-content of this segment and its randomness are calculated by the binomial test. Then, if binomial test p-value is more than significance level of 5%, length of the segment is increased; otherwise, the segment is recorded into global GC-content distribution, and next segment is selected. Finally, applying this algorithm to whole genome, starting from the first nucleotide, the nonrandom global GC-content distribution is prepared.

2.5. The $\beta$-test

In order to determine a minimal size of anomalous GC-content area, which can be identified by the Mann-Whitney U-test, beta error rate on a simulated dataset was calculated by a specific test, which we further denote as the $\beta$-test. First, in this test, the artificial sequence of length $L = 1$ Mbp with linear gradient of local GC-content from 0 to 1 is prepared. Second, using the global GC-content distribution of real genome sequence, anomalous GC-content range $[0, G_{CL}] \cup [G_{CU}, 1]$ is defined. From standard significance level of 5%, the anomalous GC-content range is defined as GC-content values falling outside of 95% symmetrical quantile range of the global GC-content distribution, i.e., $G_{CL}$ and $G_{CU}$ are quantiles 0.025 and 0.975 respectively. This range of GC-content corresponds to area $[0, L_{L}] \cup [L_{U}, 1]$ bp of artificial sequence, where $L_{L} = L \times G_{CL}$ and $L_{U} = L \times G_{CU}$. Thus, any local distribution of GC-content of this area significantly differs from the global GC-content distribution in terms of deviation from the genome-average GC-content.

A local distribution of $l \times n$ size is a sample of $n$ GC-content observations from nonoverlapping DNA segments of $l$ bp. In order to avoid foreign local GC-content heterogeneity of artificial sequence, which is a consequence of sequence generation method, each GC-content value was calculated as $\frac{1}{2} \sqrt{(P_1 + P_2)}$, where
$P_1$ and $P_2$ are start and end positions of artificial sequence segment respectively, and truncated in accordance with the segment length $P_2 - P_1 + 1$.

On the next step, all possible nonoverlapping local GC-content distributions of given size in specified range of artificial sequence position are prepared. Finally, the U-test is applied to compare such local GC-content distributions of artificial sequence with the global GC-content distribution of real genome sequence. If the U-test alpha p-value is above the alpha significance level, there is beta error. Beta p-value is calculated as ratio of beta errors to a total number of tested local distributions of given size. The minimal size of local GC-content distribution, which provides the p-value under the beta significance level, is accepted as the maximal resolution of the method.

3. RESULTS AND DISCUSSION

3.1. Analysis of GC-content by fixed-length window method

Figure 1A shows a global GC-content distribution of *E. coli* K-12 genome as calculated using the fixed-length window method (1,000 bp). One can see that this global distribution is highly heterogeneous, and therefore it is difficult to reliably identify short local areas with anomalous GC-content. Most of the genome (~95%) has GC-content which is within range of two SDs about its mean value (50.8%). It is noteworthy that the p-value by Shapiro-Wilk W-test of normality is less than $10^{-4}$ for the global GC-content distribution of *E. coli* K-12 genome, indicating that this distribution significantly deviates from a normal one. Indeed, a predominant deviation of genome GC-content towards decreased values makes the global distribution of GC-content highly unsymmetrical.

The SD of the global GC-content distribution is often used as a criterion of anomaly of local GC-content. If local GC-content goes out of a range of one or two SDs about the genome-average value, the corresponding local area is considered to be an anomalous GC-content area (Yoon et al., 2005). However, SD depends on an arbitrary window length selection (data not shown). This may cause total rearrangement of predicted local anomalous areas in bacterial genome.

Figure 1B shows the global GC-content distribution of *E. coli* K-12 genome for all feasible window lengths, which shows dependence of GC-content pattern stability on the window length. GC-content is in SD scale, which directly shows local anomalous GC-content area in terms of SD criterion. Anomalous areas with significant deviation of GC-content from the genome-average value form vertical lines on this 3D plot. Stability of a given anomalous area is proportional to size of this line on window length scale. As illustrated by Figure 1B, local anomalous areas of the global GC-content distribution of *E. coli* K-12 genome consists of two fraction: stable anomalous areas, which are identified in almost full range of window lengths, and unstable anomalous areas, which are identified only in restricted range of window lengths. Generally, the

![FIG. 1. The global GC-content distribution of E. coli K-12 genome obtained by fixed-length window method. (A) Window length was 1,000 bp. Two horizontal dashed lines show range of two standard deviations (SDs) about genome-average GC-content (50.8%); ~95% of genome GC-content are in this range. Shapiro-Wilk W-test p-value for this distribution is less than $10^{-4}$, indicating its non-normal nature. (B) Window lengths were 1–100 kbp. Genome position is shown on x-axis (only the first 1 Mbp are shown for clarity); window length is shown on y-axis (in logarithmic scale). GC-content is shown in black-and-white scale. ΔGC is difference between local GC-content and genome-average value, measured in SD units $\Delta GC = (GC - \bar{GC})/SD$. Thus, black areas are DNA segments with anomalous local GC-content in terms of SD criterion. Window length step was 20 bp for the range 1–10 kbp and 200 bp for the range 10–100 kbp.](image-url)
increasing window length tends to decrease the number of unstable anomalous areas, and they vanish at approximately 10 kbp. However, at large window lengths, where Figure 1B shows a high stability of anomalous areas, the method cannot precisely localize them due to low resolution of the global distribution of GC-content. Thus, the fixed-length window method is quick and simple, but its results depend on an arbitrary length selection and therefore cannot be fully trusted.

3.2. Nonrandom global GC-content distribution

The nonrandom global GC-content distribution has been obtained by applying the segmentation algorithm to the *E. coli* K-12 complete genome (Blattner et al., 1997) as described above and consists of 84,122 nonoverlapping segments ranging from 5 to 255 bp in length, with an average of 55 bp.

Figure 2A shows main statistical characteristics of nonrandom global GC-content distribution. Generally, the nonrandom global GC-content distribution has similar shape to the global distribution obtained by the fixed-length window method. First of all, the distribution is highly biased towards decreased GC-content values as indicated by the median shifted against the mean. Additionally, the Shapiro-Wilk W-test p-value is less than $10^{-3}$, indicating a strong deviation of the distribution from normal one.

Another important characteristic of the nonrandom global GC-content distribution is its only local dependence on selection of a start point of nonrandom distribution. In a set of 10,000 tests, a start point was randomly selected and the segmentation algorithm was applied. It was found that after a certain distance, new nonrandom global distributions coincide with the reference distribution obtained by starting the segmentation algorithm from the initial position of *E. coli* K-12 genome. The measure of deviation of these distributions is distance between a current start point and coincidence point. Figure 2B shows the histogram of distance distribution. One can see that, for *E. coli* K-12, genome distance values are exponentially distributed within a range from 0 to 7,899 bp (754 bp on average). Interestingly, preliminary estimates, done for the more complex human genome, show that this property of the algorithm is similar to that obtained for *E. coli* K-12 genome (data not shown). Thus, these computational tests show that arbitrary selection of a start point of the segmentation algorithm affects only locally the nonrandom distribution.

3.3. Analysis of local GC-content in *E. coli* K-12 genome

Many interesting objects found in DNA (such as direct and inverted repeats, protein binding sites) are as short as a few tens of base pairs (i.e., they are shorter than their corresponding nonrandom segments). Another problem with the use of nonrandom distribution in search for short anomalous GC-content areas is its local dependence on selection of a start point of the segmentation algorithm as it was discussed in the

![FIG. 2](image-url) (A) Histogram and box-and-whiskers plot of the nonrandom global GC-content distribution of *E. coli* K-12 genome (corrected for different segment lengths). Box-and-whiskers plot: plus sign shows genome-average GC-content (50.8%), two minus signs show minimal (0.0%) and maximal (100.0%) values, and labels show corresponding percentiles. Histogram: bar for GC-content greater than 100.0% is artefact of sampling and actually represents segments with 100.0% GC-content; smooth curve shows normal distribution with the same mean and standard deviation. The nonrandom global distribution is highly biased towards decreased GC-content; Shapiro-Wilk W-test p-value is less than $10^{-3}$ for the distribution, indicating its non-normal nature. (B) Convergence distance for the nonrandom global GC-content distributions, starting at different DNA positions of *E. coli* K-12 genome. Histogram of distances between current start point and coincidence point of current and reference nonrandom global GC-content distributions (total 10,000 tests). The distance is exponentially distributed within of 0–7,899 bp (754 bp on average).
previous section. Due to these two main reasons, we cannot directly use the segmentation algorithm for local GC-content analysis. These problems can be overcome by an analysis of local GC-content distributions by the fixed-length window method due to both fixed-length and the nonrandom global GC-content distributions are found to be statistically similar at certain window lengths.

First of all, one has to identify a range of window lengths, in which the nonrandom and fixed-length global distributions are statistically similar in terms of the Mann-Whitney U-test. To elucidate this question, the global GC-content distributions of *E. coli* K-12 genome were calculated by the fixed-length window method for a wide range of window lengths and then compared with the nonrandom global GC-content distribution by the nonparametric Mann-Whitney U-test. Figure 3A shows the results of these calculations. Generally, the closer the window length is to the average segment length of the nonrandom distribution (55 bp), the higher the U-test p-value is. It means that both distributions are more statistically similar. The upper limit of similarity range between these two types of distributions is difficultly determined due to p-value oscillations in area from 238 to 269 bp, that approximately corresponds to a maximal segment length of the nonrandom distribution (255 bp). A lower limit of similarity range is somewhat diffused as well. However, at short window lengths of, for example, 4, 6, 8, 9 bp, these two distributions are similar in terms of the Mann-Whitney U-test.

Additionally, dependence of the p-values on the shift of reading windows has been examined. Indeed, the shift of only 1 bp can affect all GC-content values, obtained by the fixed-length window method. Figure 3B shows Mann-Whitney U-test p-values for all possible shifts of global GC-content distributions for window lengths within a range from 2 to 20 bp. One can see that, despite some variations of p-values, no shift can change p-value sufficient for change statistical results of the U-test. The longer window lengths (50, 100, 150, 200, 250, 300 bp) were tested as well. For the first four window lengths, the p-values oscillate above the significance level of 5% (mean p-values are 0.74, 0.50, 0.24 and 0.11, respectively). For window of 300 bp, p-values are below 5% (maximal p-value is 0.0372), and for window of 250 bp, there are some fluctuations about the significance level of 5%, in accordance with data shown in Figure 3A. Thus, the fixed-length window method can be used for local GC-content analysis if window lengths are within a similarity range: {4, 6} ∪ {8, 238} ∪ {240, 242, 244} ∪ {246, 253} ∪ {255} ∪ {259, 261} ∪ {263, 265, 269}.

In order to use the Mann-Whitney U-test for local GC-content analysis, one has to define parameters of the local GC-content distribution such as the segment length and the number of GC-content values in the local distribution. Obviously, increasing these parameters makes the results more reliable, but the ability of the method to identify very short DNA segments with anomalous GC-content will deteriorate. On the other hand, an extreme decrease of local distribution size often results in loss of reliability due to the inapplicability of conventional statistical methods for datasets that are too small.

In order to validate the definition of local GC-content distribution parameters, the β-test was applied to artificial DNA sequence with position dependent GC-content. A mandatory requirement for the β-test is definition of an anomalous GC-content range in the global GC-content distribution of bacterial genome under consideration. Given a standard significance level of 5%, the anomalous GC-content range is defined as GC-content value outside of 95% symmetrical quantile range of the nonrandom global GC-content

![FIG. 3. Results of the Mann-Whitney U-test of statistical similarity between the nonrandom global GC-content distribution of *E. coli* K-12 genome and the global distributions by fixed-length window method. Each horizontal dashed line shows significance level of 5%. (A) Window lengths were from 2 to 1,000 bp. Two vertical dashed lines show average (55 bp) and maximal (255 bp) segment lengths of the nonrandom global distribution. Window lengths of 4 and 6 bp are marked by arrows. (B) Window lengths were from 2 to 20 bp with all possible shifts. Labels at each data point show window lengths.](image-url)
distribution. In the case of the *E. coli* K-12 genome, these quantiles are equal to 0.266 and 0.700, respectively. On the artificial DNA sequence, these quantiles correspond to 266 and 700 kbp positions. Thus, the \( \beta \)-test calculates dependence of probability of false positive (beta p-value) of the U-test on union of two segments of artificial sequence which GC-content is anomalous by the definition [1 bp, 266 kbp] \( \cup \) [700 kbp, 1 Mbp] on parameters of local GC-content distribution.

The \( \beta \)-test p-value is equal to 0 for a large size of local GC-content distribution (i.e., the U-test works without beta errors). In order to find minimal size of local distribution (and consequently a maximal resolution of method), we apply \( \beta \)-test for 4 \( \times \) 5 size (5 segments of 4 bp in the distribution, total 20 bp). The latter is a minimal number of segments required by the Mann-Whitney U-test, while a lower limit of similarity range is 4 bp.

**3.4. Anomalous GC-content areas in *E. coli* K-12 genome**

Scanning the GC-content of *E. coli* K-12 genome by the U-test was done using the 6 \( \times \) 5 size of local GC-content distribution. Figure 4A shows a typical pattern of arrangement of anomalous GC-content areas in the first 10 000 bp of *E. coli* K-12 genome. It was found that approximately 21% of the genome belongs to numerous anomalous GC-content areas distributed all over the genome.

The area density analysis was used to identify the anomalous GC-content areas of *E. coli* K-12 genome for further analysis as follows. The local density of anomalous areas was defined as \( \rho = L_A/L \), where \( L_A \) is the length of anomalous areas and \( L \) is the length of DNA segment under consideration. Although a common practice for density calculations is to fix the denominator and to determine the numerator, in this case we chose fixing the numerator.

Thus, the numerator was fixed to \( L_A = k \times l \times n \), where \( l \times n \) is the size of local GC-content distribution and \( k \) is the number of anomalous areas (arbitrary defined parameter); local denominator \( L_i \) was defined as a distance between \( i \)-th and \( i + k \)-th anomalous areas. Then, a density threshold was defined as \( \rho_T = 50\% \). This means that for the local distribution size 6 \( \times \) 5 and \( k = 50 \) a maximal distance \( L_i^{\text{max}} \) is equal to \( k \times l \times n/\rho_T = 50 \times 6 \times 5/0.5 = 3000 \) bp. If \( L_i > L_i^{\text{max}} \), then \( \rho < \rho_T \), and these anomalous areas were excluded from the further analysis. If there were two overlapping segments with \( \rho \geq \rho_T \), they were united. As a result, \( K \) segments with \( \rho \geq \rho_T \) were selected. The decrease in \( k \) causes an increase \( K \) and therefore can be used as an adjustable parameter to regulate desirable level of detailed elaboration. We set \( k = 50 \), and the top 23 (\( K = 23 \)) most densely populated anomalous areas of *E. coli* K-12 genome were selected (Fig. 4B) with total length of approximately 115 kbp.

A further detailed functional characterization of these areas was done by analysis of corresponding annotation data presented in the GenBank. Area 1 corresponds to cryptic prophage CP4-6, areas 3 and 4 to

---

**FIG. 4.** (A) Typical pattern of local distribution of anomalous GC-content areas. The p-value is the result of the Mann-Whitney U-test of similarity between local fixed-length distributions (6 \( \times \) 5) and nonrandom global GC-content distribution of *E. coli* K-12 genome. Only the first 10 kbp are shown for clarity. Horizontal line shows significance level of 5%; bars (bottom panel) show areas of anomalous GC-content. (B) Map of top 23 anomalous GC-content areas on *E. coli* K-12 chromosome.
DLP12, areas 5 and 6 to e14, area 10 to Qin, and area 12 to CPS-53 (Casjens, 2003; Lindsey et al., 1989; Mehta et al., 2004). Prophages, both intact and cryptic, are common agents of HGT. These cryptic prophages have homologues in different *E. coli* strains and other bacteria; some of them probably have evolved for millions years within bacterial genomes.

Areas 14 and 16 correspond to mobile *IS* elements (Mahillon and Chandler, 1998), *IS2* and *IS1* classes respectively. These elements are widely distributed among bacterial genomes and due to their mobility they can take part in HGT.

Areas 2, 8, 17, and 19 correspond to four DNA rearrangement hot spots, or *Rhs* elements (Wang et al., 1998). The *Rhs* element is a special GC-rich DNA segment, flanked by AT-rich segments, and it can be linked with one of the *vgr* genes or *IS* element. All *Rhs* elements are highly homologous to each other, what promotes its recombination. In different *E. coli* strains, the number of *Rhs* elements varies from 0 to 7. There are five *Rhs* elements in *E. coli* K-12 under consideration. The remaining element *rhsC* was not included into the list of top anomalous areas due to not sufficient anomalous area density. However, calculations show that it can exceed $\rho_f$ if $k$ is decreased to 35 ($K = 32$), indicating somewhat arbitrary character of density threshold. *Rhs* elements are known to be absent from *Salmonella enterica*, but are present in *Chromobacterium violaceum* (Almeida et al., 2004). Therefore, it is reasonable to suggest that they could be horizontally transferred mobile elements in *E. coli* genome (Sheahan et al., 2004).

Areas 11 and 20 correspond to highly variable *rfb* and *rfa* gene clusters responsible for surface polysaccharides and antigen syntheses, which were reported to be horizontally transferred (Fitzgerald et al., 2003; Heinrichs et al., 1998; Milkman et al., 2003; Mitchison et al., 1997; Tarr et al., 2000). For instance, *rfc* gene, encoding the O-antigen polymerase, has an anomalous codon usage profile as well, and, moreover, can be located outside of the *rfb* gene cluster (Wong et al., 1999). *rfaP* and *rfaB* genes in *E. coli* K-12 have homologues in *Salmonella typhimurium* LT2. However *rfaS*, which is located between these genes, is absent in *S. typhimurium* LT2 (Klena et al., 1993). Additionally, the difference in local GC-content of this DNA area may indicate multiple HGT events (Klena et al., 1992).

Areas 21 and 22 correspond to *ace* operon, involved in glyoxylate cycle, and *phn* operon, which is associated with phosphonate metabolism. The *ace* operon was also reported to be horizontally transferred (Nelson et al., 1997), while genes of the *phn* operon have homologues in *Rhizobium meliloti* (Parker et al., 1999) and *Pseudomonas stutzeri* (White and Metcalf, 2004).

Area 23 corresponds to *nanC*, *fimB*, and *fimE* genes, which code outer membrane channel and two regulators of *fim* operon, respectively. NanC and other proteins of the KdgM family are widely spread in gram-negative bacteria, including *Erwinia carotovora*, *Yersinia pestis*, *S. enterica*, *E. coli*, *Klebsiella pneumoniae*, *Vibrio halioticoli*, and *Pseudomonas* species (Condemine et al., 2005). FimB and FimE are site-specific recombinases that catalyze inversion of *fim* switch (Bryan et al., 2006).

Area 7 corresponds to *ycgF*, coding blue-light sensing proteins that use FAD (BLUF) (Hasegawa et al., 2006). BLUFs are widely distributed in photosynthetic microorganisms such as purple bacteria or cyanobacteria, in which they are involved in blue-light-related cellular signalling processes. However, these proteins have been identified in nonphotosynthetic microorganisms (including *E. coli*) as well, in which they have no known physiological functions.

Area 9 corresponds to *narU*, the first gene in *narUZYWV* operon (Clegg et al., 2006). *narUZYWV* codes the nitrate reductase Z, while *narU* and *narK* code two nitrate transport proteins with DNA sequences identity of ~75%. NarK protein is expressed on rich medium, while NarU is abundant during severe nutrient starvation.

Area 15 corresponds to *tdcA* and *tdcR* genes. These genes code transcriptional regulators of *tdc* operon (Hagewood et al., 1994), which is implicated in amino acid catabolism.

Area 18 corresponds to *yhiM*, a gene of unknown function, which could be involved in a mechanism of acid-stress defense (Bordi et al., 2003). In addition, area 13 is not yet functionally characterized. However, it is noteworthy that near this area a cryptic prophage CP4-57 is annotated. Since tRNA coding genes, including *ileY*, are common sites for alien DNA integration (Lindsey et al., 1989; Ochman et al., 2000), it is possible that this area is also associated with an HGT event.

### 4. CONCLUSION

In this work, we analyzed the GC-content distribution of the *E. coli* K-12 genome. Initially, the global GC-content distribution was obtained by a conventional fixed-length window method. It was found that this
distribution is highly unsymmetrical and significantly deviates from a normal one. It was also shown that the distribution strongly depends on an arbitrary length selection that precludes its use per se in the search for short local DNA areas with anomalous GC-content.

To overcome this problem, a new method of genome segmentation based on the binomial test has been developed. This statistical approach has only one arbitrary parameter, the significance level, and provides reliable estimation of the global GC-content distribution. It was found that, under certain conditions, both types of global GC-content distributions are statistically similar in terms of the nonparametric Mann-Whitney U-test. These allow use of the fixed-window method for estimation of a local GC-content distribution and search for short anomalous GC-content DNA segments.

The minimal size of local DNA areas suitable for analysis by the nonparametric Mann-Whitney U-test was found to be as short as 30 bp. Anomalous GC-content areas were identified by the whole genome U-test scanning. It turned out that approximately 21% of the genome belongs to anomalous GC-content areas distributed all over bacterial genome. The global distribution of anomalous GC-content areas was analyzed, and the top 23 most densely populated areas were identified. Among those, seven areas correspond to annotated prophage, four to \textit{Rhs} elements, and two to \textit{IS} elements. A remaining 10 areas contain putative horizontally transferred DNA and genes with no known function.

**ACKNOWLEDGMENTS**

We thank Prof. V. Lanzov for helpful discussions. This work was supported by the Russian Foundation for Basic Research (grant no. 07-04-00785-a), the Russian Academy of Sciences programs “Physics of Elementary Particles” (subprogram “Neutron physics,” direction “Studies of structure, dynamics and non-ordinary properties of matter by neutron scattering”), by the Ministry of Education and Science of Russian Federation (grant no. RNP.2.2.1.1.4663), the St. Petersburg Scientific Center the Russian Academy of Sciences and the governor of the Leningrad region of Russian Federation.

**DISCLOSURE STATEMENT**

No competing financial interests exist.

**REFERENCES**


Address reprint requests to:

Dr. Andrey Ilatovskiy

Division of Molecular and Radiation Biophysics,

Petersburg Nuclear Physics Institute,

Russian Academy of Sciences,

Orlova roscha,

Gatchina/St. Petersburg, 188300, Russia

E-mail: andreyi@omrb.pnpi.spb.ru