

Translation of the original article

Panyukov V.V. Mathematical Biology and Bioinformatics. 2025;20(2):494–508.

doi: [10.17537/2025.20.494](https://doi.org/10.17537/2025.20.494)

===== TRANSLATIONS OF PUBLISHED ARTICLES =====

A Method for Evaluating the Feasibility of the Chargaff's Second Rule.

Testing the Rule on Bacterial Chromosomes

Panyukov V.V.

*Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics
of Russian Academy of Sciences, Pushchino, Russia*

Abstract. In 1979, Chargaff published a mathematical formulation of his second rule, which states that in each of the complementary strands of two stranded DNA, the amount of adenine is equal to the amount of thymine and the amount of guanine is equal to the amount of cytosine. This rule has now been tested on numerous groups of sequences belonging to various organisms and, with a small number of exceptions, has been confirmed in most cases, but the conclusions about the feasibility of the rule are qualitative rather than quantitative. To our knowledge, no large-scale tests of the rule on bacterial genomes have been carried out. In the proposed work, the rule is tested on 38830 non-degenerate bacterial sequences extracted from the NCBI GenBank 2023 repository. We perform a more complete statistical analysis of the feasibility of the rule in bacterial genomes than is customary in the literature. A new approach to assessing the validity of the rule is proposed.

Key words: *Chargaff's second parity rule, bacterial genomes, Vysochanskij–Petunin inequality*

INTRODUCTION

Formulation of the Chargaff's second parity rule (PR2) rests on the grounds of the biochemical experiments involving chromosomes [1]. It is stated by the rule that for the genomic sequences longer than 100 kb, the amounts of guanine and cytosine are equal, and the amounts of adenine and thymine are equal [2].

When accumulating the sequenced data, testing of PR2 on various sequences resulted in a reliable level of compliance with this rule [3–6]. Further on, and up to now, the rule has been actively tested using the sliding windows technique, in the coding and non-coding regions, on viruses, bacteria, archaea, in human mitochondria and chromosomes [7–18]. The vast majority of the authors involved into testing have reported compliance with the PR2 rule. Still, some deviations from the rule have been detected as well. For instance, in organelles and mitochondria [11, 12], in some bacterial genes [16], as well as in some bacterial genomes [17]. Feasibility of the Chargaff's rule has stimulated the search for certain supporting mechanism [6, 7, 19–22].

Independent of availability of the numerous research papers in this field of studies, there has been no formal numerical criterion obtained to specify feasibility, or non-feasibility, of PR2 to date. This is caused by the fact that such studies evaluate only some very limited sample sequences, and there is no sense of direction in doing research in overall: the 'limited set of sequences; purely qualitative results, flawed quantitative analysis, or lack of

characterization' [9]. It needs to be added here that the methods in use to evaluate feasibility of the Chargaff's rule are not of the statistical kind, which means that applicability of PR2 is not being evaluated by conducting such studies [17].

Consistent with the above, the approach to argumentation in relation to PR2 feasibility in Prabhu's early research [3] shall be reviewed. This method has become typical because of its adaptation by the authors to follow. With 32 organisms in focus, Prabhu provides the G- and C-compositions for the corresponding strands, and reports occurrence of the linear regression between the G- and C-compositions with the coefficient of correlation being close to 1. At that, the regression coefficients have not been provided for some reason. Given the above, Prabhu concludes that PR2 for both G- and C-compositions is feasible in this case.

Our calculations prove that the following (C,G) -regression takes place, namely $C = -1065.80 + 1.023G$, the coefficient of correlation equals to 0.9974. We can see that the slope of the line differs little from one. This indicates that the G- and C-compositions are close. Nonetheless, we believe this information is insufficient enough to conclude that PR2 is feasible. In fact, we note that in this case, $|G - C|$ ranges from 20 to 3296. This proves considerable heterogeneity in the G- and C-compositions in the proposed sequences. Consequently, further data analysis could likely change our opinion on feasibility of PR2 for the given sample.

Application of the linear (G,C) -, (A,T) -regressions, in order to evaluate PR2, is discussed below in a separate section. It has been proven that this method can only function as a qualitative assessment of PR2 feasibility.

The statistical analysis of the Chargaff's rule is yet more informative. For instance, the research papers [17, 23] reveal that in the groups of chromosomes there are samples manifesting significantly different PR2 feasibility. Such studies define the regions of varying deviations from the rule based on the *a-priori* determined level of significance. The author [17] introduces one other additional, to the given one, parameter, and namely, the length of fragments into which the chromosomes are divided, so that the Kolmogorov-Smirnov test can be applied.

In the discussed research, we also find heterogeneity in feasibility of PR2 in the bacterial genome dataset under consideration. This is shown as two regions: the typical and extreme deviations. The so-called threshold zone is seen between them. At that, we do not introduce any *a-priori* values for the parameters, yet use the *a-posteriori* approach. It constitutes in determining the boundaries of the regions of varying degrees of PR2 feasibility on the grounds of the analysis of distribution of certain statistics. The algorithm for the region determination is based on EVT, i.e., the extreme value theory.

We remain hopeful, that the research under discussion will be of practical use when addressing the fundamental question arising from the multiple tastings involving the Chargaff's rule: whether this rule reflects any fundamental law of nature, as Koonin [24] reports, 'or it is just a kind of statistical regularity'.

MATERIALS

In focus of our PR2 testing are 38,830 bacterial chromosomes, free of any degenerate nucleotides (S, W, R, Y, K, M, B, D, H, V, N), with the total length of 150484510033 bp and the lengths ranging from 61247 to 16040666. They are followed by the Ch23 dataset. The nucleotide sequences of Ch23 are retrieved from NCBI GenBank as of October 2023.

The GCAT composition of a chromosome strand is a quadruple (G, C, A, T) of the amount of nucleotides of each species in the sequence $G + C + A + T = sqL$, where sqL is the sequence length.

Testing of the Chargaff's rule is undertaken by examining the values and the areas $|G_s - C_s|$ and $|A_s - T_s|$, where s runs over the chromosomes from Ch23. Table 1 below shows the values of these quantities for eight chromosomes.

Table 1. The nucleotide composition values for eight chromosomes in Ch23

NCBI Id	sqL	G-C		NCBI Id	sqL	A-T
LN890288.1	650317	0		CP011853.1	4428727	0
CP104509.1	5138011	0		CP039716.1	4789195	0
CP098482.1	5369373	173789		CP013196.1	4298263	153117
CP072117.1	8896274	175032		CP040450.1	3741012	200122

Given that the second Chargaff's rule is equivalent to the equations $|G-C| \approx 0$, $|A-T| \approx 0$, we can observe some significant variations in deviation from PR2.

COVENANTS AND NOTATIONS

Let us denote $N = |\text{Ch23}|$, so it represents the quantity of chromosomes within a dataset. The numerical sampling of the N amount is deemed to be a vector, i.e.: $x = (x_1, x_2, \dots, x_N)$. The x symbol, throughout the wording, also represents the reference to an arbitrary component of the vector x , as, for instance, it is shown below, when defining regression.

In order to define the relation between x and y , the linear regression $Y = a + bx$ and the quadratic regression $Y = a + bx + cx^2$ are used, where the coefficients a, b, c are calculated using the least squares technique. The regression error is estimated by the value as follows: $err = 100 \sum_i |1 - Y_i / y_i| / N$.

Normally, the mean and standard deviations of a sample are denoted by μ and σ , $\log() = \log_{10}()$. The p -value, for a sample value of x_i , represents the probability of exceeding the value of x_i , $p(x_i) = \Pr(t \geq x_i)$. Further on, we shall deal with the unimodal samples of x . In such a case, following the Vysochanskij–Petunin inequality [25], the following is generated:

$$p(x) \leq 4/9(1 + z^2), \quad z = (x - \mu) / \sigma.$$

By means of numbering the chromosomes in Ch23, the GCAT-composition of a dataset is generated as a quadruple of $(G, C), (A, T)$, so that $(G_s, C_s), (A_s, T_s)$ demonstrate the GCAT-composition of the selected strand of the s^{th} chromosome. At that, the GCAT-composition of the alternative strand numerically shall be given as $(C_s, G_s), (T_s, A_s)$. Consequently, we have 2^N of probable quadruples of the vectors, representing the GCAT-composition of a dataset. The findings of our research bear no dependence on selecting such a quadruple, in other words they do not depend on the choice of the chromosome strand.

LINEAR REGRESSION

Taking a particular case-study of the eight bacterial chromosomes from the dataset Ch23, we shall discuss the widely cited method of linear regression applied for evaluation of feasibility of the Chargaff's second rule of the AT-composition.

Table 2. The sample consisting of eight chromosomes

Nº	NCBI Id	sqL	A	T	A-T
1	CP040450.1	3741012	1283900	1083778	200122
2	CP013196.1	4298263	1597731	1444614	153117
3	CP098482.1	5369373	1582052	1457550	124502
4	CP090316.1	2611810	688706	567580	121126
5	CP072850.1	1036449	274363	274363	0
6	CP090495.1	3201356	934710	934710	0
7	CP040551.1	3116035	493411	493411	0
8	CP011853.1	4428727	708664	708664	0

For the purposes of evaluating PR2, in pursuance of the acknowledged methodology, it is required to calculate the linear regression $T = a + bA$, where T and A are taken from the corresponding columns of Table 2. Depending on the choice of a strand, in this particular case, we have $2^8 = 256$ table alternatives, and, consequently, the same amount of regressions. It is required to agree on the choice of any particular option. When drafting Table 2, we considered such a strand within the chromosome, for which $A \geq T$. Following the calculation of the coefficients of a and b for Table 1, it would have been reasonable to wonder about the regression coefficients a and b for the remaining 255 tables. There is no need to recourse to the table-by-table examination, yet the obvious option constitutes in regression for an optional table, that is, with the strands similar to the strands of Table 2. It is obvious, that regression for the optional table coincides with the regression of the type $A = a + bT$ for Table 2. The total of the calculations is provided in Table 3 below.

Table 3. Calculations

Sample, chromosomes	Coef. b for $T = a + bA$	Optional coef. b $A = a + bT$	Correlation coefficient	Mean value from $A-T$
1÷4	0.98	1.01	0.99637	149716.75
1÷8	0.8747	1.1246	0.99187	83525.6
1, 5÷8	0.822	1.18	0.985	

The table shows the following, yet undesirable for the PR2 feasibility, properties of the linear regression:

I) Given the information provided by the linear regression calculated for some specific table, and namely the chromosomes 1 to 8, this information does not contribute to identification of the groups of chromosomes, with significantly different degrees of the PR2 feasibility in the sample.

II) Obviously, when including the chromosomes with the zero deviation from PR2 into the sample, the rule feasibility for the extended sample, as presupposed by the average value of $A-T$, has improved. Simultaneously, the changes in the b coefficient demonstrate some increase in deviation of the regression line from the $A = T$ line. Consequently, given this perspective, feasibility of the rule has deteriorated. Some unclear contradiction is present.

III) The correlation coefficient simply reflects the presence of the chromosomes 1 to 8, with the zero deviation from PR2 in the sample.

IV) The bottom row of the table demonstrates some very significant dependence of the b coefficient on the choice of the chromosome strand in forming the AT composition.

V) Absence of the PR2 feasibility criterion makes it impossible to undertake comparison of the degree of the rule feasibility in the chromosomes of different samples.

STATEMENT OF THE PROBLEM

In the statistical framework, the PR2 rule could be defined as a null assumption, which, when applied to the GC-composition, is given as $G = C$. Let us review two options for processing this assumption.

The first option constitutes in evaluation of equivalence of the distributions G and C . In this case the null assumption turns into the $dist(G, C) > \Delta$ assumption, [26], where the Hellinger distance, the total variation, the Kulback–Leibler divergence, and so on, can function as $dist$ [27]. We find this option unacceptable for two reasons. As seen from the possible options, $dist$ is, in fact, the sum of the local distances between the components of (G_s, C_s) , $1 \leq s \leq N$. Their summing up ‘averages out, smooths over’ heterogeneity in the $dist(G_s, C_s)$ distribution, which we would expect to find in the bacterial chromosomes. And

the second reason constitutes that the numerical value of Δ specifies the PR2 feasibility criterion, which, as mentioned above in Introduction, has not been established yet.

We settled on the second option, which is supported by the extreme value theory (EVT) [28–30].

EVT is widely applied to evaluate financial risks and to predict natural disasters that are likely to follow some specific pattern. Ordinary developments of the observed phenomenon shape the right tail of the distribution's *body*, with the exponentially decreasing probability, but without reaching up to any improbable values. At that, the distribution contains *the extreme tail* of the unlikely events, which diminishes at a lesser rate than the exponential curve [31]. The densities of probability for the extreme tails are classified by the Fisher-Tippett-Gnedenko theorem [32].

Our task lies in determination of the threshold values separating the extreme tail from the body for the variations in PR2 feasibility in the GC- and AT-composition samples.

TESTING THE CHAGRAFF'S SECOND RULE

The approaches to determine the threshold values which would divide the distribution body and the extreme tail are widely discussed in the reference sources, with [33–35] among those. Still, the goal of such studies is not the threshold value, but the type of the extreme tail distribution, which is needed for assessing various risks. There is no universally related technique, and in this regard the methods described in these research studies are relevant to the data under consideration therein. Moreover, the approach for determining the threshold depends on the assumption associated with the tail distribution.

In contrast to the said, in order to determine the thresholds, we shall take the distribution body, and not the extreme tail, as the basis. It should be noted right away, that determining a specific threshold value without the a-priori assigned parameters and significance level is, generally speaking, impossible. Consequently, it is only realistic to calculate the interval (hereinafter referred to as the threshold zone) which divides the body and the extreme tail.

Success of the calculation is defined by the statistics that is sensitive to heterogeneity of PR2 performance. Let us define the statistics. We shall take $X = X_1$ or $= X_2$, where

$$X_{1s} = |G_s - C_s| / \sqrt{(G_s + C_s)} \quad \text{and} \quad X_{2s} = |A_s - T_s| / \sqrt{(A_s + T_s)}, \quad 1 \leq s \leq N.$$

The histogram of the X distribution for the dataset Ch23 is unimodal.

It is noteworthy, that if and when evaluating the (G_s, C_s) –, (A_s, T_s) –compositions of the chromosome, and following PR2, the assumption about some random equally probable occurrence of guanine and cytosine alike is put forward, as well as of the equally probable occurrence of adenine and thymine; then X_s^2 represents the distribution of the chi-square type, with one degree of freedom.

Table 4. Statistical parameters for X

X	variation X	variation p -value	μ	σ	$\mu+3\sigma$		variation
X_1	0.0÷198.71	0.00042 – 0.44	4.215	6.011	22.250	$ G-C $	0 – 175032
X_2	0.0÷130.06	0.00047 – 0.44	3.510	4.117	15.860	$ A-T $	0 – 200122

Some very insignificant p -values shall signify presence of the extreme tail in the X statistic distribution.

In order to determine the threshold zones, we shall use the following statistics: $R = R_1$ or $= R_2$:

$$R_1 = \log(X_1) \leq 2.4 \quad \text{and} \quad R_2 = \log(X_2) \leq 2.1.$$

Given the uniformity of the $\log()$ function, the transition from X to R represents, in fact, the transition onto a different scale. Consequently, the histogram of the R distribution appears unimodal as well.

The method we propose for determining the threshold zone is based on the following indication associated with presence of the tail [36]: if $F(x) = \Pr(R \geq x)$, then $\lim_{x \rightarrow \infty} F(x)e^{\varepsilon x} = \infty, \forall \varepsilon > 0$, while the rate of probability diminishment in the right tail area of the body is much higher. It can be summarized by saying that there is no chance to approximate the distribution for R with some single function. This takes place because such a function decreases rapidly, which leads to the unacceptable approximation error at the extreme tail. Choosing the function based on the Fisher-Tippett-Gnedenko theorem, however, yields some feasible approximation of the extreme tail, yet the unacceptable approximation error for the body.

Figure 1 shows the different rates of decrease in probability in the right tail of the R distribution.

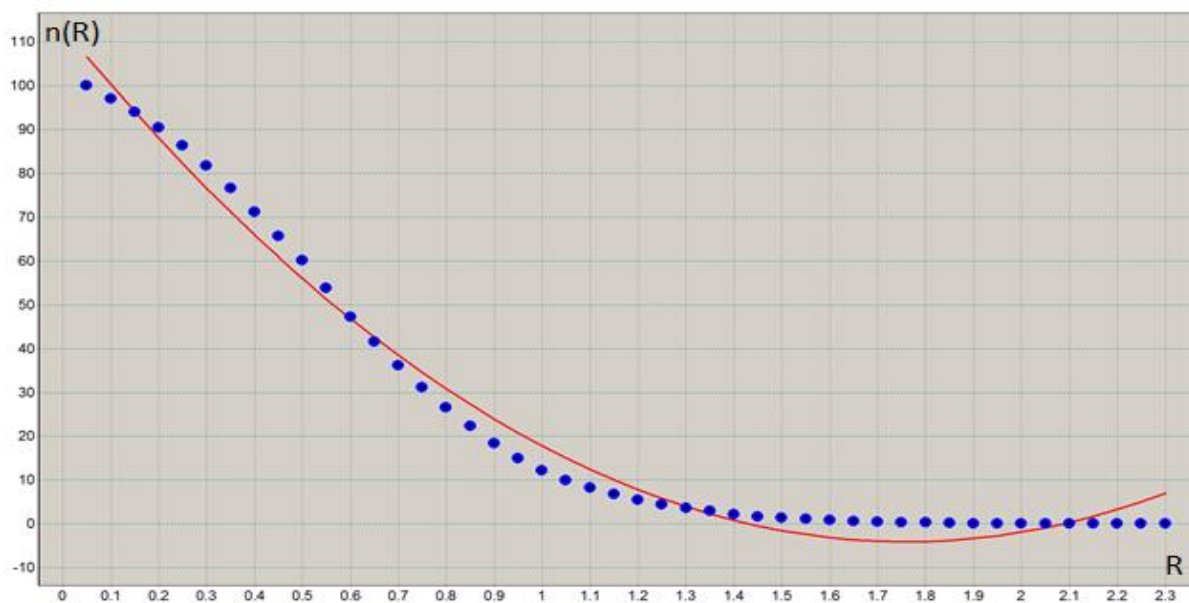


Fig. 1. The function curve $100F(x)$ for the R statistics. The red line represents the quadratic regression of the function, with an error at % 118663, and the fundamentally incorrect approximation of the extremal tail.

The algorithm for calculating the threshold zone implies generation of the truncated distributions for R by sequentially removing the rightmost point from the histogram plotted at the previous stage. The truncated distribution for R is approximated by the quadratic regression. The quality of the regression is assessed, and given that, the decision is taken whether or not to continue the process of calculation.

Further on, the informal presentation of the algorithm is provided by means of demonstration of the sequence of steps underpinning the calculations. Clarifications for the steps are accompanied by the Figures. The purpose of the algorithm is to localize the body in distribution of the R statistics.

As for the current truncated histogram hR of the R statistics, we shall calculate the quadratic regression $reghR$ (the red line). The quality of localization of the distribution body R shall be assessed by means of the approximation error $err\%$, and taking into account the coordinates of the maximum point, $maxAt$, for the regression $regh R$.

We shall review hR in the region of $R \in (0, 2.3] = aR$ (Figure 2) because we are interested in the right tail of the R distribution. The regression point on the graph (Y, X) : $Y = n(R)$ equals to the amount of the events R within the interval $(R - \Delta, R]$, $X = R, \Delta = 0.05$.



Fig. 2. Initiation of the calculations. Plotting of the initial histogram hR , $R \in (0, 2.3]$.

Figure 2 demonstrates stepping of the algorithm at which the decision to remove the rightmost point of the histogram is taken. For the purposes of further consideration, let us consider the case of the initial histogram. We shall calculate $reghR$. Figure 2 proves, that the impact from the extreme tail is quite considerable owing to $err\% = 1318.7$ and $maxAt < 0$. Whenever $maxAt$ lies outside the aR region, then the right-most point of histogram shall be removed. For the case as given, it is $(n(R), R) = (4, 2.3)$; $hR :=$ from hR the point $(4, 2.3)$ is removed.

Stepping of the algorithm, at which the decision to remove the rightmost point of the histogram is taken, is plotted in Figure 3.

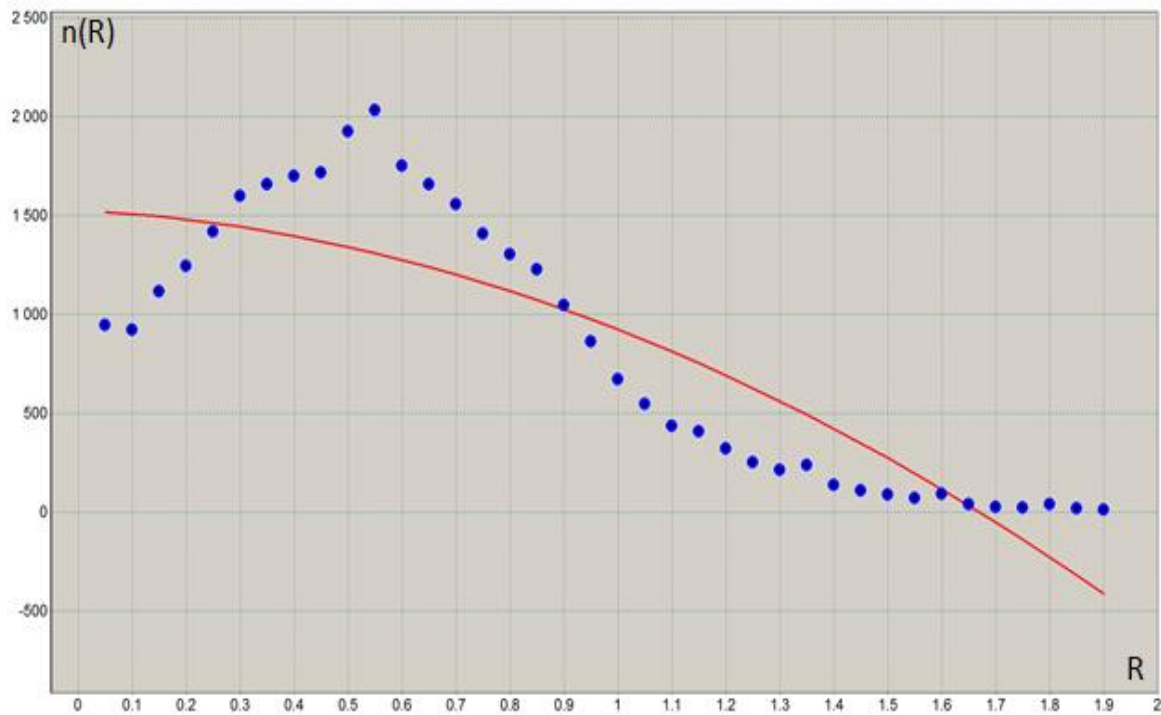


Fig. 3. The current truncated histogram. Some of the rightmost points are removed. Calculation: $reghR = 1527.67 - 141.85R - 462.84R^2$. Evaluation of the quality of regression: $err\% = 265.8$ and $maxAt < 0$. The decisions is, thus, as follows: it is required to remove the rightmost point of the histogram.

Stepping of the algorithm, at which the right boundary of the histogram is determined is plotted in Figure 4.

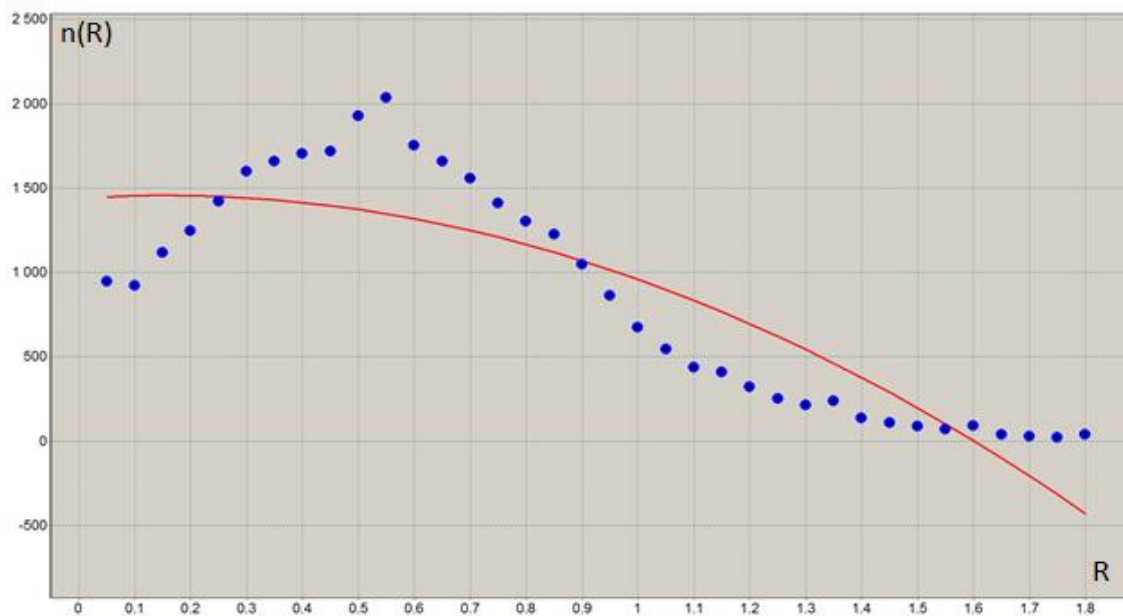


Fig. 4. The current truncated histogram. Calculation of $reghR = 1439.6 + 212.56R - 694.84R^2$, $err\% = 164.84$ and $maxAt = 0.153$ appears for the first time within the histogram region. This is justified by the increased impact from the body within the distribution of the R statistics within this area. The right boundary rtR of the threshold zone is defined, $rtR = 1.8 - \Delta = 1.75$. The value $err\%$ is significant. The body of distribution is defined unsatisfactorily. It is required to proceed further with deletion of the histogram points.

Figure 5 demonstrates stepping of the algorithm at which the body of distribution of the R statistics is localized, and in doing so the left boundary ltR of the threshold zone is determined.

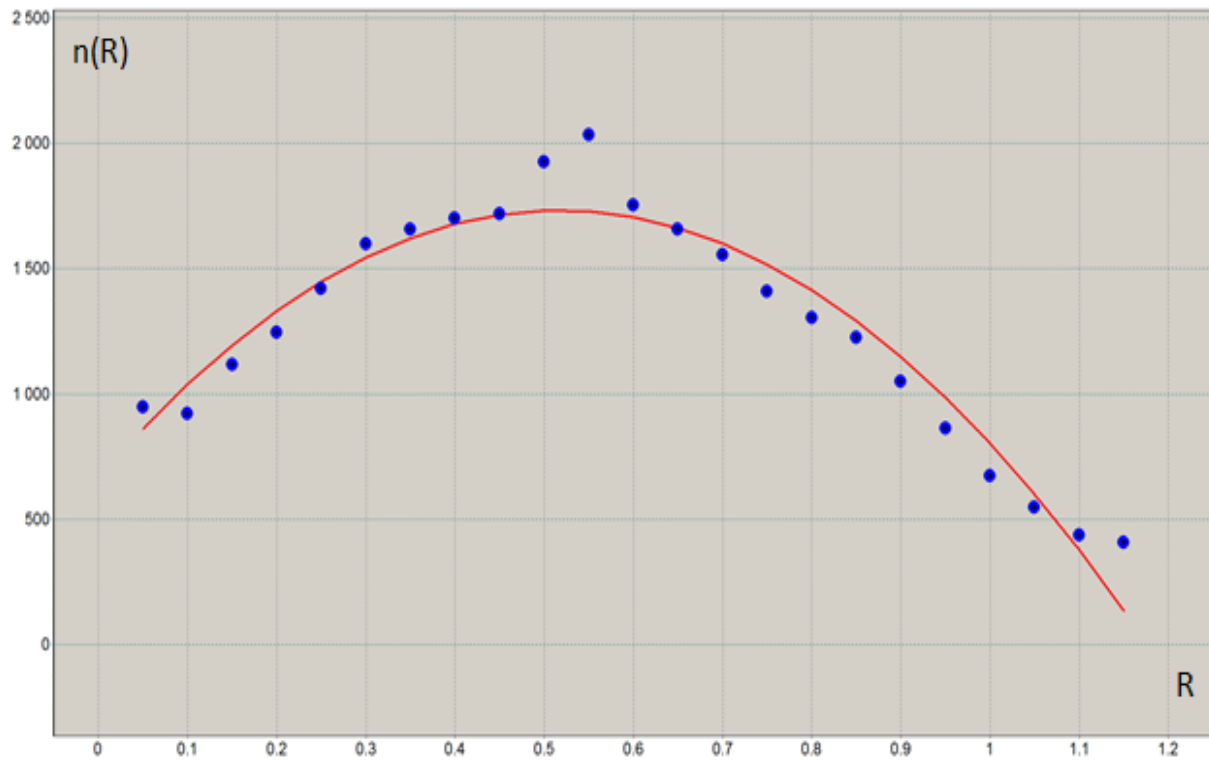


Fig. 5. The current truncated histogram. Calculation: $reghR = 666.23 + 4123.92R - 3987.5R^2$, $err\% = 10.06$ and $maxAt = 0.517$. Satisfactory approximation of the body of the distribution. Checking the stability of the approximation maximum by removing several right-hand points from the histogram (re. Figure 6).

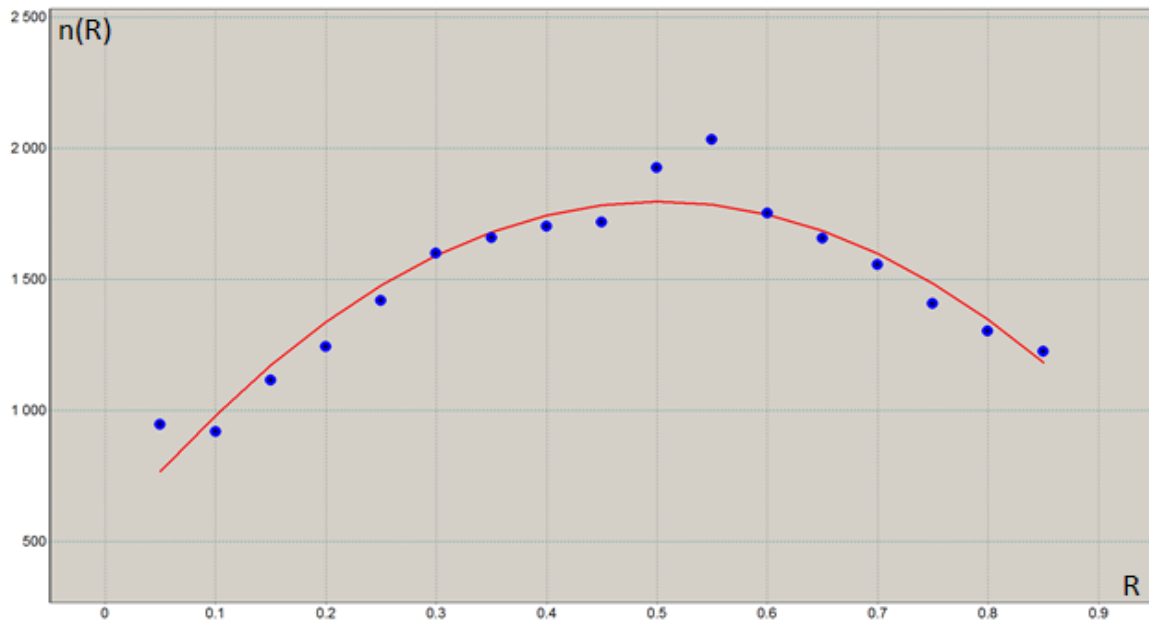


Fig. 6. Checking the stability of the approximation maximum following removal of a number of the right-hand points from the estimated distribution body for the R statistics. Calculation: $reghR = 526.65 + 5062.25R - 5045.87R^2$, $err\% = 5.11$ and $maxAt = 0.502$. Localization of $maxAt$ remains almost unchanged. Accepting: $ltR = 1.15$.

The threshold area of the statistics $R = R_1$ is determined. Similarly, the threshold zone for $R = R_2$ is calculated. In the forthcoming Pictures 7 and 8 the steps fixing the borders of rtR and ltR of the threshold area are marked.

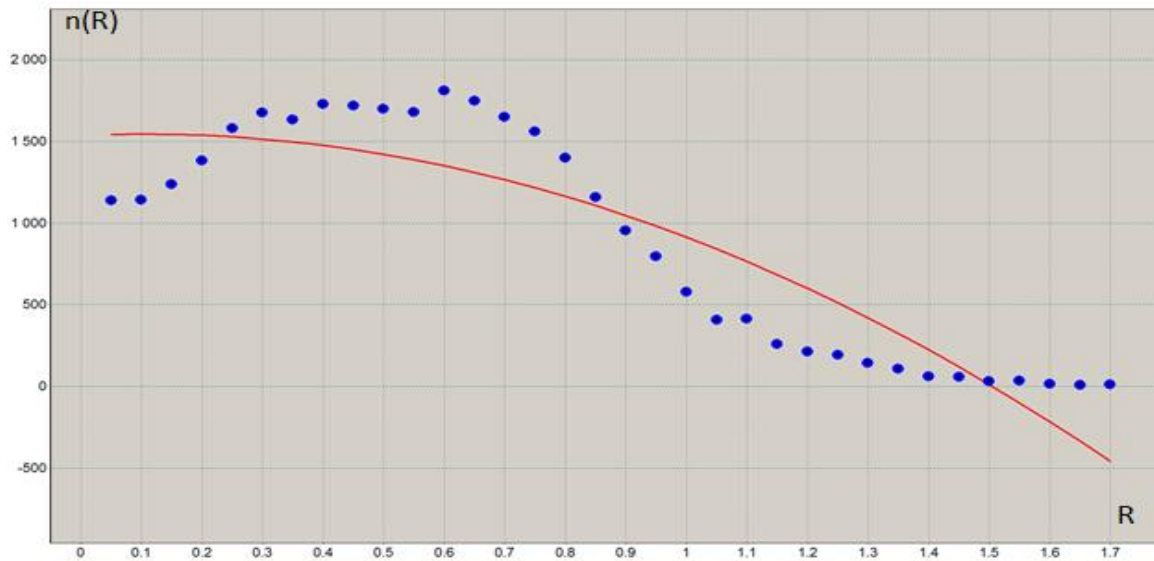


Fig. 7. The current truncated histogram. Calculation: $reghR = 1537.48 + 160.1R - 785.04R^2$, $err\% = 426.15$, $maxAt = 0.101$. The first occurrence of the regression maximum within the histogram area of hR . Accepting: $rtR = 1.7 - \Delta = 1.65$.

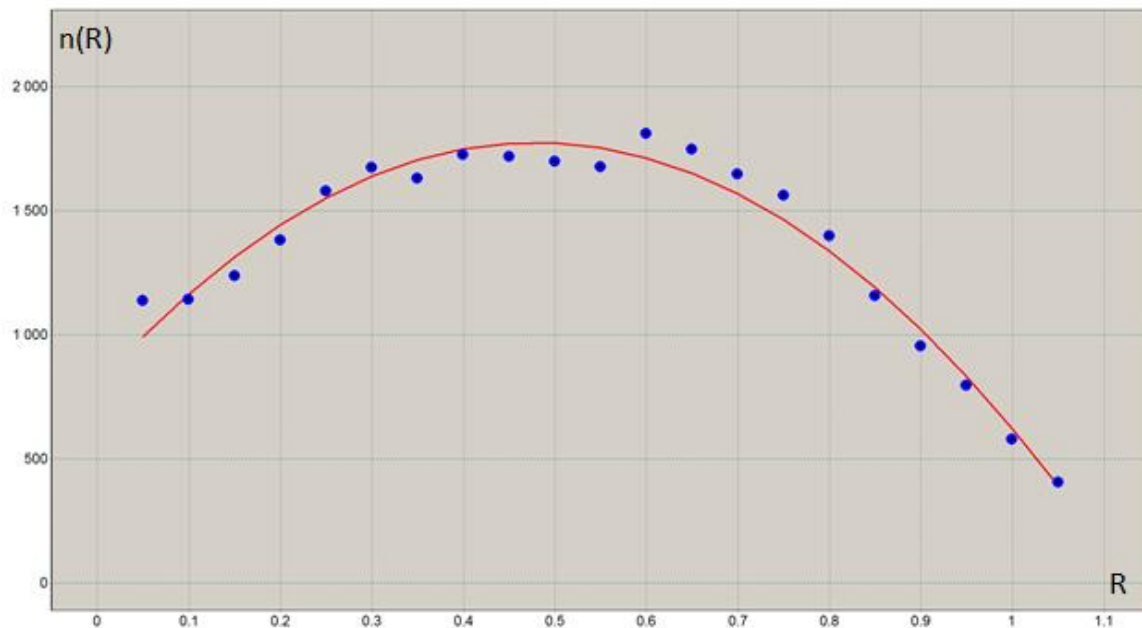


Fig. 8. The current truncated histogram. Calculation: $reghR = 798.71 + 4077.66R - 4254.66R^2$, $err\% = 4.75$, $maxAt = 0.479$. Satisfactory approximation of the body of the distribution. It was verified that localization of the approximation maximum remains almost unchanged when several right points are removed from the histogram. Accepted: $ltR = 1.05$.

The threshold area of the statistics R_2 is determined.

Tables 4 and 5 below provide the summary of the calculations, and which, on the grounds of the statistics R_1 and R_2 , outline the areas of variation of the deviations from the Chargaff's rule.

Table 5. The R_1 statistics. Deviations from PR2

deviation	R_1	X_1	$ G-C $	seqL	total in Ch23
typical	< 1.15	< 14.12	$0 \div 33849$	$61247 \div 16040666$	95.686 %
threshold	$1.15 \div 1.75$	$14.12 \div 56.23$	$2269 \div 109854$	$61238 \div 9857535$	1597
extreme	> 1.75	> 56.23	$13538 \div 175032$	$121924 \div 8896274$	78

Table 6. The R_2 statistics. Deviations from PR2

deviation	R_2	X_2	$ A-T $	seqL	total in Ch23
typical	< 1.05	< 11.22	$0 \div 22723$	$61247 \div 16040666$	96.013 %
threshold	$1.05 \div 1.65$	$11.22 \div 44.67$	$2922 \div 81120$	$61238 \div 9947447$	1519
extreme	> 1.65	> 44.67	$13550 \div 200122$	$121924 \div 5369373$	31

QUALITATIVE ASSESSMENT OF FEASIBILITY OF THE CHARGAFF'S RULE

This section touches upon evaluation of feasibility of the Chargaff's rule involving the linear regression for the (G,C)- and (A,T)-compositions.

For the pair of the vectors (y, x) the (y, x) -regression is calculated: $Y = a + bx$, with the upper regression and the lower regression being distinct.

The upper regression. We shall review the (G,C)-regression. The component (G_s, C_s) , $s \in \text{Ch23}$, depends on the chromosome strand; we shall select the one (the upper one) for which $G_s \geq C_s$, thus marking the obtained (G,C)-composition as the upper pair (G, C) , $G \geq C$. The upper pair shall be obtained in the similar manner (A, T) , $A \geq T$ for the upper (A, T) -regression.

The lower regression comes forth for the pairs (G, C) , $G \leq C$ and (A, T) , $A \leq T$. They are generated by choosing the lower strands that are alternative to the strands for the upper pairs. It is obvious that these regressions can be calculated as follows: the (C,G)- and (T,A)-regressions for the upper pairs.

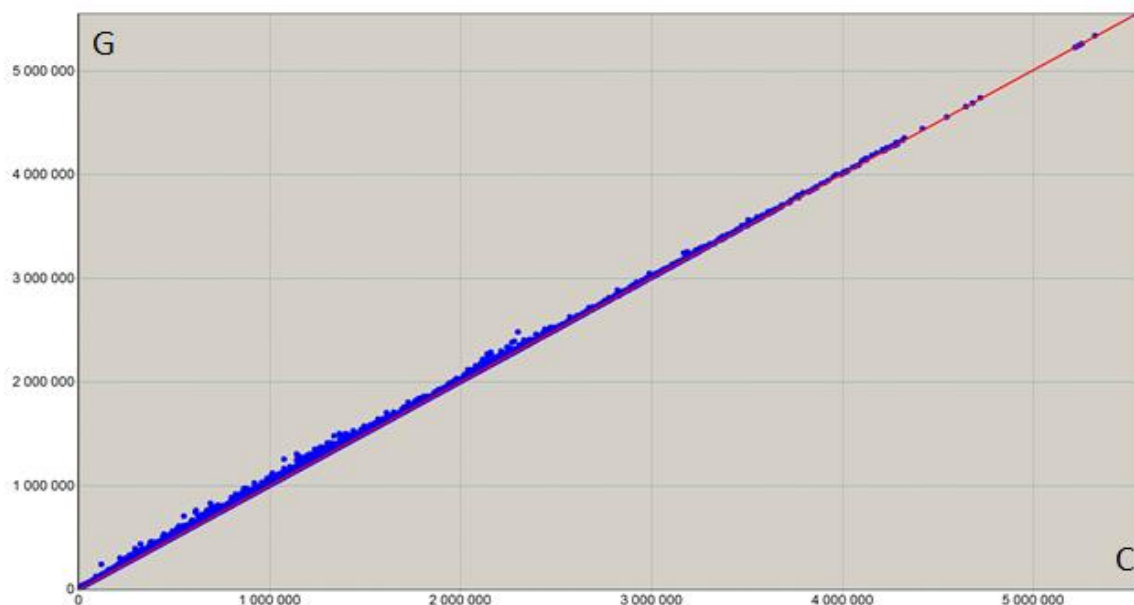


Fig. 9. The upper regression (the red line) for the pair (G, C) , $G \geq C$. $G = 4065.19 + 1.00132C$, $err = 0.63\%$, $corr = 0.99991$.

Figure 9 provides visualization of the calculated regressions.

Table 7. Coefficients of regressions for the chromosome nucleotide compositions

regression	(G,C) upper	(C,G) lower	(A,T) upper	(T,A) lower
<i>a</i>	4065.19	-3883.29	3268.25	-3050.82
<i>b</i>	1.00132	0.99851	1.00136	0.99841
<i>corr</i>	0.99991	0.99991	0.99988	0.99988
<i>err %</i>	0.63	0.65	0.43	0.44

It appears obvious that the quality of implementing the Chargaff's rule should be assessed primarily by closeness of the regression slope coefficient *b* to one.

As seen, the linear regression conceals the extreme deviations from PR2, $b = 1.00132 \approx 1$. These extreme deviations are almost invisible within the regression. This follows from $G = 2992.08 + 1.00123C$, $err = 0.41\%$, $corr = 0.99998$ as calculated for the pair (G,C), $G \geq C$ only for the typical deviations from PR2. Consequently, the sample containing the extreme deviations, as well as the sample with the extreme deviations removed from it, exhibit, to the point, the equal coefficients of the regression slope.

CONCLUSIONS

This research paper proposes the new approach to statistical evaluation of feasibility of the Chargaff's second rule (PR2) in the genomes. This approach is applied for the purposes of assessment of the PR2 stability in the bacterial genomes.

As mentioned in Introduction to this paper, validity of PR2 has been extensively tested for the chromosomes and for the specific chromosomal sites alike, in various organisms.

The overwhelming majority of cases prove agreement with the rule, yet some deviations have been identified as well. We came across only two works in which the statistical analysis of feasibility of the rule was undertaken. These are [17] and [23], where both the regions of agreement with the rule and the regions with the extreme deviations from the rule are identified in the sequence of the studied samples. On the downside, in our opinion, these works apply the subjective parameter, the statistical significance level, with the a-priori assigned value. In this case, the limits of PR2 feasibility represent the function of the significance level. Still in other studies as available to our knowledge, the linear regression is applied to assess feasibility, and which, as we have demonstrated, does not reveal the regions of the rule feasibility, yet merely reflects whether the rule is feasible, or not.

Notwithstanding numerous research papers studying the Chargaff's rule, the formal numerical criterion specifying feasibility or non-feasibility of PR2 has not yet been obtained. This is justified by the fact that such papers discuss very limited samples of the sequences, and there is no general sense of direction in that research.

Our statistical method makes it possible to identify the regions of the varying PR2 feasibility. The calculated boundaries of these regions can function as a criterion for feasibility of the Chargaff's rule. The advantages of this method constitute the a-posteriori approach to establishing the regional boundaries, and the use of the intuitive clear *R* statistic in that regard.

The a-posteriori approach comprises observation of the distribution of the selected *R* statistic. Pursuant to the extreme value theory (EVT), such distribution consists of a body (the underlying distribution) and a tail. The algorithm is proposed that distinguishes between the body in the *R* distribution and, consequently, outlines the boundaries of the different PR2 feasibility.

In order to clarify the intuitive nature of this approach, the research paper [17] is reviewed, in which, for the needs of assessing the feasibility of the $A = T$ rule, the chromosome sequence is divided into the fragments, and the A- and T-compositions are calculated for each fragment. Eventually, the distributions of composition for the

chromosomes are generated. The closeness of such distributions is assessed involving the Kolmogorov–Smirnov criterion. In our opinion, application of this criterion does not provide any intuitive sense of the extent of the rule feasibility.

Evaluation of the feasibility for the Chargaff's second rule as undertaken by us in the sample consisting of 38,830 bacterial chromosomes revealed the region of the typical PR2 feasibility, which comprised 96 % of the chromosomes, as well as the region of the extreme deviation from the rule, which included 78 chromosomes in the case of the G-, C-composition, and 31 chromosomes in the case of the A-, T-composition.

When it comes to the a-priori assigned level of significance, the researcher obtains the precise boundaries for the regions within which the Chargaff's rule is satisfied. This demonstrates considerable volatility, as it is, in principle, impossible to identify the exact value of the statistical quantity.

In the a-posteriori approach applied to determine the limits of the rule feasibility, the area separating the typical and extreme regions emerges naturally. In this research paper, this zone is referred to as the threshold zone. The threshold zone comprised 1,597 chromosomes for the G- and C-compositions, and 1,519 chromosomes for the A- and T-compositions.

In conclusion, the approach as proposed in this research paper represents a fairly efficient tool for identification of the limits of feasibility in relation to the Chargaff's second rule.

REFERENCES

1. Rudner R., Karkas J.D., Chargaff E. Separation of *B. subtilis* DNA into complementary strands. III. Direct analysis. *Proc. Natl Acad. Sci. USA*. 1968. V. 60. P. 921–922. doi: [10.1073/pnas.60.3.921](https://doi.org/10.1073/pnas.60.3.921)
2. Chargaff E. How genetics got a chemical education. *Ann. N. Y. Acad. Sci.* 1979. No. 325. P. 345–360.
3. Prabhu V.V. Symmetry observations in long nucleotide sequences. *Nucl. Acids Res.* 1993. V. 21. P. 2797–2800. doi: [10.1093/nar/21.12.2797](https://doi.org/10.1093/nar/21.12.2797)
4. Albrecht-Buehler Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *PNAS*. 2006. V. 103. No. 47. P. 17828–17833. doi: [10.1073/pnas.0605553103](https://doi.org/10.1073/pnas.0605553103)
5. Fickett J.W., Torney D.C., Wolf D.R. Base compositional structure of genomes. *Genomics*. 1992. V. 13. P. 1056–1064. doi: [10.1016/0888-7543\(92\)90019-O](https://doi.org/10.1016/0888-7543(92)90019-O)
6. Forsdyke D.R. Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *J. Mol. Evol.* 1995. V. 41. P. 573–581. doi: [10.1007/BF00175815](https://doi.org/10.1007/BF00175815)
7. Bell S.J., Forsdyke D.R. Accounting units in DNA. *J. Theor. Biol.* 1999. V. 197. P. 51–61. doi: [10.1006/jtbi.1998.0857](https://doi.org/10.1006/jtbi.1998.0857)
8. Qi D., Cuticchia A.J. Compositional symmetries in complete genomes. *Bioinformatics*. 2001. V. 17. P. 557–559. doi: [10.1093/bioinformatics/17.6.557](https://doi.org/10.1093/bioinformatics/17.6.557)
9. Baisnee P-F., Hampson S., Baldi P. Why are complementary DNA strands symmetric? *Bioinformatics*. 2002. V. 18. P. 1021–1033. doi: [10.1093/bioinformatics/18.8.1021](https://doi.org/10.1093/bioinformatics/18.8.1021)
10. Verma S.K., Das D., Satapathy S.S., Buragohain A.K., Ray S.K. Compositional Symmetry of DNA duplex in bacterial genomes. *Current Science*. 2005. V. 89. P. 374–384. URL: <https://www.jstor.org/stable/24110587> (accessed 02.12.2025).
11. Albrecht-Buehler G., Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *PNAS*. 2006. V. 103. P. 17828–17833. doi: [10.1073/pnas.0605553103](https://doi.org/10.1073/pnas.0605553103)
12. Mitchell D., Bridge R. A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun.* 2006. V. 340. P. 90–94. doi: [10.1016/j.bbrc.2005.11.160](https://doi.org/10.1016/j.bbrc.2005.11.160)
13. Nikolaou C., Almirantis Y. Deviation from Chargaff's second parity rule in organellar DNA insights into the evolution of organellar genomes. *Gene*. 2006. V. 381. P. 34–41. doi: [10.1016/j.gene.2006.06.010](https://doi.org/10.1016/j.gene.2006.06.010)

14. Sueoka N. Intra-strand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* 1995. V. 40 P. 318–25. doi: [10.1007/bf00163236](https://doi.org/10.1007/bf00163236)
15. Sueoka N. Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene*. 1999. V. 238. P. 53–58. doi: [10.1016/s0378-1119\(99\)00320-0](https://doi.org/10.1016/s0378-1119(99)00320-0)
16. Rocha E.P.C., Danchin A., Viari A. Universal replication biases in bacteria. *Mol. Microbiol.* 1999. V. 32. P. 11–16. doi: [10.1046/j.1365-2958.1999.01334.x](https://doi.org/10.1046/j.1365-2958.1999.01334.x)
17. Powdel B.R., Siddhartha Sankar Satapathy, Aditya Kumar, Pankaj Kumar Jha, Alak Kumar Buragohain, Munindra Borah, Suvendra Kumar Ray. A Study in Entire Chromosomes of Violations of the Intra-strand Parity of Complementary Nucleotides (Chargaff's Second Parity Rule). *DNA Research*. 2009. V. 16. No. 6. P. 325–343. doi: [10.1093/dnares/dsp021](https://doi.org/10.1093/dnares/dsp021)
18. Sobottka M., Hart A.G. A model capturing novel strand symmetries in bacterial DNA. *Biochem. Biophys. Res. Commun.* 2011. V. 410. P. 823–828. doi: [10.1016/j.bbrc.2011.06.072](https://doi.org/10.1016/j.bbrc.2011.06.072)
19. Lu J., Salzberg S.L. The Skew Index Test for large-scale GC Skew analysis of bacterial genomes. *PLOS Comput. Biol.* 2020. V. 16. Article No. e1008439. doi: [10.1371/journal.pcbi.1008439](https://doi.org/10.1371/journal.pcbi.1008439)
20. Lobry J.R., Lobry C. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Molecular Biology and Evolution*. 1999. V. 16. No. 6. P. 719–723. doi: [10.1093/oxfordjournals.molbev.a026156](https://doi.org/10.1093/oxfordjournals.molbev.a026156)
21. Forsdyke D.R., Bell S.J. Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Appl Bioinformatics*. 2004. V. 31. P. 3–8. doi: [10.2165/00822942-200403010-00002](https://doi.org/10.2165/00822942-200403010-00002)
22. Forsdyke D.R. Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *J. Mol. Evol.* 1995. V. 41. P. 573–581. doi: [10.1007/BF00175815](https://doi.org/10.1007/BF00175815)
23. Panyukov V.V., Ozoline O.N., Kiselev S.S. Variations in Nucleotide Composition and Deviations from Chargaff's Second Parity Rule in Bacterial Chromosomes. In: *Proceedings of the International Conference "Mathematical Biology and Bioinformatics"*. Ed. V.D. Lakhno. Vol. 10. Pushchino: IMPB RAS, 2024. Paper No. e40. doi: [10.17537/icmbb24.58](https://doi.org/10.17537/icmbb24.58)
24. Koonin E.V. Are there laws of genome evolution?. *PLoS Comput. Biol.* 2011. V. 7. Article No. 1002173. doi: [10.1371/journal.pcbi.1002173](https://doi.org/10.1371/journal.pcbi.1002173)
25. Mercadier M., Strobel F. A one-sided Vysochanskii-Petunin inequality with financial applications. *European Journal of Operational Research*. 2021. V. 295. No. 1. P. 374–377. doi: [10.1016/j.ejor.2021.02.041](https://doi.org/10.1016/j.ejor.2021.02.041)
26. Frey J. An exact multinomial test for equivalence. *The Canadian Journal of Statistics*. 2009. V. 37. No. 1. P. 47–59.
27. Alison L., Gibbs, Francis Edward Su. On Choosing and Bounding Probability Metrics. *International Statistical Review / Revue Internationale de Statistique*. 2002. P. 419–435.
28. Galambos J. *The asymptotic theory of extreme order statistics*. New York: Wiley, 1978. doi: [10.1016/B978-0-12-702101-0.50014-7](https://doi.org/10.1016/B978-0-12-702101-0.50014-7)
29. Embrechts P., Kluppelberg C., Mikosch T. *Modelling Extremal Events for Insurance and Finance*. Springer_Verlag, 1977.
30. Novak S.Y. *Extreme value methods with applications to finance*. Boca Raton. Florida. USA: CRC Press, 2012.
31. Koguta N.S., Rodionov I.V. On tests for distinguishing distribution tails. *Theory of Probability and its Applications*. 2021. V. 66. No. 3. P. 348–363. doi: [10.1137/S0040585X97T990447](https://doi.org/10.1137/S0040585X97T990447)

32. Gnedenko B. Sur la distribution limite du terme maximum d'une s'erie al'eatoire. *Ann. of Math.* 1943. V. 2. No. 44:3. P. 423–453 (in French).
33. Davison A.C., Smith R.L. Models for Exceedances Over High Thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1990. V. 52. No. 3. P. 393–425. doi: [10.1111/j.2517-6161.1990.tb01796.x](https://doi.org/10.1111/j.2517-6161.1990.tb01796.x)
34. Behrens G.N., Lopes H.F., Gamerman D. Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*. 2004. V. 4. P. 227–244. doi: [10.1191/1471082X04st075oa](https://doi.org/10.1191/1471082X04st075oa)
35. Sakthivel K.M., Nandhini V. A New Development of Threshold Selection in Extreme Value Theory. *25th Annual Conference*. 2023. February 15-17. P. 79–92. Special Proceedings: ISBN #: 978-81-950383-2-9.
36. Foss S., Korshunov D., Zachary S. *An Introduction to Heavy-Tailed and Subexponential Distributions*. 2013. Springer Series in Operations Research and Financial Engineering. Springer New York Heidelberg Dordrecht London. doi: [10.1007/978-1-4614-7101-1](https://doi.org/10.1007/978-1-4614-7101-1)

Received 21.10.2025.

Published 02.12.2025.