

Метод оценки выполнимости второго правила Чаргаффа. Тестирование правила на бактериальных хромосомах

Панюков В.В.

Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В. Келдыша Российской академии наук, Пуцино, Россия

Аннотация. В 1979 году Чаргафф опубликовал математическую формулировку своего второго правила, которое заключается в том, что в каждой из комплементарных цепей двух цепочечной ДНК наблюдается равенство количества аденина количеству тимина и равенство количества гуанина количеству цитозина. Это правило, к настоящему времени, проверено на многочисленных группах последовательностей относящихся к различным организмам и, при небольшом количестве исключений, подтверждено в большинстве случаев, но заключения о выполнимости правила носят скорее качественный чем численный характер. Насколько нам известно, масштабных тестирований правила на бактериальных геномах не проводилось. В предлагаемой работе правило тестируется на 38830 бактериальных невырожденных последовательностях извлечённых из рипозитария NCBI GenBank 2023. Мы осуществляем более полный, чем это принято в литературе, статистический, анализ выполнимости правила в бактериальных геномах. Предложен новый подход к оценке справедливости правила.

Ключевые слова: второе правило Чаргаффа, геномы бактерий, неравенство Высочанского – Петунина

ВВЕДЕНИЕ

На основании биохимических экспериментов с хромосомами [1] было сформулировано второе правило Чаргаффа (parity rule, PR2), которое заключается в том, что на геномных последовательностях длиной больше 100 тысяч нуклеотидных пар (н. п.) выполняется равенство количеств гуанина и цитозина и равенство количеств аденина и тимина [2].

С накоплением секвенированных данных, проверка PR2 на различных последовательностях продемонстрировала хороший уровень соответствия этому правилу, [3–6]. Далее и до настоящего времени правило активно тестировалось в скользящих окнах, в кодирующих и не кодирующих областях, в вирусах бактериях, археях, митохондриях и хромосомах человека [7–18]. В большинстве работ авторы тестирования сообщают о соответствии правилу PR2, но наблюдаются и отклонения от правила, например, в органеллах и митохондриях [11, 12], в некоторых генах бактерий [16] и в некоторых бактериальных геномах [17]. Выполнение правила Чаргаффа стимулировало поиск механизма, его поддерживающего [6, 7, 19–22]. Несмотря на обилие работ в данном направлении, формальный численный критерий, характеризующий выполнимость или невыполнимость PR2, не получен и до настоящего времени. Это происходит из того, что в работах изучаются весьма ограниченные по объёму выборки последовательностей и общей направленности в

исследованиях не наблюдается: “limited set of sequences; purely qualitative results, flawed quantitative analysis, or lack of characterization” [9]. Сюда следует добавить, что методы оценки справедливости правила Чаргаффа не являются статистическими, то есть, область применимости правила PR2 в работах не оценивается [17].

В подтверждение сказанного выше рассмотрим метод аргументации выполнения PR2 в ранней работе Prabhu [3], этот метод стал типичным, вследствие его применения последующими авторами. Для 32 организмов автор приводит G– и C–состав для соответствующих нитей и сообщает, что между G– и C–составом имеет место линейная регрессия с коэффициентом корреляции близким к 1, но коэффициенты регрессии при этом не приводятся. На этом основании делается вывод, что в данном случае PR2 для G– и C–состава выполняется. Наши вычисления показывают, что имеет место следующая (G,C)–регрессия $C = -1065.80 + 1.023G$, $corr = 0.9974$. Коэффициент наклона прямой близок к единице, что говорит о близости G– и C–составов. Но этих сведений, по нашему мнению, недостаточно для вывода о выполнимости PR2. В рассматриваемом случае $|G - C|$ варьирует от 20 до 3296, что говорит о существенной неоднородности G–, C–состава в анализируемых последовательностях. Поэтому дополнительный анализ данных необходим для проверки выполнимости PR2 в данной выборке.

Применение линейных (G,C)–, (A,T)–регрессий для оценки PR2 обсуждается нами ниже в отдельном параграфе. Показано, что этот метод может служить качественной оценкой выполнимости PR2.

Статистический анализ правила Чаргаффа более информативен, так работы [17, 23] обнаруживают в изучаемых выборках хромосом группы с существенным различием выполнимости PR2. Области различного отклонения от правила определяются в этих работах на основе априорно назначаемого уровня значимости, а в [17] используется ещё один дополнительный параметр, длина фрагментов, на которые разбивались хромосомы, с целью применения критерия Колмогорова – Смирнова.

В этой работе мы также обнаруживаем неоднородность выполнения PR2 в наборе бактериальных хромосом, что выражается в виде двух областей типичных и экстремальных отклонений, между которыми находится так называемая пороговая зона. При этом мы не вводим априори значений каких-либо параметров, а используем апостериорный подход, который заключается в том, что границы областей с различной степенью выполнимости PR2 определяются на основе анализа распределения некоторых статистик. Алгоритм выделения областей базируется на теории экстремальных величин.

Предлагаемый подход решает основной вопрос, вытекающий из множественных тестирований правила Чаргаффа, а именно, отражает ли это правило фундаментальный закон природы, как пишет Кунин [24], «или является, лишь, своего рода статистической закономерностью».

МАТЕРИАЛЫ

Объектом нашего тестирования правила являются 38830 бактериальных хромосом, не содержащие вырожденные нуклеотиды S, W, R, Y, K, M, B, D, H, V, N с суммарной длиной 150 484 510 033 н.п. и с диапазоном длин 61247–16040666 н.п., далее именуемые набором Ch23. Нуклеотидные последовательности Ch23 извлечены из GenBank по состоянию на октябрь 2023 года.

GCAT-состав нити хромосомы это четвёрка (G,C,A,T) количества нуклеотидов каждого вида в последовательности, $G + C + A + T = sqL$, sqL – длина последовательности.

Тестирование правила Чаргаффа осуществляется посредством изучения величин $|G_s - C_s|$ и $|A_s - T_s|$, где s – номер хромосомы в Ch23. В таблице 1 показаны значения этих величин для четырёх хромосом.

Таблица 1. Величины нуклеотидного состава для восьми хромосом в 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

Так как второе правило Чаргаффа равносильно равенствам $|G-C| \approx 0$, $|A-T| \approx 0$, то мы наблюдаем существенные вариации отклонения от PR2.

ДОГОВОРЁННОСТИ И ОБОЗНАЧЕНИЯ

Обозначаем $N = |\text{Ch23}|$ – количество хромосом в наборе. Численная выборка объёма N рассматривается как вектор $x = (x_1, x_2, \dots, x_N)$, символ x по контексту, также означает ссылку на произвольную компоненту вектора x .

Для установления связи между x и y используется линейная $Y = a + bx$ и квадратичная регрессия $Y = a + bx + cx^2$, где коэффициенты a, b, c вычисляются методом наименьших квадратов. Ошибка регрессии оценивается величиной $err = 100 \sum_i |1 - Y_i / y_i| / N$.

Как обычно, μ и σ обозначаем среднее значение и стандартное отклонение выборки, $\log() = \log_{10}()$. Величина p -value для выборочного значения x_i это вероятность превысить значение x_i , $p(x_i) = \Pr(t \geq x_i)$. Мы будем работать с унимодальными выборками x , в этом случае по теореме Высочанского - Петунина [25]

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

Пронумеровав хромосомы в Ch23, мы получаем GCAT-состав набора в виде четвёрки векторов $(G, C), (A, T)$ так, что $(G_s, C_s), (A_s, T_s)$ есть GCAT-состав выбранной нити s -той хромосомы, при этом GCAT-состав альтернативной нити численно будет таким $(C_s, G_s), (T_s, A_s)$. Таким образом, мы имеем 2^N возможных четверок векторов, представляющих GCAT-состав набора геномов. Результаты нашей работы не зависят от выбора такой четвёрки, то есть, не зависят от выбора нити хромосомы.

ЛИНЕЙНАЯ РЕГРЕССИЯ

На конкретном примере из восьми бактериальных хромосом набора Ch23 рассмотрим метод линейной регрессии для оценки выполнимости второго правила Чаргаффа AT-состава.

Для оценки PR2, в соответствии с принятой методикой, нужно вычислить линейную регрессию $T = a + bA$, где T и A берутся из соответствующих столбцов таблицы 2. В зависимости от выбора нити мы, в данном случае, имеем $2^8 = 256$ вариантов таблиц и, следовательно, столько же регрессий. Нужна договорённость о выборе варианта. При формировании таблицы 2 мы брали в хромосоме ту нить у которой $A \geq T$. После вычисления коэффициентов a и b для таблицы 2 естественно поинтересоваться, а каковы коэффициенты регрессии a и b для остальных 255-ти таблиц. Перебор не имеет смысла, но вариант, который напрашивается, это регрессия для альтернативной таблицы, то есть, с нитями, альтернативными нитям таблицы 2. Очевидно, что

регрессия для альтернативной таблицы совпадает с регрессией вида $A = a + bT$ для таблицы 2. Совокупность вычислений собрана в нижеследующей таблице 3.

Таблица 2. Выборка из восьми хромосом

№	NCBI Id	seqL	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

Таблица 3. Вычисления

Выборка, хромосомы	Коэф. b для $T = a + bA$	Альтерн. коэф. b $A = a + bT$	Коэффициент корреляции	Среднее $A-T$
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	

Таблица 3 демонстрирует следующие, нежелательные для оценки PR2, свойства линейной регрессии:

I) По информации, которую предоставляет линейная регрессия, вычисленная для конкретной таблицы, хромосомы 1–8, не позволяет выявить в выборке групп хромосом с существенно различной степенью выполнимости PR2.

II) Видно, что, при включении в выборку хромосом с нулевым отклонением от PR2, выполнимость правила расширенной выборкой, судя по среднему, от величины $A-T$ улучшилось. В то же время изменения коэффициента b говорят об увеличении отклонения линии регрессии от прямой $A=T$, то есть, с этой точки зрения выполнимость правила ухудшилась. Налицо непонятное противоречие.

III) Коэффициент корреляции практически не отражает наличие в выборке 1–8 хромосом с нулевым отклонением от PR2.

IV) Последняя строка таблицы демонстрирует весьма заметную зависимость коэффициента b от выбора нити хромосомы при формировании AT -состава.

V) Отсутствие критерия выполнимости PR2 не позволяют сравнивать степень выполнимости правила в различных выборках хромосом.

ПОСТАНОВКА ЗАДАЧИ

В рамках статистики, правило PR2 может быть сформулировано как нулевая гипотеза, которая, применительно к GC-составу имеет вид $G=C$. Обсудим два варианта работы с этой гипотезой.

Один вариант это оценка эквивалентности распределений G и C . В этом случае нулевая гипотеза переходит в гипотезу $dist(G, C) > \Delta$, [26], где в качестве $dist$ могут выступать расстояния Хеллингера, полной вариации, Кúльбака - Лёйблера и другие [27]. Этот вариант мы находим неприемлемым по двум причинам. Как видно из возможных вариантов, $dist$ – это, вообще говоря, сумма локальных расстояний между компонентами (G_s, C_s) , $1 \leq s \leq N$. А суммирование «усредняет, затушёвывает» неоднородность в распределении $dist(G_s, C_s)$, которую мы ожидаем в бактериальных

хромосомах. И вторая причина – это то, что численное значение Δ задаёт критерий выполнимости PR2, который, как говорилось во введении выше, ещё не установлен.

Мы остановились на другом варианте основанном на технике экстремальных значений EVT, [28–30].

EVT широко используется для оценки финансовых рисков и для предсказания природных катаклизмов, которые протекают по следующей схеме. Обычные события наблюдаемого явления образуют правый хвост *тела* распределения с экспоненциально убывающей вероятностью, не достигая при этом маловероятных значений, в то же самое время в распределении имеется *экстремальный хвост* из маловероятных событий, который спадает гораздо медленнее экспоненты [31]. Плотности вероятности экстремальных хвостов классифицируются теоремой Гнеденко – Фишера – Типпетта [32].

Мы ставим своей задачей определение пороговых значений, отделяющих экстремальный хвост от тела при вариации выполнимости PR2 в выборках GC– и AT–состава.

ТЕСТИРОВАНИЕ ВТОРОГО ПРАВИЛА ЧАРГАФФА

Способы определение пороговых значений, разделяющих тело распределения и экстремальный хвост, широко представлены в литературе [33–35]. Но целью таких работ является не величина порога, а вид распределения экстремального хвоста, что необходимо для оценки разного рода рисков. Общего метода здесь нет и поэтому методы работ соотносятся с рассматриваемыми данными, при этом способ определения порога зависит от гипотезы относительно распределения хвоста.

В противоположность сказанному, мы, для определения порогов, берём за основу не экстремальный хвост, а тело распределения. Сразу заметим, что определить конкретную величину порога, без априори назначаемых параметров типа уровня значимости, вообще говоря, невозможно. Поэтому мы вычисляем интервал (далее: пороговая зона), который разделяет тело и экстремальный хвост.

Успех вычисления обусловлен статистикой, чувствительной к неоднородности выполнения PR2. Определим статистики X :

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

Гистограммы распределения этих статистик по набору Ch23 унимодальны.

Заметим, что, если при измерении (G_s, C_s) –, (A_s, T_s) –состава хромосомы, следуя PR2, выдвинуть гипотезу о случайном равновероятном появлении как гуанина и цитозина, так и о равновероятном появлении аденина и тимина, то X_{1s}^2 и X_{2s}^2 имеют χ^2 -распределение с одной степенью свободы.

Таблица 4. Параметры статистик X

X	вариация X	вариация p-value	μ	σ	$\mu + 3\sigma$		вариация
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

Весьма малые значения p-value в таблице 4 говорят о наличии экстремального хвоста в распределении статистики X .

Для определения пороговых зон будем использовать статистики R :

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

В силу монотонности функции $\log()$, переход от X к R – это переход к другому масштабу, поэтому гистограмма распределения R также унимодальна.

Предлагаемый нами метод определения пороговой зоны опирается на следующий признак наличия хвоста [36]: если $F(x) = \Pr(R \geq x)$, то $\lim_{x \rightarrow \infty} F(x)e^{\varepsilon x} = \infty, \forall \varepsilon > 0$, в то время, как скорость уменьшения вероятности в области правого хвоста тела гораздо выше. Это приводит к тому, что невозможно аппроксимировать распределение для R одной функцией, поскольку такая функция – либо быстро убывающая, и тогда на экстремальном хвосте получается неприемлемая ошибка аппроксимации. Если же выбрать функцию, основываясь на теореме Гнеденко – Фишера – Типпетта, то это даст удовлетворительную аппроксимацию экстремального хвоста, но неприемлемую ошибку аппроксимации тела.

Рисунок 1 демонстрирует разные скорости падения вероятности в правом хвосте распределения R .

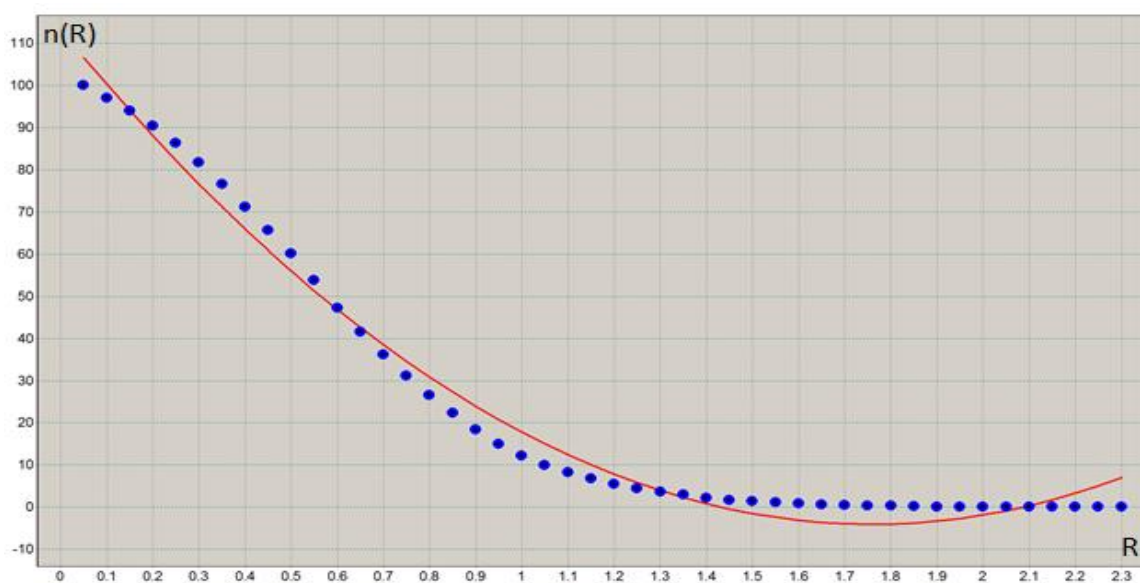


Рис. 1. График функции $100F(x)$ для статистики R . Красная линия – это квадратичная регрессия функции с ошибкой 118663 % и с принципиально неправильной аппроксимацией хода экстремального хвоста.

Алгоритм вычисления пороговой зоны заключается в формировании усеченных распределений для R посредством последовательного удаления крайней правой точки из той гистограммы, которая получена на предыдущем шаге. Усечённое распределение для R аппроксимируется квадратичной регрессией. Оценивается качество регрессии и на основании этого принимается решение о продолжении процедуры вычисления.

Далее предлагается неформальное изложение алгоритма посредством демонстрации ряда шагов, на которых базируется суть вычислений. Пояснение шагов сопровождается рисунками. Цель алгоритма локализовать тело в распределении статистики R .

Для текущей усечённой гистограммы hR статистики R вычисляем квадратичную регрессию $reghR$ (красная линия). Качество локализации тела распределения R оцениваем посредством ошибки аппроксимации $err\%$, с учётом координаты точки максимума $maxAt$ у регрессии $reghR$.

Поскольку нас интересует правый хвост у распределения R , то рассматриваем hR в области $R \in (0, 2.3] = aR$ (рисунок 2). Точка регрессии на графике (Y, X) : $Y = n(R)$ равно количеству событий R в интервале $(R - \Delta, R]$, $X = R, \Delta = 0.05$.



Рис. 2. Инициация вычислений. Вычислена начальная гистограммы hR , $R \in (0, 2.3]$.

На рисунке 2 демонстрируется шаг алгоритма, в котором принимается решение по удалению крайней правой точки гистограммы. Рассмотрим на примере начальной гистограммы. Вычисляем $reghR$. На рисунке 2 влияние экстремального хвоста очень сильное ввиду $err\% = 1318.7$ и $maxAt < 0$. Если $maxAt$ лежит за пределами области aR , удаляем крайнюю правую точку гистограммы, в данном случае это $(n(R), R) = (4, 2.3)$, hR : =из hR удалить $(4, 2.3)$.



Рис. 3. Текущая усечённая гистограмма. Часть правых точек удалена. Вычисляем $reghR = 1527.67 - 141.85R - 462.84R^2$. Оценка качества регрессии: $err\% = 265.8$ и $maxAt < 0$. Решение таково: следует удалить крайнюю правую точку гистограммы.

Шагу алгоритма, на котором принимается решение по удалению крайней правой точки гистограммы, соответствует рисунок 3.

Шаг алгоритма, в котором определяется правая граница пороговой зоны, представлен на рисунке 4.

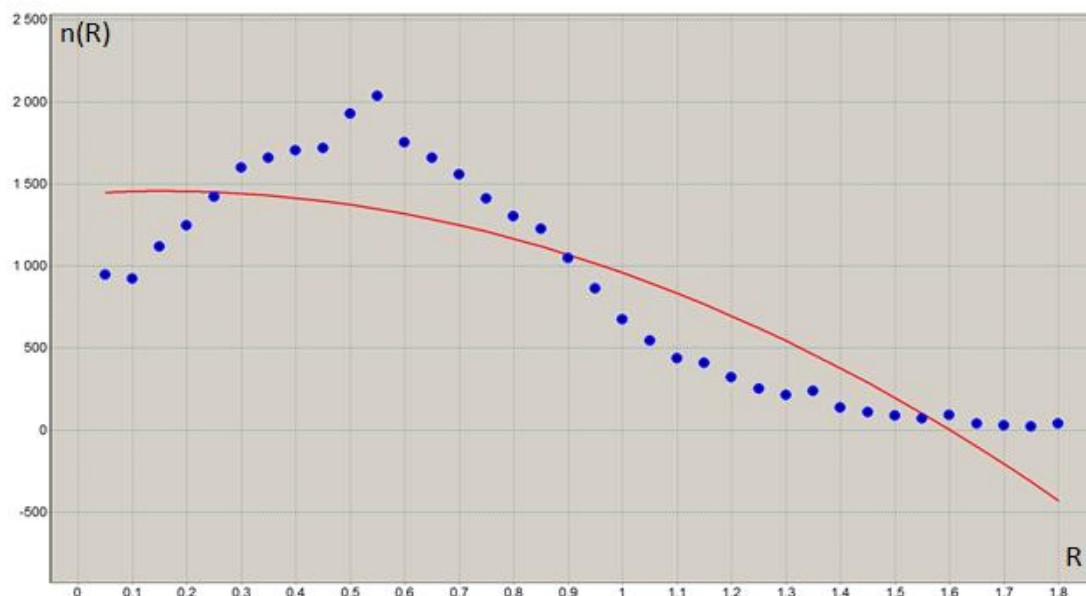


Рис. 4. Текущая усечённая гистограмма. Вычисляем $reghR = 1439.6 + 212.56R - 694.84R^2$, $err\% = 164.84$ и $maxAt = 0.153$ впервые появляется в области гистограммы, что обусловлено возросшим вкладом тела в распределение статистики R в данной области. Определяем правую границу rtR пороговой зоны, $rtR = 1.8 - \Delta = 1.75$. Величина $err\%$ значительна. Тело распределения определено неудовлетворительно. Следует продолжить удаление точек гистограммы.

Шаг алгоритма, на котором производится локализация тела распределения статистики R и тем самым определяется левая граница ltR пороговой зоны, рисунок 5.

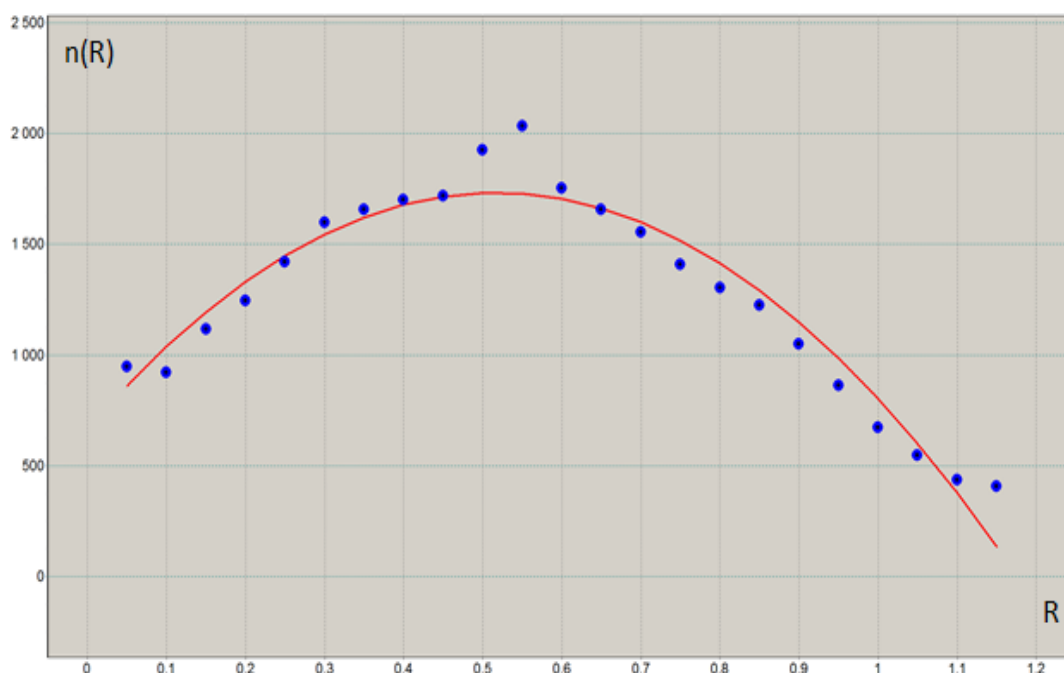


Рис. 5. Текущая усечённая гистограмма. Вычисляем $reghR = 666.23 + 4123.92R - 3987.5R^2$, $err\% = 10.06$ и $maxAt = 0.517$. Удовлетворительная аппроксимация тела распределения. Проверяем стабильность положения максимума аппроксимации, посредством удаления несколько правых точек из гистограммы (см. рисунок 6).

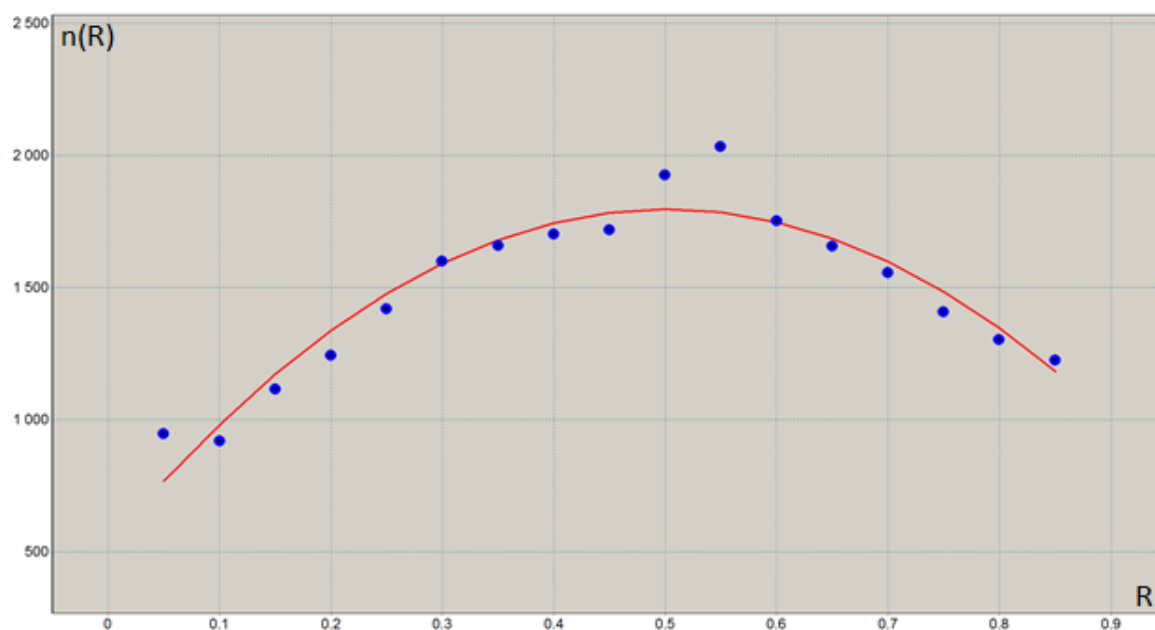


Рис. 6. Проверка стабильности положения максимума аппроксимации, после удаления ряда правых точек из предполагаемого тела распределения у статистики R . Вычисляем $reghR = 526.65 + 5062.25R - 5045.87R^2$, $err\% = 5.11$ и $maxAt = 0.502$. Положение $maxAt$ практически не поменялось. Принимаем $ltR = 1.15$.

Пороговая зона статистики $R = R_1$ определена. Аналогичным образом вычисляем пороговую зону для $R = R_2$. На следующих двух рисунках 7 и 8 показаны шаги фиксирования границ rtR и ltR пороговой зоны.

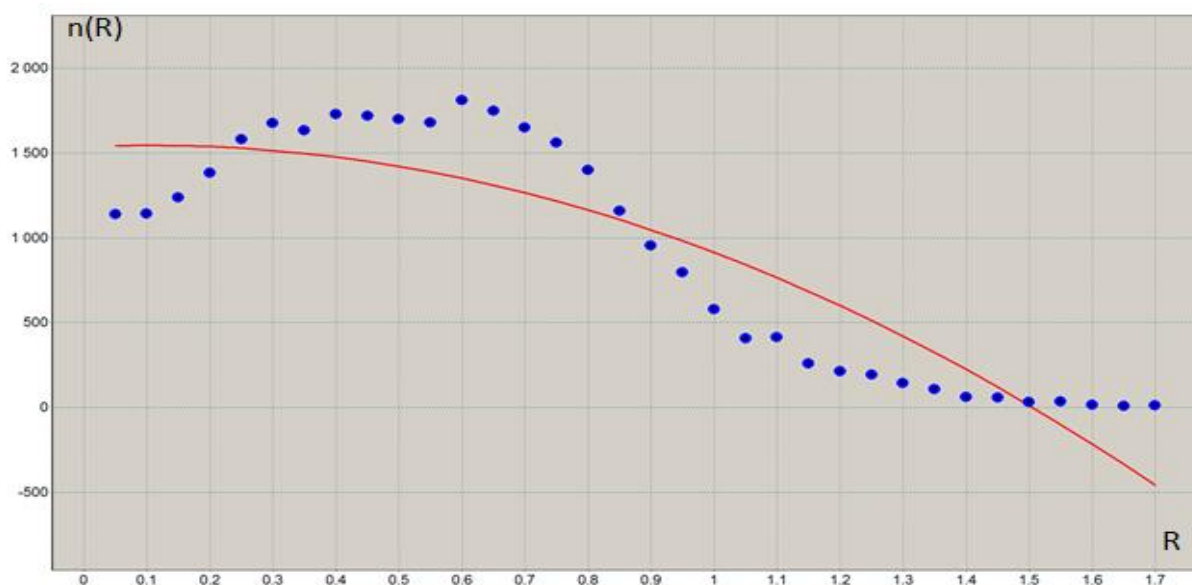


Рис. 7. Текущая усечённая гистограмма. Вычисляем $reghR = 1537.48 + 160.1R - 785.04R^2$, $err\% = 426.15$, $maxAt = 0.101$. Первое появление максимума регрессии в области гистограммы hR . Принимаем $rtR = 1.7 - \Delta = 1.65$.

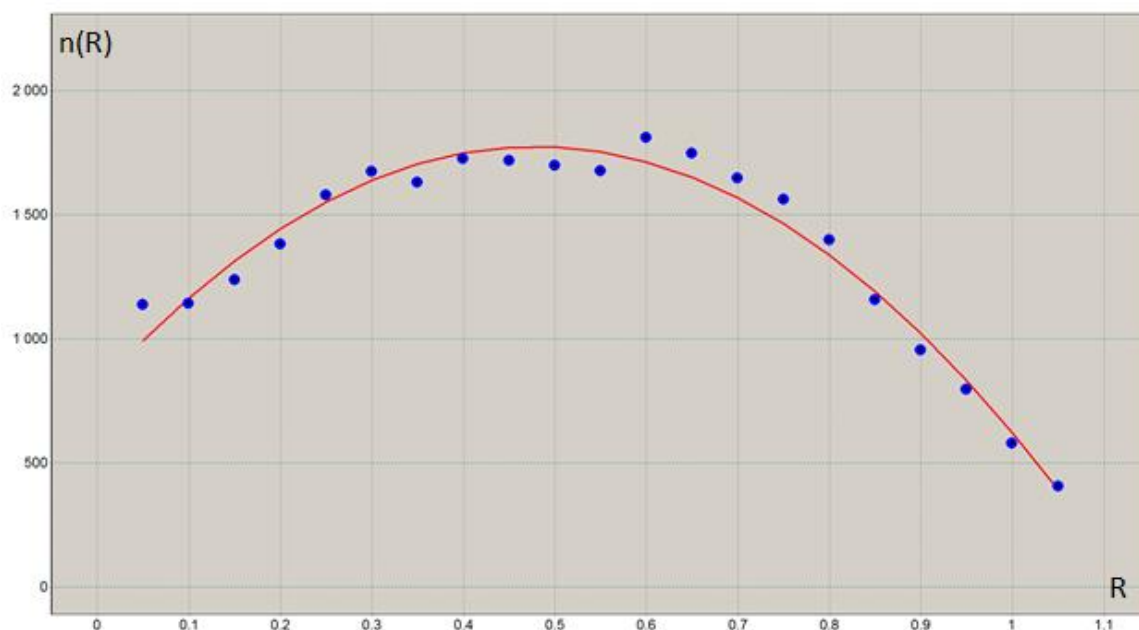


Рис. 8. Текущая усечённая гистограмма. Вычисляем $reghR = 798.71 + 4077.66R - 4254.66R^2$, $err\% = 4.75$, $maxAt = 0.479$. Удовлетворительная аппроксимация тела распределения. Проверено, что положение максимума аппроксимации практически не меняется при удалении нескольких правых точек из гистограммы. Принимаем $ltR = 1.05$.

Пороговая зона статистики R_2 определена.

Сводка вычислений представлена в ниже следующих таблицах 5 и 6, в которых, на основе статистик R_1 и R_2 , представлены области вариации отклонений от правила Чаргаффа.

Таблица 5. Статистика R_1 . Отклонения от PR2

отклонение	R_1	X_1	$ G-C $	$seqL$	КОЛ-ВО В Ch23
типичное	< 1.15	< 14.12	$0 \div 33849$	$61247 \div 16040666$	95.686 %
пороговое	$1.15 \div 1.75$	$14.12 \div 56.23$	$2269 \div 109854$	$61238 \div 9857535$	1597
экстремальное	> 1.75	> 56.23	$13538 \div 175032$	$121924 \div 8896274$	78

Таблица 6. Статистика R_2 . Отклонения от PR2

отклонение	R_2	X_2	$ A-T $	$seqL$	КОЛ-ВО В Ch23
типичное	< 1.05	< 11.22	$0 \div 22723$	$61247 \div 16040666$	96.013 %
пороговое	$1.05 \div 1.65$	$11.22 \div 44.67$	$2922 \div 81120$	$61238 \div 9947447$	1519
экстремальное	> 1.65	> 44.67	$13550 \div 200122$	$121924 \div 5369373$	31

КАЧЕСТВЕННАЯ ОЦЕНКА ВЫПОЛНИМОСТИ ПРАВИЛА ЧАРГАФФА

В этом разделе мы оценим выполнимость правила Чаргаффа посредством линейной регрессии для (G,C)– и (A,T) –составов.

Для пары векторов (y, x) вычисляется регрессия $Y = a + bx$, различаем верхнюю и нижнюю регрессии.

Верхняя регрессия. Рассмотрим (G, C) –регрессию. Компонента (G_s, C_s) , $s \in \text{Ch23}$, зависит от нити хромосомы, выбираем нить (верхнюю) у которой $G_s \geq C_s$, помечая полученный (G, C) –состав как верхнюю пару (G, C) , $G \geq C$. Аналогично формируем верхнюю пару (A, T) , $A \geq T$ для верхней (A, T) –регрессии.

Нижняя регрессия возникает для пар (G, C) , $G \leq C$ и (A, T) , $A \leq T$, которые формируются на основе выбора *нижних* нитей, которые альтернативны нитям для верхних пар. Очевидно, что эти регрессии можно вычислять как (C, G) – и (T, A) –регрессии для верхних пар.

Таблица 7. Коэффициенты регрессий для нуклеотидных составов хромосом

регрессия	(G, C) верхняя	(C, G) нижняя	(A, T) верхняя	(T, A) нижняя
a	4065.19	–3883.29	3268.25	–3050.82
b	1.00132	0.99851	1.00136	0.99841
$corr$	0.99991	0.99991	0.99988	0.99988
$err \%$	0.63	0.65	0.43	0.44

Представление о вычисленных регрессиях даёт нижеследующий рисунок 9.

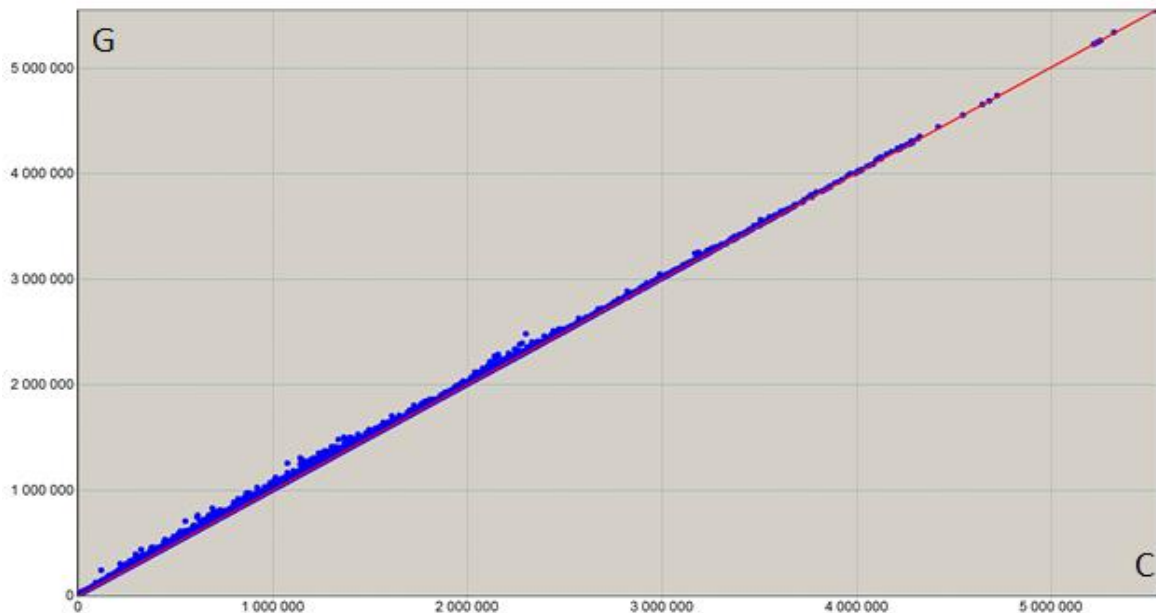


Рис. 9. Верхняя регрессия (красная линия) для пары (G, C) , $G \geq C$. $G = 4065.19 + 1.00132C$, $err = 0.63\%$, $corr = 0.99991$.

Качество выполнения правила Чаргаффа следует оценивать близостью коэффициента наклона линии регрессии b к единице.

Мы видим, что линейная регрессия маскирует экстремальные отклонения от PR2, $b = 1.00132 \approx 1$. Эти экстремальные отклонения в регрессии, практически, себя никак не проявляют, что вытекает из $G = 2992.08 + 1.00123C$, $err = 0.41\%$, $corr = 0.99998$ вычисленной для пары (G, C) , $G \geq C$ только для типичных отклонений. То есть, как выборка с экстремальными отклонениями, так и выборка с удалёнными экстремальными отклонениями имеют, практически, равные коэффициенты наклона линии регрессии.

ЗАКЛЮЧЕНИЕ

В работе предлагается новый метод статистической оценки выполнимости второго правила Чаргаффа (PR2) в геномах. Этот метод применён для оценки выполнимости PR2 в бактериальных геномах.

Справедливость PR2 проверялась рядом авторов как у целых хромосом, так и у отдельных сайтов различных организмов. В подавляющем большинстве случаев наблюдается согласие с правилом, но обнаружены и отклонения. Мы нашли только две работы где проводится статистический анализ выполнимости правила это [17] и сообщение [23], где в исследуемых выборках последовательностей выявляются как области согласия с правилом так и области с экстремальными отклонениями от правила. Недостатком этих работ является использование субъективного параметра, статистического уровня значимости, с априорно назначаемой для него величиной и, в таком случае, границы выполнимости PR2 оказываются функцией от уровня значимости. В остальных, известных нам работах, для оценки выполнимости используется линейная регрессия. Этот метод не выявляет областей выполнения правила а лишь сигнализирует о выполнении или невыполнении правила.

Несмотря на обилие работ по изучению правила Чаргаффа, формальный численный критерий, характеризующий выполнимость или невыполнимость PR2, не получен до настоящего времени. Это связано с тем, что в работах изучаются весьма ограниченные по объёму выборки последовательностей и общей направленности в исследованиях не наблюдается.

Наш статистический метод позволяет выделить области различной выполнимости PR2. Вычисляемые границы областей могут служить в качестве критерия выполнимости правила Чаргаффа. Достоинствами метода являются апостериорный подход к установлению границ областей и использование интуитивно понятной статистики R .

Апостериорный подход состоит в том, что мы наблюдаем распределение выбранной статистики R , которое согласно теории экстремальных величин EVT, состоит из тела (основное распределение) и хвоста. Предложен алгоритм, который выделяет тело в распределении R и, тем самым, определяет границы различной выполнимости PR2.

Для пояснения интуитивности рассмотрим работу [17], где для оценки выполнимости правила $A = T$ последовательность хромосомы разбивается на фрагменты и для каждого фрагмента вычисляется A -, T -состав, в результате для хромосомы получаем распределения составов. Близость этих распределений оценивается критерием Колмогорова – Смирнова. Использование этого критерия, на наш взгляд, не даёт интуитивного ощущения степени выполнимости правила.

Проведённая нами оценка выполнимости второго правила Чаргаффа в выборке из 38830 бактериальных хромосом выявила область типичного выполнения PR2, куда вошло 96 % хромосом, и область экстремального отклонения от правила куда вошло 78 хромосом в случае G -, C -состава и 31 хромосома в случае A -, T -состава.

В случае априорно назначаемого уровня значимости исследователь получает точные границы областей степени выполнения правила Чаргаффа. Это дань значительной условности, поскольку точное значение статистической величины определить, в принципе, невозможно.

В апостериорном подходе, который мы используем для определения границ выполнения правила, естественным образом возникает зона, разделяющая типичную и экстремальную области, в данной работе называемая пороговой зоной. В пороговую зону вошло 1597 хромосом в случае G -, C -состава и 1519 хромосом в случае A -, T -состава.

Таким образом, предлагаемый в данной работе метод является достаточно эффективным инструментом для выявления границ выполнимости второго правила Чаргаффа.

СПИСОК ЛИТЕРАТУРЫ

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 G. 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 26.09.2025).
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. Панюков В.В., Озолин О.Н., Киселев С.С. Вариации нуклеотидного состава и отклонения от второго правила Чаргаффа в бактериальных хромосомах. *Доклады Международной конференции “Математическая биология и биоинформатика”*. Под ред. В.Д. Лахно. Т.10. Пушино: ИМПБ РАН. 2024. Статья № е40. 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. Галамбош Я. *Асимптотическая теория экстремальных порядковых статистик*. М.: Наука, 1984.
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. Когут Н.С., Родионов И.В. О критериях различения хвостов распределений, *Теория вероятн. и ее примен.* 2021. Т. 66. № 3. С. 433–453.
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.
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.

Рукопись поступила в редакцию 14.09.2025, переработанный вариант поступил 30.09.2025.
Дата опубликования 09.10.2025.

A Method for Evaluating the Feasibility of 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, Vyschanskij–Petunin inequality*