УДК 573.22 + 575.89

ВТОРОЕ ПРАВИЛО ЧАРГАФФА И СИММЕТРИЯ ГЕНОМОВ

[1]Гребнев Я.В., [2]Садовский М.Г.

[1]ФГАОУ ВПО «Сибирский федеральный университет», Институт фундаментальной биологии и биотехнологии, Красноярск, e-mail: yaroslav.grebnev@gmail.com;

[2]ФГБУН «Институт вычислительного моделирования» Сибирского отделения Российской академии наук, Красноярск, e-mail: msad@icm.krasn.ru

В работе представлены предварительные результаты исследования нарушения суперсимметрии в геномах различных организмов. Под суперсимметрией понимается т.н. второе правило Чаргаффа, устанавливающее равенство частот олигонуклеотидов, читающихся одинаково в противоположных направлениях, с учётом замены нуклеотидов по правилу комплементарности. Представлены предварительные результаты исследования геномов различных организмов и органелл, от вирусов до высших животных. Вычислялась мера нарушения симметрии в пределах одного стренда ДНК для указанных геномов. Всего проанализировано более 1500 последовательностей, проведён сравнительный анализ данных, полученных для разных последовательностей: как внутригеномный, так и межгеномный. Установлено, что внутригеномная вариабельность меры нарушения второго правила Чаргаффа падает по мере роста длины слов, для которых она определяется, и сопоставима с межгеномной вариабельностью для коротких слов.

Ключевые слова: палиндромы, частота, классификация, корреляция, таксономия, эволюция

CHARGAFF'S SECOND RULE AND SYMMETRY IN GENOMES

[1]Grebnev Y.V., [2]Sadovskiy M.G.

[1]Siberian Federal University, Institute of Fundamental Biology and Biotechnology, Krasnoyarsk, e-mail: yaroslav.grebnev@gmail.com;

[2]Institute of Computational Modeling of Siberian Branch of Russian Academy of sciences, Krasnoyarsk, e-mail: msad@icm.krasn.ru

Some preliminary results are provided towards the study of the violation of genomic super-symmetry; that latter is the so called Second Chargaff's rule. The rule stipulates that oligonucleotides that could be read equally in opposite directions with respect to the symbol change according to the complimentary law (complimentary palindromes) should exhibit pretty close frequency. We have checked the genomes of organisms of various taxa ranging from viruses via bacteria, yeasts, animals, plants, etc.; more than 1500 genetic sequences had been studied, totally. The measure for the second rule violation was calculated for a single strand. Both intragenomic, and intergenomic studies have been carried out. It was found that intragenomic variability decays, as the length of string grows up. The intergenomic variability is comparable to the intragenomic one, for considerably short strings.

Keywords: palindromes, frequency, classification, correlation, taxonomy, evolution

Первое правило Чаргаффа устанавливает равенство в молекуле ДНК количества тимина (T) и количества аденина (A), а также соответствующее равенство для гуанина (G) и цитозина (C). Позже было установлено, что аналогичное правило выполняется и для одного стренда ДНК; данное равенство было названо вторым правилом Чаргаффа. Нарушение второго правила Чаргаффа зависит от длины анализируемого участка генома и может характеризовать сам геном. Для целой хромосомы высших эукариот характерная ошибка в (приблизительных) равенствах $A \approx T$, $G \approx C$ составляет $10^{-3}...10^{-2}$. Правила Чаргаффа – это универсальные правила, и им подчинены геномы всех организмов от низших растений до высших животных, не являются исключением в том числе и внеклеточные формы жизни.

В настоящее время исследованию нарушения второго правила Чаргаффа посвящено не так много работ [1–4], несмотря на фундаментальный характер этого факта. Основная цель настоящей работы – оценка степени нарушения второго правила Чаргаффа в геномах различных организмов.

Материалы и методы исследования

В настоящей работе производилось исследование поведения невязки для геномов различных организмов. Геномы организмов представляли собой расшифрованные тексты нуклеотидных последовательностей, которые были взяты из EMBL банка данных (http://www.ebi.ac.uk/genomes/). В работе использовались геномы различных организмов и вирусов; проанализированы следующие последовательности геномов: дрожжи 81 061 875 нуклеотидов, грибы 269 875 059 нуклеотидов, бактерии 36 358 967, митохондрии 243 981, вирусы 29142, а также такие организмы, как *Giberella moniliformis* 41 104 290, комар 230 466 657, бык 2 629 841 282, 21 хромосома человека 33 216 610, 1 хромосома макаки 232 296 185, *Arabidopsis taliana* 93 654 490,

дрозофилы 125 566 102, шимпанзе 106 544 938 и гориллы 9 140.

Для определения невязки составлялись частотные словари последовательностей. Частотный словарь – это множество всех символьных подпоследовательностей заданной длины, встречающихся в изучаемой последовательности, вместе с указанием частоты их встречаемости [5–7]. В рамках настоящей работы составлялись частотные словари толщины от 1 до 8 (т.е. содержащие слова длины 1, 2, …, 8).

Показателем, характеризующим степень нарушения второго правила Чаргаффа, была величина

$$\mu = \frac{2}{4^q}\sqrt{\sum_{\Omega^* \in \omega}\left(f_\omega - f_{\bar{\omega}}\right)^2}, \qquad (1)$$

где q – длина слов в рассматриваемом словаре; Ω – множество всех слов, являющихся прямыми; ω – слово; $\bar{\omega}$ – комплементарное слово; f_ω – частота прямого слова; $f_{\bar{\omega}}$ – частота комплементарного слова.

Для анализа поведения величины невязки (1) оценим её поведение для случайной нескоррелированной последовательности. Предположим, что в ней точно выполняется второе правило Чаргаффа:

$$p(A) = p(T); \quad p(C) = p(G). \qquad (2)$$

Тогда $\forall q$ невязка (1) равна нулю и второе правило выполняется с абсолютной точностью.

Пусть теперь соотношение (2) выполняется не точно, а с некоторой погрешностью:

$$p(A) = p_W + \varepsilon;\ p(T) = p_W - \varepsilon;$$

$$p(C) = p_S + \delta;\ p(G) = p_S - \delta. \qquad (3)$$

Здесь $2p_W = p(A) + p(T)$ и $2p_S = p(C) + p(G)$. Тогда невязка для словаря толщины $q = 1$ определяется выражением $\mu = 2(\varepsilon + \delta)$. Для частотных словарей при $q = 2$ имеем три комбинации типов нуклеотидов для всех мыслимых слов: WW ⇔ WW, SS ⇔ SS и SW ⇔ SW; здесь буквы обозначают слабые и сильные нуклеотиды (т.е. $W = \{A, T\}$ и $S = \{C, G\}$), а порядок не важен. Для случаев WW ⇔ WW и SS ⇔ SS возможны по три случая невязки (1) с каждой стороны палиндрома (т.е. для каждого из слагаемых в скобках в (1):

$$p_w^2 - \varepsilon^2; \quad p_w^2 + 2p_w\varepsilon + \varepsilon^2 \quad \text{и} \quad p_w^2 - 2p_w\varepsilon + \varepsilon^2, \qquad (4)$$

и аналогично

$$p_s^2 - \delta^2; \quad p_s^2 + 2p_s\delta + \delta^2 \quad \text{и} \quad p_s^2 - 2p_s\delta + \delta^2. \qquad (5)$$

Отбрасывая члены порядка ε^2, δ^2 и выше, получаем оценку для каждого палиндрома вида

$$\frac{\max\{\varepsilon, \delta\}}{2}.$$

Поскольку общее число палиндромов в частотном словаре составляет $0{,}5\times 4^q$, постольку окончательное выражение для оценки величины невязки (1) определяется выражением

$$\mu \cong \frac{\max\{\varepsilon, \delta\}}{2^q}, \qquad (6)$$

где q – толщина словаря.

Результаты исследования и их обсуждение

В таблице представлены результаты вычислений показателей нарушения второго правила Чаргаффа для различных организмов. Видно, что наибольшее количество нарушений второго правила Чаргаффа наблюдается для митохондрий *Equus caballus* breed Appalosa, далее по степени нарушения второго правила Чаргаффа следуют митохондрии парнокопытных, что еще раз свидетельствуют о том, что в митохондриях различных геномов происходит наибольшее количество нарушений второго правила Чаргаффа. Затем по степени нарушения второго правила Чаргаффа следуют геномы высших животных, в частности геном гориллы, далее можно выделить геномы внеклеточных форм жизни, в частности геном вируса табачной мозаики. Следующими нарушителями второго правила Чаргаффа являются грибковые организмы аскомицеты, затем – насекомые. Наименьшее нарушение второго правила Чаргаффа наблюдается для растений.

Данные таблицы свидетельствуют об экспоненциальном убывании невязки (1) с ростом толщины словаря для различных таксономических групп. Следует отметить, что вариабельность невязки при малых значениях толщины словаря весьма велика, но с ростом толщины словаря она падает, что согласуется с оценкой (6), произведённой выше. Наибольшее количество нарушений второго правила Чаргаффа среди исследованных нами организмов наблюдалось у митохондрий и внеклеточных форм жизни.

Справедливость оценки (6) подтверждается также рисунком, на котором показан ход значений отношения двух последовательных значений невязки (1), полученной для той или иной группы геномов. Хорошо видно, что по мере роста толщины словаря (при приближении толщины к $q = 8$) отношение двух последовательных значений невязки (1) стремится к значению, равному двум, что полностью согласуется с оценкой (6). По-видимому, можно ожидать, что точность приближения этого отношения к 2 будет лишь возрастать по мере роста толщины частотных словарей, взятых в рассмотрение.

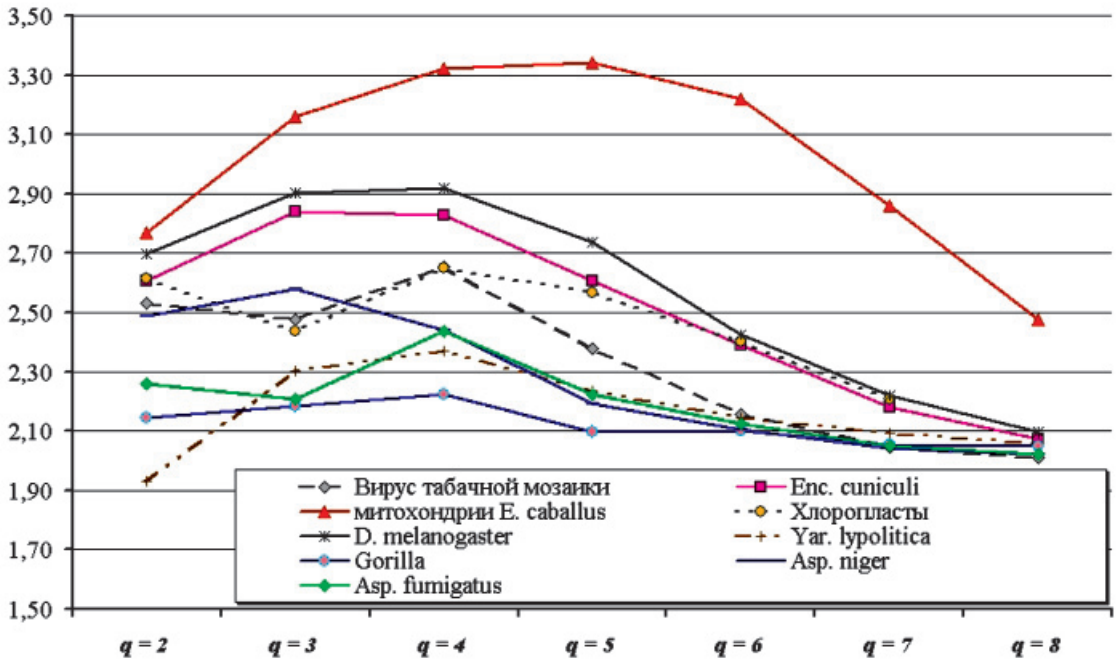
Особого внимания заслуживает более детальное изучение поведения самой невязки (1) для сравнительно малых значений q: $1 \le q \le 4$. Если стремление отношения двух последовательных значений невязки при росте q можно объяснить в том числе и эффектами конечности исследуемой символьной последовательности: действительно, число различных слов в частотном словаре растёт экспоненциально, что ведёт к быстрому падению числа тех слов, которые встречаются более чем в одной копии, – то поведение невязки на сравнительно малых

длинах слов ($1 \le q \le 4$) скорее всего отражает биологические особенности анализируемых генетических последовательностей в наибольшей степени.

Величина невязки (1) для различных организмов; N_1 – количество исследованных организмов; N_2 – количество исследованных нуклеотидов, млн. пар; μ_{min} – минимальное значение невязки; μ_{max} – максимальное значение невязки; $\langle\mu\rangle$ – среднее значение невязки σ_μ – стандартное отклонение

Организм	N_1	N_2	$\mu_{min}\cdot10^6$	$\mu_{max}\cdot10^2$	$\langle\mu\rangle\cdot10^2$	$\sigma_\mu\cdot10^3$
Ashbya gossypi	7	9095747	4,1767600	0,74800000	0,0508000	0,474000
Aspergilus fumigatus	8	29384958	2,5095100	0,15261870	0,0163000	0,155000
Aspergilus niger	19	33975768	2,9484400	0,55338040	0,0380000	0,246000
Aspergilus nudilans	8	29828291	2,5080100	0,18392510	0,0193000	0,102513
Candida albicans	9	12061552	3,1211000	0,28585600	0,0328000	0,149000
Candida dublinesis	8	14618422	3,2505800	0,18711890	0,0225000	0,133100
Candida glabrata	13	12318245	4,8216400	0,39477620	0,0437000	0,183616
Cryptococcus JEC21	14	19051922	3,6409500	0,30546400	0,0270489	0,159054
Fusarium oxysporum	15	57720560	3,6954100	0,27849940	0,0257665	0,159516
Giberella moniliformis	11	41104290	4,5612800	0,15603360	0,0200162	0,0096240
Giberella zeae	4	36358967	3,3075800	0,05998620	0,0136769	0,0013179
Kluweromyces lactis	6	10689156	3,4477200	0,28333580	0,0366872	0,159453
Lachancea kluyveri	8	10394259	3,8284100	0,41620820	0,0460825	0,236253
Lachancea thermotolerns	8	9705144	4,2600000	0,17290000	0,0242000	0,103000
Pichia	8	15441179	3,0831100	0,24417120	0,0276118	0,124950
Schizosaccharamyces pombe	3	12495682	2,3659500	0,13228910	0,0193245	0,0077696
Schizosaccharamyces pombe (штамм 2)	3	12571820	2,3591500	0,13612510	0,0180149	0,0088554
Yarrowia lypolitica	6	20502981	2,8100000	0,13650000	0,0167000	0,0070900
Zygosacharamicues	7	9764635	4,2600000	0,23900000	0,0363000	0,1140000
Anofeles	5	230466657	3,6101500	0,05241630	0,0104760	0,0035282
Arabidopsis	6	93654490	1,5500000	0,14300000	0,0212090	0,7727100
Drosophila melanogaster	6	125566102	1,1100000	0,61600000	0,0312000	0,4350000
Cryptococcus B3501A	14	19699782	3,6800000	0,32400000	0,0298000	0,1720000
Bos taurus	5	81698	48,700000	9,90000000	1,8821000	1,1882100
Equus caballus	57	8993004	50,200000	11,8000000	2,2284000	2,2284000
Eremothecium gossypii	7	9119312	4,1800000	0,74800000	0,0508000	0,5080000
Gorilla	10	9140	136,00000	10,3000000	1,0436000	10,436000
Encephalitozoon cuniculi	11	2497519	9,7000000	1,63000000	0,1360000	1,3600000
African cassava mosaic virus	3	8273	105,00000	4,23000000	0,7672000	0,7938000
Хлоропласты	463	$\approx 4\cdot10^9$	13,600000	7,99000000	0,1528000	0,0011460
Митохондрии (различных организмов)	2004	$\approx 9\cdot10^9$	26,400000	26,9000000	2,1318000	10,403000

Представленные в статье результаты также косвенно опровергают одну из гипотез [1–4] происхождения второго правила Чаргаффа, а именно гипотезу удвоений. Согласно этой гипотезе, второе правило Чаргаффа возникло в результате серии удвоений длинных и сверхдлинных участков ДНК. При этом сама по себе последовательность, которая подвергалась удвоениям, предполагалась близкой по своим свойствам к случайной.



Отношение двух последовательных (по толщине словаря) значений невязки для различных таксономических групп

Однако оценка (6) показывает, что для случайной последовательности – при условии почти точного выполнения второго правила Чаргаффа на уровне мононуклеотидного состава – второе правило Чаргаффа выполняется и для слов большей длины. Более того, можно ожидать, что длинные и сверхдлинные повторы будут приводить скорее к нарушениям второго правила Чаргаффа, по крайней мере в среднем по всей последовательности; возможно возникновение заметной гетерогенности последовательности по показателю невязки (1), определяемому для разных фрагментов исходной последовательности, однако этот вопрос выходит за рамки настоящей работы.

Список литературы/References

1. Albrecht-Bühler G. Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences // Genomics, 2008, vol. 90, pp. 297–305.

2. Nikolaou C, Almirantis Y. Deviations from Chargaff›s second parity rule in organellae DNA Insights into the evolution of organellar genomes // Gene, 2006; 381:34-41.

3. Mitchell D. GC content and genome length in Chargaff compliant genomes. // Biochem. Biophys. Res. Commun. 2007; 353(1):207–10.

4. Rapoport A.E., Trifonov E.N. Compensatory nature of Chargaff›s second parity rule. // J. Biomol. Struct. Dyn. 2013; 31(11):1324–36.

5. Sadovsky M.G. Information capacity of nucleotide sequences and its applications // Bulletin of Math. Biol., 2006, vol. 68, no. 2, pp. 156–178.

6. Sadovsky M.G., Shchepanovsky A.S., Putintzeva Yu.A. Genes, Information and Sense: Complexity and Knowledge Retrieval // Theory in Biosciences, 2008, vol. 127, pp. 69–78.

7. Gorban A.N., Popova T.G., Sadovsky M.G. Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy // Open Syst. & Information Dyn., 2000, v.7, no. 1, pp. 1–17.

Рецензенты:

Садовский В.М., д.ф.-м.н., профессор, зав. лабораторным отделом вычислительной механики деформируемых сред Института вычислительного моделирования СО РАН, г. Красноярск;

Кратасюк В.А., д.б.н., профессор, зав. кафедрой биофизики ИФБиБТ СФУ, г. Красноярск.

Работа поступила в редакцию 19.12.2014.