***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 9. Пущино: ИМПБ РАН, 2022. Статья № e33. doi: 10.17537/icmbb22.46*

Длинные прямые повторы в бактериальных хромосомах

Панюков В.В.[1], Озолинь О.Н.[2], Киселев С.С.[2]

[1]*Институт математических проблем биологии РАН – филиал Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»*
[2]*Институт биофизики клетки РАН – обособленное подразделение Федерального государственного бюджетного учреждения науки «Федеральный исследовательский центр «Пущинский научный центр биологических исследований Российской академии наук»*

panyukov@itaec.ru

Как мелкие, так и крупномасштабные геномные перестройки (дупликации, инверсии и делеции) могут возникать в результате гомологичной рекомбинации между повторами в нуклеотидной последовательности. В работе проведено исследование распределения точных максимальных прямых повторов в 23748 хромосомах бактерий, не содержащих вырожденные нуклеотиды, и в 3 независимых наборах негеномных последовательностей. Обнаружено, что бактериальные хромосомы сильно отличаются от негеномных последовательностей по длинам максимальных повторов. У 33 штаммов, преимущественно у *Bordetella pertussis* и *Escherichia coli*, присутствуют прямые повторы длиной более 100 тысяч пар нуклеотидов, в то время как в негеномных последовательностях их длина не превышает 32 пар нуклеотидов. Показано отсутствие корреляции значений длины наибольшего повтора с длиной хромосомы и GC-составом. Предложено уточнение нижнего порогового уровня для классификации геномов по длине наибольшего повтора. С помощью неравенства Высочанского–Петунина определена граница значимости для длины максимального повтора (640 нуклеотидных пар).

Ключевые слова: *длинные прямые повторы, геномы бактерий, неравенство Высочанского – Петунина.*

Long Direct Repeats in Bacterial Chromosomes

Panyukov V.V.[1], Ozoline O.N.[2], Kiselev S.S.[2]

[1]*Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences*
[2]*Institute of Cell Biophysics of the Russian Academy of Sciences*

Both small and large-scale genomic rearrangements (duplications, inversions, deletions) can occur due to homologous recombination between repeats in nucleotide sequence. We are studied the distribution of exact maximal direct repeats in 23748 bacterial chromosomes that do not contain degenerated nucleotides, and in 3 independent sets of non-genomic sequences. It was found that bacterial chromosomes are considerably different from non-genomic sequences in the lengths of the longest repeats. In 33 strains, mainly in *Bordetella pertussis* and *Escherichia coli*, direct repeats of more than 100 thousand base pairs was detected, while in non-genomic sequences their length does not exceed 32 base pairs. The lack of correlation between the value of longest repeat length and chromosome length or GC-content is shown. The clarification of the lower threshold level is proposed for the classification of genomes by the longest repeat length. The boundary of the significance for the maximal repeat length (640 base pairs) is determined with the inequality of Vysochanskij–Petunin.

Key words: long direct repeats, bacterial genomes, Vysochanskij–Petunin inequality.

1. Введение

В 1970 г. была выдвинута гипотеза о том, что эволюционный прогресс обусловлен дупликацией генетического материала с последующей дивергенцией копий [1]. Крупномасштабные хромосомные делеции, дупликации и инверсии могут возникать вследствие гомологичной рекомбинации и транспозиций. Данные события могут быть обусловлены наличием повторов в бактериальной ДНК [2, 3]. После десятилетнего

периода накопления данных было обнаружено, что геномы прокариот содержат огромное число повторяющихся последовательностей [3–5]. При увеличении количества доступных для анализа хромосом бактерий стало известно, что длина максимального повтора (прямого или инвертированного) может достигать значительной величины и даже превышать 100 тысяч нуклеотидных пар (н.п.). В 2007 г. при анализе генома *Methylobacillus flagellatus* KT выявлен прямой повтор длиной 143032 н.п. [6]. В 2008 г. был полностью секвенирован геном лабораторного штамма *Escherichia coli* DH10B, в котором идентифицирована крупномасштабная дупликация длиной 113260 н.п. [7]. При изучении линейных хромосом бактерий из рода *Streptomyces* было обнаружено, что они содержат на концах терминальные инвертированные повторы, которые могут достигать длины до нескольких сотен тысяч н.п. [8–10].

Мы проанализировали распространение максимальных прямых повторов с идентичными копиями в 23748 полностью секвенированных хромосомах бактерий. Во время проведения данной работы была опубликована статья [11], в которой исследовалось распространение прямых и инвертированных повторов длиной более 100 н.п. в 6387 бактериальных геномах.

2. Материалы и методы

2.1. Бактериальные хромосомы

В качестве источника использовалась локальная копия базы данных NCBI GenBank по состоянию на октябрь 2021 г., включающая в себя 24629 последовательностей бактериальных хромосом, не содержащих вырожденные нуклеотиды (S, W, R, Y, K, M, B, D, H, V, N). Было исследовано 23748 хромосом длиной не менее 1 млн. н.п. Суммарная длина этих хромосом составляет 92671604930 н.п. Наибольшей длиной (16040666 н.п.) обладает геном *Minicystis rosea* DSM 24000 (номер доступа в GenBank CP016211.1) [12]. В подавляющем большинстве случаев эти хромосомы являются кольцевыми, поэтому в наших дальнейших вычислениях повторов все нуклеотидные последовательности рассматривались как кольцевые.

2.2. Негеномные последовательности

Было исследовано распределение прямых повторов в трёх независимых наборах негеномных последовательностей длиной менее 10 млн. н.п. (это обусловлено тем, что только 68 бактериальных хромосом обладают длиной свыше 10 млн. н.п.):

а) «bactMirr». Для каждой из оставшихся 23680 хромосом генерировалась негеномная последовательность идентичной длины посредством использования десятичных знаков числа пи. Диапазон вариации длин последовательностей от 1000000 н.п. до 9997872 н.п.

б) «bactShuff». Каждая из нуклеотидных последовательностей 23680 бактериальных хромосом перемешивалась посредством случайной перестановки без неподвижных нуклеотидов. Необходимо обратить внимание на то, что для данного набора негеномных последовательностей сохраняется не только длина, но и AT/GC-состав исходного генома.

в) «randByMt». 19 групп негеномных последовательностей (по 500 штук с одинаковой длиной в каждой, длина для отдельной группы от 1 млн. н.п. до 10 млн. н.п. с интервалом 500 тыс. н.п.) было получено с помощью встроенного в Microsoft Visual Studio 2010 генератора псевдослучайных чисел mt19937 (вихрь Мерсенна [13]).

2.3. Программное обеспечение

Поиск максимальных прямых повторов в невырожденных кольцевых нуклеотидных последовательностях (содержащих только A, C, G, T) осуществлялся с помощью разработанной нами программы FindMaxRepAll.

3. Результаты

В каждой из 23748 бактериальных хромосом были найдены максимальные повторы. Их длина варьировала в диапазоне от 27 до 391982 н.п. Подавляющее большинство хромосом (23639, т.е. 99.54 %) содержало максимальные повторы длиной до 40 тыс. н.п., однако в 76 хромосомах выявлены повторы длиной от 40 тыс. н.п. до 100 тыс. н.п., а в 33 хромосомах — от 100 тыс. н.п. до 391982 н.п.

Всякий фрагмент последовательности однозначно определяется своей длиной и положением своего начального символа (нуклеотида в геномной последовательности). Если данный фрагмент не имеет копий в последовательности, то мы называем его одиночным, если же в последовательности имеется не менее двух копий данного фрагмента, то фрагмент, как известно, называется повтором, а число его копий – кратностью повтора.

Повтор называется максимальным, если всякий его содержащий фрагмент является одиночным. Таким образом, максимальный повтор перестаёт быть повтором при увеличении его длины.

Программа FindMaxRepAll получает на входе таксон (таблицу базы данных) и для каждой кольцевой последовательности таксона выдаёт файл-таблицу с колонками «позиция, длина». В колонке «позиция» перечисляются те позиции данной последовательности, в которых находятся максимальные повторы и для этих позиций сообщается длина соответствующего повтора. Такой файл даёт достаточно подробные сведения о максимальных повторах в нуклеотидной последовательности.

Таблица 1. Статистические параметры максимальных повторов в нуклеотидных последовательностях

		Хромосомы бактерий	«bactMirr»
Длина наибольшего повтора, н.п.	Диапазон вариации	27–391982	17–30
	Среднее арифметическое	5248	21
	Стандартное отклонение	8408	3.474
Средняя длина максимального повтора, н.п.	Диапазон вариации	10.26–14.31	10–12
	Среднее арифметическое	11.81	11.00
	Стандартное отклонение	0.553	0.315
		«bactShuff»	«randByMt»
Длина наибольшего повтора, н.п.	Диапазон вариации	18–30	17–27
	Среднее арифметическое	21.872	21.35
	Стандартное отклонение	1.454	3.985
Средняя длина максимального повтора, н.п.	Диапазон вариации	10–13	10.07–11.75
	Среднее арифметическое	11.28	11.19
	Стандартное отклонение	0.403	0.466

Ввиду малых значений соответствующих стандартных отклонений величины средних арифметических для средней длины максимального повтора (таблица 1) можно рассматривать как константы, характеризующие распределения максимальных повторов в бактериальных хромосомах. Геномы сильнее всего отличаются от 3 наборов негеномных последовательностей по значениям длины наибольшего повтора. Несмотря на значительную длину наибольших повторов, средняя длина максимальных повторов в хромосомах бактерий остаётся довольно малой (11.81), что говорит об отсутствии влиянии длинных максимальных повторов на статистику распределения. Коэффициенты корреляции между количеством максимальных повторов и длиной последовательности, а также между суммой длин максимальных повторов и длиной последовательности оказались около 0.99 (как для бактериальных хромосом, так и для 3 наборов негеномных последовательностей). В свою очередь, длина наибольшего повтора не коррелирует с длиной нуклеотидной последовательности. Для бактериальных хромосом данный коэффициент корреляции равен 0.047. Значение коэффициента корреляции между длиной наибольшего повтора и GC-составом хромосомы составляет 0.0027.

Чаще всего наибольшие повторы длиной более 100 тыс. н.п. встречаются у *Bordetella pertussis* (11 хромосом) и *Escherichia coli* (4 хромосомы) (таблица 2). Это соответствует результатам, полученным в работе [14], где было обнаружено, что в 300 хромосомах из 9331 исследованных содержатся повторы длиной более 30 тыс. н.п.

Таблица 2. Бактериальные хромосомы, содержащие наибольшие прямые повторы длиной свыше 100000 пар нуклеотидов

№ доступа GenBank	Название	Длина повтора, н.п.
CP010133.1	*Escherichia coli* C11	391982
CP006996.1	*Rhodococcus pyridinivorans* SB3094	365622
AP022639.1	*Bradyrhizobium diazoefficiens* H12S4	357603
CP010235.1	*Escherichia coli* S40	285417
CP017405.1	*Bordetella pertussis* J448	258999
CP014470.1	*Thiomicrospira* sp. S5	218241
AP022638.1	*Bradyrhizobium diazoefficiens* F07S3	204187
CP000967.2	*Xanthomonas oryzae* pv. *oryzae* PXO99A	185758
CP009741.1	*Burkholderia mallei* Turkey9 chromosome 1	185501
CP026997.1	*Bordetella pertussis* J085	174044
CP025001.1	*Bacillus siamensis* SCSIO 05746	169775
CP025530.1	*Bordetella pertussis* D236	167593
CP046994.1	*Bordetella pertussis* J299	166714
CP042254.1	*Legionella longbeachae* B3526CHC	165245
LT991957.1	*Enterobacter cloacae* complex bacterium C45	165065
CP032736.1	*Bordetella pertussis* J196	158184
CP046987.1	*Bordetella pertussis* J737	157669
CP046981.1	*Bordetella pertussis* ATCC 12742	155358
CP010157.1	*Escherichia coli* D10	153570
CP032735.1	*Bordetella pertussis* J318	149511
CP011126.1	*Coxiella*-like endosymbiont CRt	147808
CP020000.1	*Acinetobacter calcoaceticus* CA16	145069
CP000284.1	*Methylobacillus flagellatus* KT	143034
CP017873.1	*Campylobacter coli* WA333	141717
CP017404.1	*Bordetella pertussis* J447	130021
CP045427.1	*Shewanella* sp. YLB-09	128217
CP046991.1	*Bordetella pertussis* J349	126766
CP025527.1	*Bordetella pertussis* J139	121151
CP007620.1	*Pseudomonas putida* DLL-E4	118617
CP053604.1	*Escherichia coli* NEB10-beta	113300
CP041626.1	*Dolosigranulum pigrum* 83VPs-KB5	105458
CP009731.1	*Burkholderia mallei* Turkey4 chromosome 1	104397
CP013913.1	*Serratia fonticola* GS2	100913

Ранее была предложена классификация бактериальных хромосом по содержанию повторов [15]. Она основывается на том, что одним из самых длинных многокопийных участков является оперон, кодирующий гены рибосомных РНК (его длина составляет около 5–7 тысяч н.п.). Особый интерес представляют геномы класса III (которые содержат наибольший повтор длиной свыше 7 тысяч н.п.). В анализируемой нами выборке к данному классу относятся 10.88 % хромосом.

В качестве минимального порогового уровня для учёта повторов в обсуждаемой статье [15] была выбрана длина 500 н.п. Нами была проведена оценка границ значимости максимальных повторов на основании выборочных распределений для негеномных последовательностей из набора «randByMt». При этом использовалось неравенство Высочанского-Петунина [16], условием применения которого является унимодальность выборочных распределений. Оказалось, что из 9500 частотных гистограмм максимальных повторов только 156 гистограм имеют два локальных максимума, в то время как остальные гистограмы имеют один локальный максимум и, следовательно, унимодальны. При более подробном анализе оказалось, что для 156 гистограмм значение отношения частот локальных максимумов менее 0.00004, следовательно данные гистограммы также являются унимодальными. Из неравенства Высочанского-Петунина вытекает, что

$$L \geq m + \frac{2\sigma}{3P^{0.5}},$$

где L – граница типичности, m – среднее арифметическое, σ – стандартное отклонение, P – вероятность.

При значимости $P = 10^{-6}$ получаем границу типичности L, равную 637.2. Окончательно принимаем, что значимые максимальные повторы имеют длину ≥ 640 н.п.

Для бактериальных хромосом доля значимых максимальных повторов довольно мала (от 0 до 0.00077). Они присутствуют в 23021 хромосоме. Доля сумм длин значимых максимальных повторов варьирует в диапазоне от 0 до 0.0919.

4. Список литературы

1. Оно С. *Генетические механизмы прогрессивной эволюции.* М.: Мир, 1973. 228 с. (Перевод с англ. Ohno S. *Evolution by gene duplication.* Springer-Verlag, 1970).
2. Achaz G., Rocha E.P.C., Netter P., Coissac E. Origin and fate of repeats in bacteria. *Nucleic Acids Res.* 2002. V. 30. P. 2987–2994. doi: 10.1093/nar/gkf391
3. Treangen T.J., Abraham A.-L., Touchon M., Rocha E.P.C. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* 2009. V. 33. P. 539–571. doi: 10.1111/j.1574-6976.2009.00169.x
4. Coenye T., Vandamme P. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Res.* 2005. V. 12. P. 221–233. doi: 10.1093/dnares/dsi009
5. Mrazek J., Guo X., Shah A. Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. USA.* 2007. V. 104. P. 8472–8477. doi: 10.1073/pnas.0702412104
6. Chistoserdova L., Lapidus A., Han C., Goodwin L., Saunders L., Brettin T., Tapia R., Gilna P., Lucas S., Richardson P.M., Lidstrom M.E. Genome of *Methylobacillus flagellatus*, molecular basis for obligate methylotrophy, and polyphyletic origin of methylotrophy. *J. Bacteriol.* 2007. V. 189. P. 4020–4027. doi: 10.1128/JB.00045-07
7. Durfee T., Nelson R., Baldwin S., Plunkett III G., Burland V., Mau B., Petrosino J.F., Qin X., Muzny D.M., Ayele M., Gibbs R.A., Csorgo B., Posfai G., Weinstock G.M., Blattner F.R. The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse. *J. Bacteriol.* 2008. V. 190. P. 2597–2606. doi: 10.1128/JB.01695-07
8. Hopwood D.A. Soil to genomics: the *Streptomyces* chromosome. *Annu. Rev. Genet.* 2006. V. 40. P. 1–23. doi: 10.1146/annurev.genet.40.110405.090639
9. Choulet F., Gallois A., Aigle B., Mangenot S., Gerbaud C., Truong C., Francou F.-X., Borges F., Fourrier C., Guerineau M., Decaris B., Barbe V., Pernodet J.-L., Leblond P. Intraspecific variability of the terminal inverted repeats of the linear chromosome of *Streptomyces ambofaciens*. *J. Bacteriol.* 2006. V. 188. P. 6599–6610. doi: 10.1128/JB.00734-06
10. Tidjani A.-R., Bontemps C., Leblond P. Telomeric and sub-telomeric regions undergo rapid turnover within a *Streptomyces* population. *Sci. Rep.* 2020. V. 10. Article № 7720. doi: 10.1038/s41598-020-63912-w
11. Malhotra N., Seshasayee A.S.N. Replication-dependent organization constrains positioning of long DNA repeats in bacterial genomes. *Genome Biol. Evol.* 2022. V. 14. Article № evac102. doi: 10.1093/gbe/evac102
12. Pal S., Sharma G., Subramanian S. Complete genome sequence and identification of polyunsaturated fatty acid biosynthesis genes of the myxobacterium *Minicystis rosea* DSM 24000^{T}. *BMC Genomics*. 2021. V. 22. Article № 655. doi: 10.1186/s12864-021-07955-x
13. Matsumoto M., Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. on Modeling and Computer Simulations*. 1998. V. 8. P. 3–30. doi: 10.1145/272991.272995
14. Schmid M., Frei D., Patrignani A., Schlapbach R., Frey J.E., Remus-Emsermann M.N.P., Ahrens C.H. Pushing the limits of *de novo* genome assembly for complex prokaryotic genomes harboring very long, near identical repeats.

Nucleic Acids Res. 2018. V. 46. P. 8953–8965. doi: 10.1093/nar/gky726
15. Koren S., Harhay G.P., Smith T.P.L., Bono J.L., Harhay D.M., Mcvey S.D., Radune D., Bergman N.H., Phillippy A.M. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* 2013. V. 14. Article № R101. doi: 10.1186/gb-2013-14-9-r101
16. Высочанский Д.Ф., Петунин Ю.И. Обоснование правила 3σ для одномодальных распределений. *Теория вероятностей и мат. статистика.* 1980. Вып. 21. С. 25–36.