УДК 575.852

© А. В. Илатовский, В. А. Ланцов

Петербургский институт ядерной физики им. Б. П. Константинова РАН

ПАРНЫЕ ПОВТОРЫ В СТРУКТУРЕ БАКТЕРИАЛЬНОГО ГЕНОМА И РЕКОМБИНОГЕННАЯ АКТИВНОСТЬ КЛЕТКИ

❀ **В настоящей работе мы исследовали встречаемость парных совершенных прямых и инвертированных повторов в ряде бактериальных геномов. Кумулятивные профили встречаемости повторов для 36 бактериальных штаммов показали, что распределения повторов неслучайны и имеют ряд характерных особенностей. Сравнение геномов обычной по рекомбиногенной активности *E. coli* K-12 и несущей гиперрекомбиногенный белок RecA *P. aeruginosa* показало обратную зависимость между числом прямых повторов и рекомбиногенной активностью. В целом, геномы непатогенных бактерий содержат большее по сравнению с патогенами число прямых повторов.**

❀ **Ключевые слова**: бактериальный геном; анализ последовательностей; нуклеотидные повторы; белок RecA.

ВВЕДЕНИЕ

Повторяющиеся нуклеотидные последовательности присутствуют в геноме любого организма, от бактерий до человека. Структурно полинуклеотидные повторы **(ПНП)** различаются длиной (короткие или длинные), взаимной ориентацией (прямые или инвертированные), степенью идентичности (совершенные или несовершенные), структурой (тандемные или интерспейсерные) и количеством копий повторяющейся последовательности (множественные или парные). Функционально ПНП служат сайтами узнавания разнообразных ДНК-связывающих ферментов, в частности, ферментов гомологической рекомбинации и репликации, обеспечивая динамику эволюции геномов. Структурно-функциональные характеристики нуклеотидных повторов стали одним из активно развивающихся разделов геномики, который оперирует разнообразными подходами и генерирует интересные гипотезы пластичности геномов (см. обзоры (Смирнов, 2007; Berg et al., 2002; Kawano et al., 2002; Petrillo et al., 2006; Rocha, 2003)). Эволюционная изменчивость бактериальных геномов определяется рядом событий, среди которых горизонтальный перенос ДНК, редуцирование или инвертирование части бактериального генома напрямую связаны с гомологической рекомбинацией **(ГР)**, ее активностью в данной бактериальной клетке и наличием сайтов рекомбинации — гомологических нуклеотидных последовательностей, которые могут присутствовать в геноме в виде прямых или инвертированных повторов.

ГР, являясь частью механизма репарации ДНК, задействована во многих ключевых процессах жизнеобеспечения бактериальной клетки, связанных как со структурой ее генома, так и с ее функционированием (обеспечение правильной сегрегации хромосом, преодоление коллапса репликации и т. п.) (Бабынин, 2007; Cox, 2002). Все эти процессы нуждаются в строго дозированной ГР, контроль над которой осуществляется не с помощью ее индуцируемости, а за счет ее стимулируемости. Действительно, в экспоненциально делящейся бактериальной клетке всегда присутствует избыток главного белка ГР — ДНК-связывающего белка RecA, составляющий до 10 000 молекул на клетку (Moreau, 1987; Sassanfar et al., 1990), что, в принципе, позволяет покрыть одновременно до 30 000 нуклеотидов. При этом белок RecA находится в виде спиральных филаментов, компактизованных межфиламентным взаимодействием, как это было выявлено в кристаллах белка RecA из *Escherichia coli* (Story et al., 1992). Остается только активировать белок RecA путем его полимеризации (в 5'→3' направлении) на однонитевой ДНК **(онДНК)** в момент стимулирования ГР, причем очевидным стимулятором является появляющаяся в клетке (например, в результате работы комплекса RecBCD) онДНК.

Система бактериальной конъюгации у *E. coli* K-12 позволяет *in vivo* количественно оценить активность систем ГР клетки или, иными словами, рекомбиногенную активность ее белка RecA **(RecAEc)**, выраженную в частоте рекомбинационных обменов **(ЧРО)** на единицу длины ДНК. Выяснилось,

Поступила в редакцию 16.04.2010.
Принята к публикации 03.11.2010.

что величина ЧРО для *E. coli* K-12 достаточно низка (5 обменов в расчете на геном) (Lanzov, 2003). Выяснилось также, что не только мутантные белки RecAEc, но и белки RecA из других бактерий показывают в *E. coli* различные величины ЧРО. Так, белок RecA из патогена *Pseudomonas aeruginosa* **(RecAPa)** показал гиперрекомбинацию, увеличив ЧРО в 6–8 раз, а гибридный белок RecAX53 (в котором 12 аминокислот центрального домена белка были из RecAPa, а остальные 340 аминокислот из RecAEc) увеличивал ЧРО в 9 раз (Bakhlanova et al., 2001). Отметим, что белки RecAPa и RecAX53 по сравнению с RecAEc имеют повышенные биохимические активности, важные для рекомбинации (Baitin et al., 2008).

Попадание патогенной бактерии в организм человека, как правило, сопряжено с активным сопротивлением организма и борьбой бактерии за выживание. Усиление системы рекомбинационной репарации ДНК может быть одним из факторов, используемых патогенной бактерией в этой борьбе (Dorer et al., 2010; Merterns et al., 2008). В этом случае можно ожидать усиления рекомбиногенной активности белка RecA у такой бактерии.

В геноме любого микроорганизма имеются повторы нуклеотидов различной длины, которые могут стимулировать ГР. В случае прямых повторов ГР приводит к делетированию фланкированной ими части генома, а при инвертированных повторах — к инверсии соответствующей части генома. Оба события дестабилизируют геном. У бактерий ГР активна, и их геномы оптимизированы эволюционным отбором (в том числе и специализированным «ПН-выбором» (Смирнов, 2007)) для поддержания как стабильности, так и функциональной активности генома. Увеличение стабильности генома требует уменьшения числа повторов, тогда как нормальное функционирование бактерии может не считаться с этим требованием, и есть примеры сохранения сложной системы повторов, защищенных от структурных изменений специальными белками клетки (Buchet et al., 1999). Этим объясняется тот факт, что отбор ПНП не всегда сопряжен с уменьшением размера генома и числа повторов в нем (Смирнов, 2007). Тем не менее, представляется разумным предположение о том, что должна существовать обратная корреляция между числом парных повторов и рекомбиногенной активностью клетки: чем выше эта активность, тем меньше повторов должно быть в ее геноме. При этом важно анализировать такие геномные (суммарные) характеристики ПНП, которые бы коррелировали с требованиями, предъявляемыми белком RecA к минимальным размерам ДНК для образования стабильного пресинаптического комплекса.

Начиная с 1992 г., когда впервые был клонирован ген *recA* из *Proteus mirabilis* и показана его способность компенсировать отсутствие гена *recA* в *E. coli* (Eitner et al., 1992), было секвенировано более 370 геномов патогенных и непатогенных бактерий (Benson et al., 2007). Они представляют основательную базу для сравнительных структурных исследований бактериального генома в целом и его рекомбинационных систем в частности. В настоящей работе мы попытались ответить на три следующих вопроса:

1) Какие характеристики генома, учитывающие совершенные парные ПНП, информативны для обсуждаемых проблем?
2) Действительно ли геном *P. aeruginosa* отличается от генома *E. coli* уменьшенным количеством повторов, которые могут дестабилизировать геном?
3) Если да, то сколь широко ограничения на число повторов в геноме распространены у патогенных бактерий?

МАТЕРИАЛЫ И МЕТОДЫ

Список бактерий формировался исходя из результатов поиска аминокислотной гомологии при помощи программы BLAST (McGinnis et al., 2004). Для анализа содержания повторов были отобраны бактерии, идентичность аминокислотной последовательности белка RecA которых с белком RecAEc была не ниже, чем у белка RecAPa (всего 36 штаммов, из них 21 — патогены): **Aeromonas hydrophila* ATCC 7966 (76 %, CP000462), *Baumannia cicadellinicola* Hc (82 %, CP000238), *Colwellia psychrerythraea* 34H (74 %, CP000083), **Erwinia carotovora* SCRI1043 (88 %, BX950851), **Escherichia coli* CFT073 (100 %, AE014075), ***E. coli* K-12 (100 %, U00096)**, **E. coli* O157:H7 Sakai (99 %, BA000007), **Haemophilus influenzae* 86-028NP (73 %, CP000057), *H. influenzae Rd KW20* (73 %, L42023), *H. somnus* 129PT (72 %, CP000436), *Mannheimia succiniciproducens* MBEL55E (72 %, AE016827), **Pasteurella multocida* Pm70 (74 %, AE004439), *Photobacterium profundum* SS9 (77 %, CR354531), **Photorhabdus luminescens* TTO1 (86 %, BX470251), *Pseudoalteromonas haloplanktis TAC125* (72 %, CR954246), ****Pseudomonas aeruginosa* PAO1 (71 %, AE004091)**, **P. aeruginosa* UCBPP-PA14 (71 %, CP000438), *Psychromonas ingrahamii* 37 (75 %, CP000510), **Salmonella typhimurium* LT2 (97 %, AE006468), *Shewanella amazonensis* SB2B (82 %, CP000507), **S. oneidensis* MR-1 (84 %, AE014299), *S. sp.* MR-4 (86 %, CP000446), *S. sp.* MR-7 (86 %, CP000444), S. sp. W3-18-1 (84 %, CP000503), **Shigella boydii* Sb227 (100 %, CP000036), **S. dysenteriae* Sd197 (100 %, CP000034), **S. flexneri* 2a 2457T (100 %, AE014073), **S. flexneri* 5b 8401 (100 %, CP000266), **S. sonnei* Ss046 (100 %, CP000038), *Sodalis glossinidius morsitans* (88 %, AP008232), **Vibrio cholerae* O1 El Tor N16961 (79 %, AE003852), **V. parahaemolyticus* RIMD 2210633 (79 %, BA000031), **V. vulnificus* YJ016 (80 %, BA000037), *Wigglesworthia glossinidia* (72 %, BA000021), **Yersinia pestis* CO92 (89 %, AL590842), **Y. pseudotuberculosis* IP 32953 (89 %, BX936398); символом * указаны патогенные штаммы, в скобках приведены степень идентичности аминокислотных последо-

вательностей белков RecA данной бактерии и RecAEc и номер нуклеотидной последовательности генома бактерии в базе GenBank (Benson et al., 2007).

Кумулятивное распределение (профиль) вероятности встречаемости ПНП в геноме бактерии

ПНП характеризуется тремя параметрами: размером повторяющихся последовательностей L, расстоянием между центрами повторяющихся последовательностей S и степенью идентичности этих последовательностей I. Для исследования встречаемости повторов в бактериальных геномах мы разработали компьютерную программу поиска повторов в последовательности ДНК и статистической обработки полученных данных.

Поиск повтора осуществляется следующим образом: сначала выбирается участок ДНК минимального размера $L_{\min} = 10$ п. н. Затем в области, ограниченной параметром $S_{\max} = 2\,000$ п. н., проводится поиск последовательности, идентичной первоначальной со степенью идентичности не хуже I. Если такая последовательность найдена, длина повторяющейся последовательности увеличивается и проверяется степень идентичности. В случае, когда такое увеличение уменьшает степень идентичности ниже I, текущий повтор отбраковывается, а в итоговом списке повторов будет зарегистрирован повтор, полученный на предыдущем шаге. В противном случае длина повторяющейся последовательности будет увеличиваться до тех пор, пока дальнейшее увеличение не приведет к падению степени идентичности ниже I.

Предполагается, что размер L определяет только вероятность успешного поиска гомологии рекомбинационным комплексом и не влияет непосредственно на вероятность акта рекомбинации (за исключением пороговой длины L_C, ниже которой рекомбинация невозможна). Тогда адекватным количественным описанием встречаемости повтора в данном геноме является кумулятивное распределение (профиль) вероятности встречаемости повторов в этом геноме по параметру L со стороны больших значений с равным статистическим весом для каждого значения L. Действительно, в рамках данного предположения вероятность рекомбинации практически полностью зависит от успешного поиска гомологии в последовательности ДНК, и поэтому, в первую очередь, она будет происходить на повторе с бóльшим размером L. Итак, для количественной оценки встречаемости повторов мы использовали кумулятивный профиль.

Кумулятивный профиль рассчитывается следующим образом. После прогона программы поиска повторов для данного генома получается распределение числа повторов от длины повторяющейся части $N(L)$. Для учета длины генома L_G, распределение $N(L)$ умножается на нормирующий множитель, равный $1/L_G$, что эквивалентно частоте встречаемости в расчете на 1 п. н. Таким образом, кумулятивный профиль определяется как

$$P(L_i) = \sum_{L \geq L_i} N(L)/L_G.$$

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

ПНП — составная и необходимая часть любого генома, которая служит как для его организации, так и функционирования. Например, укладка бактериального генома в петлеобразные структуры (у *E. coli* насчитывается до 100 петель (Krawiec et al., 1990)) также направляется определенными повторами, рассредоточенными по геному (Tolstorukov et al., 2005).

Среди возможных подходов в описании встречаемости парных ПНП в бактериальном геноме кумулятивный профиль вероятности встречаемости таких повторов оказался наиболее информативным с точки зрения задач настоящего исследования. Спейсерную область ДНК между повторами выбирали из расчета, чтобы повторы с большой вероятностью попадали в одну петлю укладки ДНК, т. е. чтобы спейсеры были много меньше размера петли у *E. coli*, составляющей ~46 000 п. н. Такой подход позволил получить непрерывное распределение искомой вероятности P от длины повтора L с шагом в 1 п. н.

Кумулятивные профили распределения парных совершенных прямых повторов для геномов 36 бактериальных штаммов (21 из которых патогенные) приведены на рисунке 1. Как видно из рисунка, зависимость вероятности встречаемости прямых повторов в геноме от их длины имеет три четко выраженных части. Резкое уменьшение вероятности найти в геноме повтор длиной от 10 до 18–20 п. н. сменяется умеренным спадом кривой для повторов 20–100 п. н., которое, в свою очередь, сменяется резким спадом вероятности для длинных повторов от 100 до ~1 000 п. н. (область редких событий).

Интересно сравнить полученные профили с рассчитанными теоретически в предположении о случайном распределении нуклеотидов в бактериальном геноме (см. Приложение). Теоретический анализ зависимости числа повторов от частот встречаемости нуклеотидов каждого типа в данном геноме показал, что чем ближе частоты к равномерному распределению, тем меньше будет повторов. В предельном случае, когда все частоты равны, число повторов минимально. К такому случаю весьма близко распределение нуклеотидов в последовательности генома *E. coli* K-12. Профиль, полученный для подвергнутой случайному перемешиванию последовательности генома *E. coli* K-12, полностью описывается теоретической кривой, приведенной на рисунке 1. Для геномов с отличным от равномерного распределением частот встречаемости нуклеотидов можно ожидать значительно большего числа повторов. Так, для генома *W. glossinidia*, содержание в котором пар ГЦ является самым низким среди известных (22,5 %), расчетное число повторов в 5 раз больше. Сравнение с реальными профилями показало, что количество коротких (10–20 п. н.) повторов в 2–3 раза больше, чем расчетное для случайного распределения нуклеотидов в последовательности. Тем не менее, характер зависимости количества повторов от отклонения частот встречаемости

нуклеотидов от равномерного распределения соответствует результатам теоретического анализа. Так, геном *W. glossinidia* содержит в 6,3 раз больше коротких повторов, чем геном *E. coli* K-12. Это может свидетельствовать о неслучайности последовательности генома и об отсутствии специального контроля за числом прямых повторов (в сторону уменьшения) со стороны клетки.

Важной особенностью рассматриваемой зависимости является резкое изменение ее характера, когда длина повтора становится 18–20 п. н. и более. Такие повторы являются неслучайными (теоретическое рассмотрение дает вероятность их существования на порядки меньше) и значимы для генома. Общепринято, что активирование белка RecA связано с образованием филамента на онДНК, точнее, одного полного оборота спирали, что составляет 6 молекул белка, полимеризующихся приблизительно на 18 нуклеотидах (Kowalczykowski, 2000; West, 2003). Насколько нам известно, экспериментально найденная минимальная последовательность, способная инициировать ГР *in vivo*, имеет длину 26 п. н. (Ogawa et al., 1992), что близко к теоретически ожидаемой величине 18 п. н.

Согласно генетическим и биохимическим данным, приведенным выше, белок RecAPa существенно более активен в инициировании ГР. Можно было бы предположить, что в клетках *P. aeruginosa* повышенная активность RecAPa компенсируется большей активностью конкурирующего белка SSBPa. Однако это не так, последний по сродству к онДНК оказался менее сильным, чем SSBEc (Baitin et al., 2003). Это предполагает, что RecAPa все же реализует свои нестандартные рекомбинационные активности *in vivo*. Следовательно, можно ожидать, что вероятность присутствия повторов от 20 до 100 п. н. у *P. aeruginosa* должна быть меньшей, чем у *E. coli*. Рисунок 1 подтверждает, что одной из форм стабилизации генома могут быть ограничения, налагаемые на число прямых повторов в нем.

Вопрос о том насколько и другие патогенные бактерии подобно *P. aeruginosa* стабилизируют свой геном путем ограничения числа потенциальных сайтов для ГР или, иными словами, используют более активную рекомбинационную репарацию как защиту от повреждений в ДНК, остается открытым. Косвенным ответом может стать сравнение профилей встречаемости повторов в геномах патогенных и непатогенных бактерий. Рисунок 1 показывает кумулятивные профили встречаемости прямых повторов для 36 бактериальных штаммов, из которых 21 — патогенные. Как видно, искомые профили в основных чертах подобны друг другу (за исключением отдельных случаев, которые будут рассмотрены далее), и в целом число повторов в геномах непатогенов больше, чем в геномах патогенов. Это наблюдение также подтверждает тезис об ограничении числа повторов в ходе борьбы патогенных бактерий за сохранность своих геномов и их приспособляемость к окружающим условиям.

Ряд бактериальных геномов имеет особенности в распределении прямых повторов, отличные от основных тенденций. Во-первых, сильно деградированные геномы бактерий *B. cicadellinicola* и *W. glossinidia*, являющихся облигатными эндосимбионтами цикадки *Homalodisca coagulata* (Wu et al., 2006) и мухи цеце *Glossina brevipalpis* (Akman et al., 2002). Длина геномов составляет 686 194 и 697 724 п. н., содержание пар ГЦ составляет 33,2 % и 22,5 %, максимальная длина прямых повторов — 38 и 20 п. н. соответственно. Специфический характер окружающей среды кардинальным образом отразился на строении геномов этих бактерий: деградация до минимального набора генов привела не только к весьма малому размеру последовательности, но и к отсутствию достаточно длинных повторов. Во-вторых, геном *S. glossinidius*, бактерии, которая является эндосимбионтом мухи цеце *G. morsitans*, но может существовать и в свободном виде, и, возможно, представляет собой переходную форму (Toh et al., 2006). Длина генома *S. glossinidius* составляет 4 171 146 п. н., содержание пар ГЦ составляет 54,7 %, что типично для свободноживущих бактерий, однако, его кодирующая плотность аномально низкая — 51 %, в то время как характерное значение — 85–90 %. Кумулятивный профиль также показывает отклонение — пониженное по сравнению с геномами других бактерий содержание прямых повторов. В-третьих, геном *P. ingrahamii*, выделенного из арктических льдов экстремального психрофила, который способен расти при температуре 12°C (Auman et al., 2006). Уникальные условия существования и пониженное содержание пар ГЦ (40,1 %) могут существенным образом влиять на функционирование различных систем метаболизма ДНК, в т. ч. и белков, участвующих в рекомбинации (Riley et al., 2008). Кумулятивный профиль говорит о том, что данный геном содержит повышенное число прямых повторов по сравнению с другими геномами.

На рисунке 2 приведены кумулятивные профили распределения совершенных инвертированных повторов для всех исследованных геномов. В отличие от распределения прямых повторов (рис. 1), в большинстве случаев наблюдается резкий спад вероятности встречаемости повторов практически до нуля при длине 30–40 п. н., более длинные повторы являются единичными событиями. Характер спада, как и в случае прямых повторов, описывается теоретически рассчитанной кривой, с учетом различий в частотах встречаемости индивидуальных нуклеотидов в геноме — число повторов возрастает от наименьших величин для *E. coli* K-12 до наибольших для *W. glossinidia*. Аналогично ситуации с прямыми повторами, вероятность встречаемости инвертированных повторов больше ожидаемой в предположении о случайном распределении нуклеотидов, причем это отличие увеличивается с ростом длины повтора, что говорит о вероятной биологической роли инвертированных повторов, например, в качестве составных частей IS элементов (Mahillon et al., 1998).

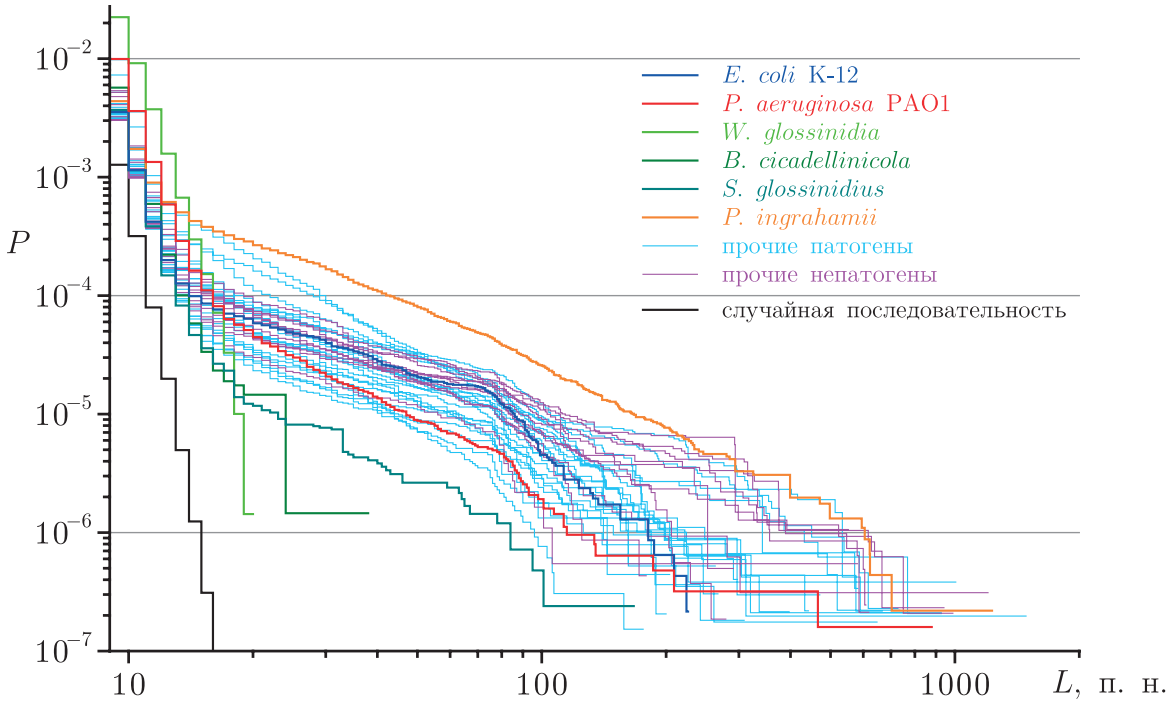

Рис. 1. Кумулятивные профили распределения совершенных прямых повторов. Различными цветами обозначены профили для последовательностей геномов: основных штаммов исследования *E. coli* K-12 и *P. aeruginosa* PAO1; эндосимбионтов *W. glossinidia*, *B. cicadellinicola* и *S. glossinidius*; экстремального психрофила *P. ingrahamii*; остальных патогенных и непатогенных штаммов (всего 36). Для сравнения приведена рассчитанная в предположении о случайном распределении нуклеотидов теоретическая кривая, соответствующая «минимальному» случаю равномерного распределения частот встречаемости индивидуальных нуклеотидов (выделена черным цветом)

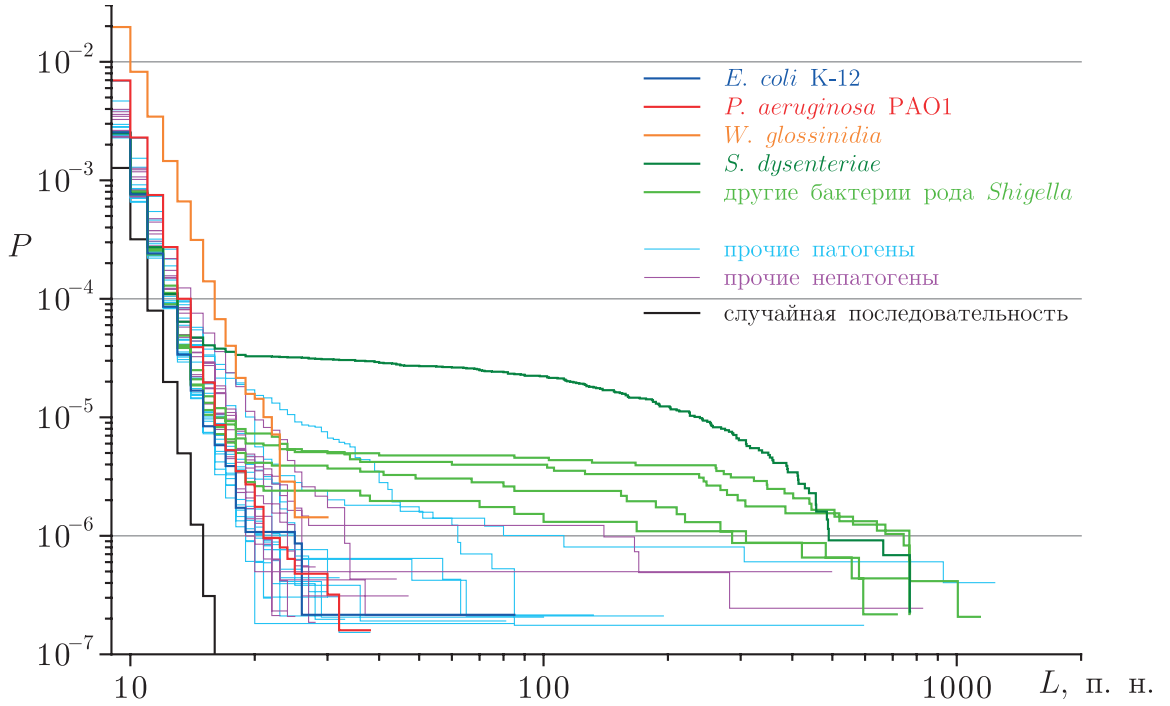

Рис. 2. Кумулятивные профили распределения совершенных инвертированных повторов. Различными цветами обозначены профили для последовательностей геномов: основных штаммов исследования *E. coli* K-12 и *P. aeruginosa* PAO1; эндосимбионта *W. glossinidia*; *S. dysenteriae* и других бактерий рода *Shigella*; остальных патогенных и непатогенных штаммов (всего 36). Для сравнения приведена рассчитанная в предположении о случайном распределении нуклеотидов теоретическая кривая, соответствующая «минимальному» случаю равномерного распределения частот встречаемости индивидуальных нуклеотидов (выделена черным цветом)

В противоположность кумулятивным профилям для прямых повторов, распределения вероятности встречаемости инвертированных повторов для патогенов (за исключением 5 штаммов рода *Shigella*) и непатогенных бактерий обладают близкими характеристиками как в области малых длин повторов, так и в области длинных повторов и не имеют четко выраженных отличий. Геномы бактерий рода *Shigella*, в особенности *S. dysenteriae*, содержат аномально высокое (на порядок величины) количество длинных повторов в области нескольких сотен п. н. Считается, что многочисленные внутригеномные перестройки являются одним из путей развития патогенности бактерий рода *Shigella* (Yang et al., 2005).

ЗАКЛЮЧЕНИЕ

В настоящей работе мы исследовали ряд бактериальных геномов на наличие в них совершенных прямых и инвертированных повторов, которые, в принципе, могли бы стать сайтами инициирования гомологической рекомбинации, приводящей к драматическим изменениям нуклеотидной последовательности генома. Из стабильности бактериальных геномов вытекает предположение о существовании соответствия между количеством потенциальных сайтов интрамолекулярной гомологической рекомбинации и рекомбиногенной активностью гомологической рекомбинации, определяемой белком RecA.

Кумулятивный профиль вероятности встречаемости повторов в зависимости от их длины оказался удобным инструментом для исследования содержания повторов в последовательности генома. Построенные кумулятивные профили для 36 бактериальных штаммов (рис. 1 и 2) показали, что распределения прямых и инвертированых повторов неслучайны и имеют ряд характерных особенностей — описываемый теоретически быстрый спад вероятности встречаемости коротких повторов (10–20 п. н.), медленный спад профиля в области до 80–100 п. н. в случае прямых повторов, область редких событий (единичные повторы длиной ~1 000 п. н.). Указанными свойствами обладает большинство исследованных последовательностей, однако есть ряд ярких исключений — кумулятивный профиль отражает индивидуальные особенности эволюции бактериального генома, направляемой как ПН выбором, так и отбором окружающей среды.

Сравнение геномов обычной по рекомбиногенной активности *E. coli* K-12 и гиперрекомбиногенной *P. aeruginosa* (точнее, несущей гиперрекомбиногенный белок RecAPa) показало обратную зависимость между числом прямых повторов и ожидаемой активностью рекомбинационной репарации клетки. Т. к. *P. aeruginosa* — известный патоген, и повышенная рекомбинационная репарация и, как следствие этого, пониженное число потенциальных сайтов рекомбинации может быть свойственно и другим патогенам, мы сравнили профили распределения повторов в геномах различных патогенных и непатогенных бактерий. Оказалось, что, аналогично соотношению в паре *E. coli* K-12/*P. aeruginosa*, геномы непатогенных бактерий содержат большее по сравнению с патогенами число прямых повторов.

Взаимное расположение профилей можно интерпретировать как следствие контроля количества повторов со стороны системы рекомбинации: чем ниже расположен профиль, тем больше ожидаемая рекомбиногенная активность белка RecA. К сожалению, к настоящему времени нет достаточной информации о свойствах белков RecA из различных микроорганизмов, тем не менее, используя представленные профили, можно очертить круг микроорганизмов, полезных для сравнительного анализа.

Работа поддержана грантом СПбНЦ РАН (проект «Анализ структуры бактериальных геномов для сравнения рекомбиназных активностей у патогенных и непатогенных бактерий»).

ПРИЛОЖЕНИЕ

Вероятность $p(L)$ нахождения совершенного повтора с повторяющейся последовательностью нуклеотидов длиной L в данной точке генома равна:

$$p(L) = h_1 h_2 \ldots h_L,$$

где h_i — вероятность нахождения данного нуклеотида в данной позиции последовательности. Предполагая независимость рапределения нуклеотидов, т. е. учитывая только мононуклеотидные частоты встречаемости f_i, выражение для вероятности $p(L)$ можно переписать следующим образом:

$$p(L) = f_A^{N_A} f_C^{N_C} f_G^{N_G} f_T^{N_T},$$

где N_i — количество нуклеотидов данного типа в последовательности, $\sum_i N_i = L$. Предполагая распределение нуклеотидов в геноме равномерным, можно получить явное выражение для N: $N_i = f_i L$. Таким образом, вероятность $p(L)$ равна

$$p(L) = [f_A^{f_A} f_C^{f_C} f_G^{f_G} f_T^{f_T}]^L$$

и зависит только от длины последовательности L. Константа $B = f_A^{f_A} f_C^{f_C} f_G^{f_G} f_T^{f_T} < 1$ полностью определяется мононуклеотидными частотами встречаемости, специфичными для каждого генома.

Пусть S — расстояние между повторяющимися последовательностями. Принимая во внимание тот факт, что вероятность обнаружения ответной подпоследовательности не зависит от S, при поиске в интервале значений $S < S_{\max}$ вероятность обнаружения повтора увеличивается в $\langle S \rangle$ раз. Имеем:

$$p_S(L) = \langle S \rangle B^L.$$

Кумулятивное распределение вероятности обнаружения повторов, построенное со стороны бóльших значений, определяется выражением:

$$P_S(L \geq L_0) = \sum_{L \geq L_0} p_S(L) =$$
$$= \langle S \rangle \sum_{L \geq L_0} B^L \approx \langle S \rangle B^{L_0} (1 + B + B^2 + ...).$$

Т. к. $B < 1$, то выражение в скобках может быть вычислено как сумма членов убывающей геометрической прогресии:

$$1 + B + B^2 + ... = 1 / (1 - B).$$

Отсюда:

$$P_S(L \geq L_0) = \langle S \rangle B^{L_0} / (1 - B).$$

Логарифмируя, имеем:

$$\lg P_S(L \geq L_0) = L_0 \lg B - \lg (1 - B) + \lg \langle S \rangle.$$

Рассмотрим зависимость величины B от частот f_i. Без ограничения общности, зафиксируем частоты f_G и f_C, т. е. $B = f_A^{f_A} f_T^{f_T} B_0$, где $B_0 = f_C^{f_C} f_G^{f_G}$. Исключим частоту f_T. Пусть $f_A + f_T = D$. Имеем:

$$B = f_A^{f_A} (D - f_A)^{D - f_A} B_0,$$

т. е. $B = B(f_A)$. Проанализируем эту зависимость. Применяя логарифмическое дифференцирование, в частности, учитывая, что

$$[x^x]' = x^x (1 + \ln x)$$

и приводя подобные члены, находим производную:

$$[B(f_A)]' = B(f_A) \ln f_A / (D - f_A).$$

Единственный ноль производной — $f_A = ½ D$. При $f_A <(>) ½ D$ получаем $[B(f_A)]' <(>) 0$. Таким образом, $f_A = ½ D$ или, иными словами, $f_A = f_T$ — точка минимума функции $B(f_A)$. Т. к. величина B симметрична относительно всех частот f_i, то ее минимум достигается в точке $f_i = ¼$ и равен ¼. Весьма близки к такому равномерному случаю частоты распределения нуклеотидов в геноме *E. coli* K-12.

Литература

1. *Бабынин Э.,* 2007. Молекулярный механизм гомологичной рекомбинации в мейозе: происхождение и биологическое значение // Цитология. Т. 49. С. 182–193.
2. *Смирнов Г.,* 2007. Почему редуцируются бактериальные геномы? // Молекулярная генетика, биофизика и медицина сегодня / под ред. В. А. Ланцова. СПб.: Изд. ПИЯФ РАН. С. 34–60.
3. *Akman L., Yamashita A., Watanabe H. et al.,* 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia* // Nat. Genet. Vol. 32. P. 402–407.
4. *Auman A., Breezee J., Gosink J. et al.,* 2006. *Psychromonas ingrahamii sp. nov.,* a novel gas vacuolate, psychrophilic bacterium isolated from Arctic polar sea ice // Int. J. Syst. Evol. Microbiol. Vol. 56. P. 1001–1007.
5. *Baitin D., Zaitsev E., Lanzov V.,* 2003. Hyper-recombinogenic RecA protein from *Pseudomonas aeruginosa* with enhanced activity of its primary DNA binding site // J. Mol. Biol. Vol. 328. P. 1–7.
6. *Baitin D., Bakhlanova I., Chervyakova D. et al.,* 2008. Two RecA protein types that mediate different modes of hyperrecombination // J. Bacteriol. Vol. 190. P. 3036–3045.
7. *Bakhlanova I., Ogawa T., Lanzov V.,* 2001. Recombinogenic activity of chimeric *recA* genes *(Pseudomonas aeruginosa/Escherichia coli)*: A search for RecA protein regions responsible for this activity // Genetics. Vol. 159. P. 7–15.
8. *Benson D., Karsch-Mizrachi I., Lipman D. et al.,* 2007. Genbank // Nucleic Acids Res. Vol. 35. P. D21–D25.
9. *Berg O., Kurland C.,* 2002. Evolution of microbial genomes: Sequence acquisition and loss // Mol. Biol. Evol. Vol. 19. P. 2265–2276.
10. *Buchet A., Nasser W., Eichler K., Mandrand-Berthelot M.,* 1999. Positive co-regulation of the *Escherichia coli* carnitine pathway *cai* and *fix* operons by CRP and the CaiF activator // Mol. Microbiol. Vol. 34. P. 562–575.
11. *Cox M.,* 2002. The nonmutagenic repair of broken replication forks via recombination // Mutat. Res. Vol. 510. P. 107–120.
12. *Dorer M., Fero J., Salama N.,* 2010. DNA damage triggers genetic exchange in *Helicobacter pylori* // PLoS Pathog. Vol. 6. P. e1001026.
13. *Eitner G., Adler B., Lanzov V., Hofemeister J.,* 1992. Interspecies *recA* protein substitution in *Escherichia coli* and *Proteus mirabilis* // Mol. Gen. Genet. Vol. 185. P. 481–486.
14. *Kawano M., Kanaya S., Oshima T. et al.,* 2002. Distribution of repetitive sequences on the leading and lagging strands of the *Escherichia coli* genome: Comparative study of long direct repeat (LDR) sequences // DNA Res. Vol. 9. P. 1–10.
15. *Kowalczykowski S.,* 2000. Initiation of genetic recombination and recombination-dependent replication // Trends Biochem. Sci. Vol. 25. P. 156–165.
16. *Krawiec S., Riley M.,* 1990. Organization of the bacterial chromosome // Microbiol. Rev. Vol. 54. P. 502–539.
17. *Lanzov V., Bakhlanova I., Clark A.,* 2003. Conjugational hyperrecombination achieved by derepressing the LexA regulon, altering the properties of RecA protein and inactivating mismatch repair in *Escherichia coli* K-12 // Genetics. Vol. 163. P. 1243–1254.
18. *Mahillon J., Chandler M.,* 1998. Insertion sequences // Microbiol. Mol. Biol. Rev. Vol. 62. P. 725–774.
19. *McGinnis S., Madden T.,* 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools // Nucleic Acids Res. Vol. 32. P. W20–W25.
20. *Mertens K., Lantsheer L., Ennis D., Samuel J.,* 2008. Constitutive SOS expression and damage-inducible AddAB-mediated recombinational repair systems for *Coxiella burnetii* as potential adaptations for survival within macrophages // Mol. Microbiol. Vol. 69. P. 1411–1426
21. *Moreau P.,* 1987. Effects of overproduction of single-stranded DNA-binding protein on RecA protein-dependent processes in *Escherichia coli* // J. Mol. Biol. Vol. 194. P. 621–634.

22. *Ogawa T., Shinohara A., Ogawa H., Tomizawa J.,* 1992. Functional structures of the RecA protein found by chimera analysis // J. Mol. Biol. Vol. 226. P. 651–660.
23. *Petrillo M., Silvestro G., Nocera P. et al.,* 2006. Stem-loop structures in prokaryotic genomes // BMC Genomics. Vol. 7. P. 170.
24. *Riley M., Staley J., Danchin A. et al.,* 2008. Genomics of an extreme psychrophile, *Psychromonas ingrahamii* // BMC Genomics. Vol. 9. P. 210.
25. *Rocha E.,* 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: From duplications to genome reduction // Genome Res. Vol. 13. P. 1123–1132.
26. *Sassanfar M., Roberts J.,* 1990. Nature of the SOS-inducing signal in *Escherichia coli*: The involvement of DNA replication // J. Mol. Biol. Vol. 212. P. 79–96.
27. *Story R., Weber I., Steitz T.,* 1992. The structure of the *E. coli recA* protein monomer and polymer // Nature. Vol. 355. P. 318–325.
28. *Toh H., Weiss B., Perkin S. et al.,* 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host // Genome Res. Vol. 16. P. 149–156.
29. *Tolstorukov M., Virnik K., Adhya S., Zhurkin V.,* 2005. A-tract clusters may facilitate DNA packaging in bacterial nucleoid // Nucleic Acids Res. Vol. 33. P. 3907–3918.
30. *West S.,* 2003. Molecular views of recombination proteins and their control // Nat. Rev. Mol. Cell Biol. Vol. 4. P. 435–445.
31. *Wu D., Daugherty S., Aken S. et al.,* 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters // PLoS Biol. Vol. 4. P. e188.
32. *Yang F., Yang J., Zhang X. et al.,* 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery // Nucleic Acids Res. Vol. 33. P. 6445–6458.

DNA REPEATS IN BACTERIAL GENOME AND INTRACELLULAR ACTIVITY OF HOMOLOGOUS RECOMBINASE

A. V. Ilatovskiy, V. A. Lanzov

❀ **SUMMARY: In present work distribution of perfect direct and inverted repeats in a set of bacterial genomes was analysed. Complementary cumulative distribution functions of repeat frequency for 36 bacterial strains are nonrandom and have certain characteristic features. Inverse relation of direct repeats frequency to recombinogenic activity is shown for reference *E. coli* K-12 strain and *P. aeruginosa* strain with hyperrecombinogenic RecA protein. In general, direct repeat frequency is higher in non-pathogenic strains than that in pathogens.**

❀ **KEY WORDS:** bacterial genomes; sequence analysis; DNA repeats; RecA protein.

❀ Информация об авторах

Илатовский Андрей Владимирович — м. н. с.
Лаборатория биофизики макромолекул, отделение молекулярной и радиационной биофизики, Петербургский институт ядерной физики им. Б. П. Константинова РАН.
188300, Ленинградская область, г. Гатчина, Орлова роща, ОМРБ ПИЯФ РАН.
E-mail: andreyi@omrb.pnpi.spb.ru.

Ланцов Владислав Александрович — д. б. н., профессор, заведующий.
Лаборатория молекулярной генетики, отделение молекулярной и радиационной биофизики, Петербургский институт ядерной физики им. Б. П. Константинова РАН.
188300, Ленинградская область, г. Гатчина, Орлова роща, ОМРБ ПИЯФ РАН.

Ilatovskiy Andrey Vladimirovich —
Division of Molecular and Radiation Biophysics.
Petersburg Nuclear Physics Institute, the Russian Academy of Sciences.
Orlova Roscha, Gatchina, 188300, Russia.
E-mail: andreyi@omrb.pnpi.spb.ru.

Lanzov Vladislav Alexandrovich —
Division of Molecular and Radiation Biophysics.
Petersburg Nuclear Physics Institute, the Russian Academy of Sciences.
Orlova Roscha, Gatchina, 188300, Russia.