Описание бактериального генома

Обзор генома и протеома бактерии Bacillus paralicheniformis Bac84

Беляев Г. Д.[1]
[1]Факультет биоинженерии и биоинформатики, Московский Государственный Университет им. М. В. Ломоносова, Москва, Россия.

РЕЗЮМЕ

Данная работа является обзором генома и протеома бактерии Bacillus paralicheniformis штамма Bac84. Ее цель – описание и выявление закономерностей в геноме и протеоме с использованием методов MS Excel и python.

Ключевые слова: Bacillus paralicheniformis, геном, протеом, бактерии.

1 ВВЕДЕНИЕ

Bacillus paralicheniformis – как и другие представители данного рода – вид грамположительных факультативно анаэробных бактерий, способных образовывать эндоспоры. Впервые был выделен в 2015 году из твенджана – ферментированной пасты на основе соевых бобов [1]. Видовое название дано из-за фенотипического и филогенетического сходства с Bacillus licheniformis.

Род Bacillus имеет крайне высокое значение в биотехнологии, так как большинство представителей являются экстремофилами, что позволяет выделять из них ферменты с различными функциями, работающие в нестандартном диапазоне условий. В частности, применяются рестриктаза BamHI, барназа, детергенты α-амилаза, субтилизин и другие гидролазы [2]. Это свидетельствует о перспективности изучения представителей данного рода для поиска новых противомикробных средств и ферментов.

В данной работе производится анализ генома и протеома Bacillus paralicheniformis штамма Bac84, который был выделен из микробного мата лагуны Рабиг-Харбор в Саудовской Аравии [3]. По данным из базы NCBI геном данной бактерии состоит из 4376831 пар оснований и содержит 4193 белок кодирующих гена, 110 РНК кодирующих генов и 84 псевдогена [4], [5].

2 МАТЕРИАЛЫ И МЕТОДЫ

2.1 Методы

В данном обзоре были использованы следующие методы по работе с таблицами MS Excel:

- Импорт текстового файла в таблицу с преобразованием в ячейки с помощью мастера импорта.
- Фильтрация данных по значению с использованием фильтра и расширенного фильтра.
- Выделение частей таблицы с помощью мыши и выделение всей таблицы с помощью инструмента в левом верхнем углу.
- Горячие клавиши:
    - Ctrl+C – копирование выделенного диапазона;
    - Ctrl+V – вставка данных из буфера обмена;
    - Ctrl+X – вырезание данных из выделенной области с помещением в буфер обмена;
    - Ctrl+A – выделение всей таблицы.
- Использование $ для фиксации диапазона при адресации (клавиша F4).
- Использование простейших арифметических формул.
- Использование формул:
    - ЕСЛИ – проверка необходимого условия.
    - ВПР – связь между таблицами и получение необходимых данных по ключу.
    - СУММ – суммирование выделенного диапазона.
    - СЧЁТЕСЛИ/СЧЁТЕСЛИМН – подсчет количества ячеек, для которых выполнено условие или условия соответственно.
    - ТРАНСП – транспонирование выделенного диапазона.
    - ЕСЛИОШИБКА – возвращает указанное значение при появлении ошибки.
    - НАЙТИ – возвращает индекс первого вхождения подстроки в строку.
    - БИНОМ.РАСП – возвращает значение биномиального распределения.
- Распространение формул Ctrl+D, Ctrl+R.
- Специальная вставка для избавления от формул.
- Создание примечаний.
- Создание ссылки на таблицу для скачивания.
- Создание гистограмм

Для получения информации о расположении ориджина и терминатора репликации был использован онлайн-сервис Genskew (http://genskew.csb.univie.ac.at/)

2.2 Материалы

Материалом для работы стала директория с информацией о геноме Bacillus paralicheniformis на сайте NCBI [4].

3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Размер генома

Размер генома – 4 376 831 пара оснований, что незначительно отличается от данных, приведенных на странице NCBI, посвященной B. Paralicheniformis [5]. Код программы для получения этой информации в таблице с сопроводительными материалами на листе genome_size.

3.2 Процент кодирующей области

С учетом перекрывания некоторых генов кодирующие последовательности занимают 3 863 784 пары нуклеотидов, что соответствует 88,28% от общей длины генома. Но в этих данных не учтены котранскрибируемые РНК, расположенные возле оперонов. Они также выполняют некоторые функции в клетке, но не учитываются в этой работе. Более подробная информация доступна в таблице с сопроводительными материалами на листе % CDS.

3.3 GC-состав

GC-состав – 45,84%. Количество нуклеотидов на одной цепи можно увидеть на Рисунке 1. Исходя из полученных данных можно сделать вывод о выполнении второго правила Чаргаффа для данной бактерии. Код программы для получения этой информации в таблице с сопроводительными материалами на листе genome_size.

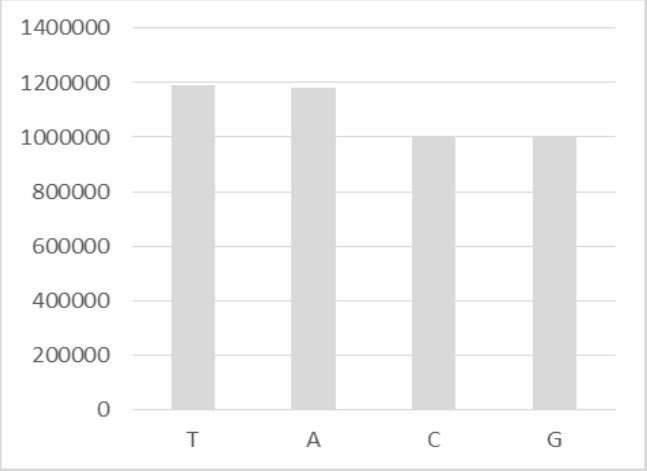

Рис.1 Гистограмма количества нуклеотидов на одной цепи

3.4 Типы генов и их распределение по цепям

Общее количество генов – 4 387. Из них 4193 белок кодирующих, 84 псевдогена и 110 РНК кодирующих. Распределение генов по типам, а также по цепям и случайность этого распределения можно увидеть в Таблице 1 и на Рисунке 2. Неслучайность распределения по цепям белок кодирующих генов довольно сложно объяснить, так как зачастую положение гена на + или – цепи определяет ориентация плазмиды, транспозона, генетического материала бактериофага и других мобильных генетических элементов при вставке в хромосому. Эта ориентация могла бы быть случайной, если бы не оперонная организация генома у прокариот, когда несколько генов регулируются одним промотором, что позволяет встраиваемой конструкции экспрессироваться вне зависимости от необходимости в клетке, благодаря чему в дальнейшем она с большей вероятностью станет кодирующей, а не мусорной ДНК. Это предположение согласуется

B. Paralicheniformis содержит 24 рРНК и 62 рибосомальных белка. Их полный список доступен в таблице с сопроводительными материалами на листе ribosomal. У бактерий гены, кодирующие рРНК, собраны в опероны, состоящие из 3 генов: 16S, 23S и 5S рибосомальных РНК, и преимущественное расположение этих оперонов на одной из цепей, как и в случае с тРНК, довольно трудно объяснимо. Подробнее гены рРНК и тРНК будут рассмотрены в соответствующих разделах обзора. Говорить о случайности распределения остальных видов РНК представляется невозможным из-за их крайне малого количества – 5 генов: 2 гена ncRNA, 1 ген SRP_RNA, 1 ген tmRNA и 1 ген RNase_P_RNA.

Случайность распределения псевдогенов по цепям обусловлена отсутствием отбирающего фактора на образование подобных нефункциональных аналогов действующих генов. По разным причинам псевдогены могут появляться на любой из цепей с поправкой лишь на то, что на цепях разное количество генов.

Кодирует:	всего:	Ориентация:		p-value:	Случайно:
		+1	-1		
Белок	4193	2025	2168	0,01415	Нет
Псевдоген	84	36	48	0,11493	Да
тРНК	81	53	28	0,00363	Нет
рРНК	24	21	3	0,00013	Нет
Другие РНК	5	2	3	0,5	Да

Табл.1 Распределение типов генов по цепям и случайность этого распределения.

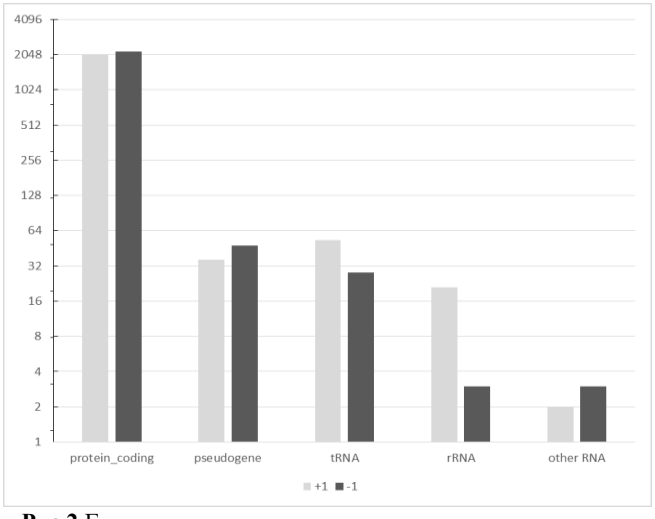

Рис.2 Гистограмма количества различных типов генов.

3.5 Ориджин и терминатор репликации

Для поиска ориджина и терминатора репликации был использован сервис Genskew (http://genskew.csb.univie.ac.at/). Поиск основан на анализе неоднородности GC-состава. OriC соответствует минимуму содержания гуанина и цитозина на участках в 1000 нуклеотидов, а ter – максимуму их содержания. Расчет идет по формуле:

Skew = (N(C) – N(G)) / (N(C) + N(G))

Графический результат поиска представлен на Рисунке 3. Предполагаемое место начала репликации находится в районе 26 257 пары нуклеотидов, а терминатора – в районе 2 074 225 пары, тем самым разделяя кольцевую ДНК на две почти равные части: первая состоит из 2 047 968 п.н., что составляет 46,79% от всего генома; вторая содержит 2 328 863 п.н., что соответствует 53,21% генома. Данные согласуется с функциональными требованиями для двунаправленности репликации у бактерий, так как эти части практически равны.

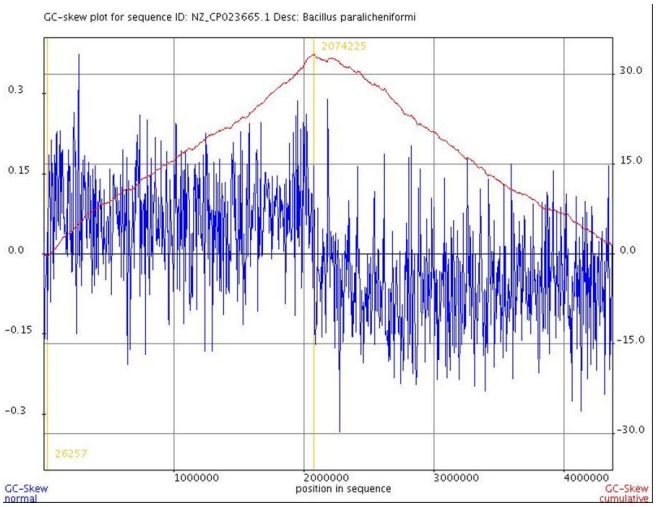

Рис.3 График GC-состава на участках в 1000 нуклеотидов

3.6 Гипотетические белки

Протеом B. licheniformis содержит 419 гипотетических белков, что составляет 9,99% от общего их числа. Такое большое число белков с неизвестной функцией обусловлено относительной лабораторной неизученностью данного вида, и других видов близкородственных бактерий, из анализа геномов которых, пришли к выводу о существовании подобных белков. Более подробная информация доступна в таблице с сопроводительными материалами на листе hypothetical.

3.7 Длины белков

Протеом Bacillus licheniformis состоит из 4 193 белков. Такое большое количество обычно не характерно для бактерий, но является нормальным в роде Bacillus. Больше всего белков (1680 шт.) находятся в диапазоне от 256 до 512 аминокислот. Мода для полученных данных – 89 аминокислот, медиана – 258. Самый длинный белок – синтаза нерибосомального пептида-антибиотика бацитрацина – состоит из 6 357 аминокислот. Самый короткий белок – сигнальный пептид для синтеза соединения, обеспечивающего устойчивость к эритромицину – состоит из 14 аминокислот. Распределение белков по длинам представлено на Рисунке 4. Более подробная информация доступна в таблице с сопроводительными материалами на листе protein length.

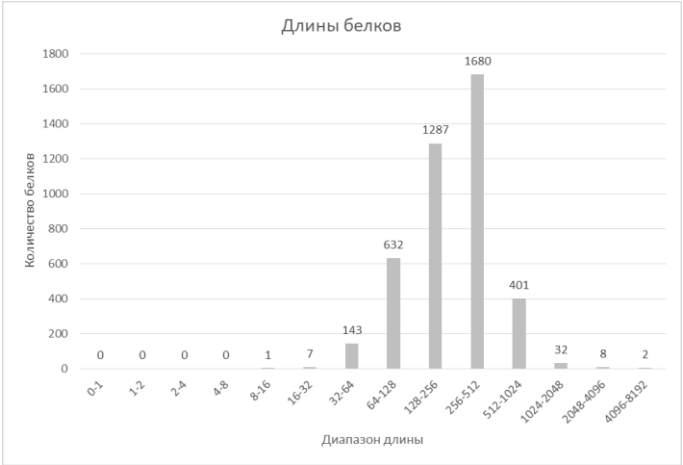

Рис.4 Гистограмма длин белков

3.8 Рибосомальные белки и рРНК

В геноме B. licheniformis было обнаружено 24 гена, кодирующих рРНК, и 62 гена, кодирующих рибосомальные белки. рРНК бактерий закодирована в разбросанных по всему геному рДНК-кластерах. Они имеют строго упорядоченную структуру и состоят из последовательно идущих генов 16S, 23S и 5S рРНК, ориентированных по направлению к терминатору. Также зачастую между генами 16S и 23S рРНК расположен спейсер, состоящий из регуляторных последовательностей и тРНК. Информация о расположении подобных оперонов и наличии в них тРНК доступна в Таблице 2. Более подробная информация доступна в таблице с сопроводительными материалами на листе ribosomal.

оперон:	диапазон:	цепь:	тРНК
1	36477-41548	+	tRNA-Ile; tRNA-Ala
2	61192-66263	+	tRNA-Ile; tRNA-Ala
3	121970-126866	+	-
4	127857-132752	+	-
5	192428-197323	+	-
6	682247-687145	+	-
7	1004694-1009628	+	-
8	3308382-3303425	-	-

Табл.2 Расположение оперонов рРНК и наличие в них тРНК

4 ВЫВОДЫ

В итоге проделанной работы можно сделать вывод о том, что Bacillus licheniformis обладает рядом черт характерным либо для большинства бактерий, либо для представителей рода Bacillus:

- Размер генома – 4 376 831 пара оснований. Эта длина больше, чем у большинства бактерий, но является характерной для рода Bacillus.
- Процент кодирующей области - 88,28%.
- GC-состав 45,84%.
- Гены распределены по цепям неравномерно.
- Терминатор удален от ориджина примерно на одинаковое расстояние с обеих сторон
- Количество и длины белков являются характерными для рода Bacillus.

Данная бактерия особенно интересна тем что 10% ее протеома представлено гипотетическими генами, и каждый из них потенциально может стать новым антибиотиком или применимым ферментом.

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Таблица с сопроводительными материалами: https://kodomo.fbb.msu.ru/~gennady.belyaev/term1/Belyaev_supple-fin.xlsx

СПИСОК ЛИТЕРАТУРЫ

1. Dunlap CA, Kwon SW, Rooney AP, Kim SJ. “Bacillus paralicheniformis sp. nov., isolated from fermented soybean paste.” Int J Syst Evol Microbiol. 2015 Oct;65(10):3487-3492.
2. Barrett AJ, Rawlings ND, Woessnerd. Handbook of proteolytic enzymes (2nd ed.). Elsevier Academic Press. 2004
3. Ghofran O, Salim B, Rozaimi R. “In silico exploration of Red Sea Bacillus genomes for natural product biosynthetic gene clusters.” BMC Genomics. 2018 May 22;19(1):382.
4. Директория с данными о геноме B. paralicheniformis Bac84 на сайте NCBI ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/993/925/GCF_002993925.1_ASM299392v1
5. Страница на NCBI с данными о B. Paralicheniformis https://www.ncbi.nlm.nih.gov/genome/41499
6. Xizeng Mao, Han Zhang, Yanbin Yin and Ying Xu, The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces//Nucleic Acids Research, 2012, Vol. 40, No. 17, P:8210–8218.