Биоинформатический анализ генома Bacteroides faecis (штамм BFG-108)

Олег Капшай[1]

[1] Факультет биоинженерии и биоинформатики, Московский государственный университет

1. ВВЕДЕНИЕ

В 2010 году из человеческих фекалий впервые выделили 2 штамма бактерий *Bacteroides faeces*. Они анаэробы, неподвижные, неспорообразующие, грамотрицательные, оксидаза-, каталаза- и уреаза-отрицательные. Растут при 25-42 C (оптимум 37 C), могут образовывать кислоты из моно- и олигосахаридов, но не из соответствующих полиспиртов. Ближайшим родственником является B. thetaiotaomicron [1].

В дальнейшем, были выявлены случаи бактериемии, вызванные *B. faeces* и др., после хирургических операций по удалению раковых опухолей в ЖКТ. Выделенные штаммы были устойчивы к 2 антибиотикам [2]. В 2021 был найден штамм, устойчивый к карбапенемам благодаря энзиму металло-B-лактамазе, появившийся в крови пациента на 5 день лихорадки. Данный вид резистентности распространен среди Bacteroides [3].

В данной работе определено количество некоторых групп генов, стоп кодонов в белок кодирующих участках. Рассмотрены статистические данные о кольцевой хромосомы, белковых последовательностей и перекрываниях генов. Также определены участки ter и oriC при помощи GC-skew, выполнено выравнивание последовательностей длинных перекрываний генов.

2. МАТЕРИАЛЫ И МЕТОДЫ

Таблица особенностей, статистика сборки и геномная последовательность взяты с сайта NCBI [4]. Также был использован discontiguous megablast NCBI. Для вычисления GC-skew использовалась программа GenSkew [5]. Для построения диаграмм, таблиц и сортировки таблицы особенностей использовались Гугл Таблицы. Статистическая значимость оценивалась при помощи критерия Пирсона [6] по формуле:

$$\chi^2 = \sum \frac{(O-E)^2}{O} \qquad (*),$$

где O - ожидаемое значение величины, E - наблюдаемое значение. Далее полученное значение сравнивается с критическим по таблице из учебника [6] и делается вывод о достоверности гипотезы.

Для анализа генома использовались скрипты на bash и python, которые хранятся на сервере кодомо в папке /home/students/y22/kaps/term1/minireview_scripts:

i) anything_stops - рассчитывает количество стоп-кодонов в генах. На вход принимает имя файла последовательности (chr, pl1 и pl2)(в файле находится только последовательность без пробелов, переносов строк и т.п.) и имя файла таблицы координат (chr_st_en_st, pl1_st_en_st, pl1_st_en_st)(таблица не содержит заголовка, координаты хранятся в виде <старт> <конец> <цепочка ДНК>, разделение через Tab. Скрипт выводит количество встреч стоп-кодонов и возможные ошибочные кодоны (не стоп) по данным координатам второй строкой.

ii) в подпапках fasta и table находятся скрипты, которые принимают на вход геном в FASTA формате и таблицу особенностей данной бактерии соответственно. Выводят скрипты те данные, которые указаны в названии, например, fasta_description выводит описание хромосом.

iii) в подпапке overlaps:

a) anything_overlaps принимает имя файла с последовательностью(chr, pl1, pl2), имя файла таблицы перекрываний (chr_overlaps, pl1_overlaps, pl2_overlaps). На выход выводятся последовательности перекрываний (файлы chr_seq_ov, pl1_seq_ov, pl2_seq_ov).

b) local_stops принимает строку последовательностей перекрываний и выводит встречаемость стоп-кодонов.

c) usual_overlaps принимает имя файла последовательностей перекрываний и разделяет последовательности в зависимости от типа перекрывания (конец с концом, страт с концом и т.п.)(см п. 3.5.1).

3. РЕЗУЛЬТАТЫ

3.1 Распределение стоп-кодонов

Геном *B. faeces* состоит из трех последовательностей - хромосомы и 2 плазмид unnamed1 и unnamed2. В табл 1 представлены их длины и GC состав и встречаемость стоп-кодонов генах белков.

Табл 1. Статистические данные о молекулах ДНК

	Длина,пн	GC состав, %	TGA	TAG	TAA
хромосома	628242	42,5	945	715	3097
unnamed1	36882	32	10	5	33
unnamed2	32924	47	12	5	20

Выдвинем гипотезу, что стоп-кодоны на хромосоме распределены случайным образом. Количество степеней свободы 2. Зафиксируем здесь и далее уровень значимости 0,05 доверительным, тогда критическое значение равно 5,99. Вероятность встретить кодон TGA (и кодон TAG, тк они имеют одинаковый нуклеотидный состав) равна

P = 0,575 * 0,425 * 0,575 / 8 = 0,01756,

для TAA вероятность встречи равна

P = 0,575 * 0,575 * 0,575 / 8 = 0,02376.

Сумма вероятностей 0,05888. Рассчитаем математическое ожидание для каждого стоп-кодона: для TGA и TAG

$$O = 4757 * \frac{0,01756}{0,05888} = 1418,7,$$

для TAA

$$O = 4757 * \frac{0,02376}{0,05888} = 1919,6.$$

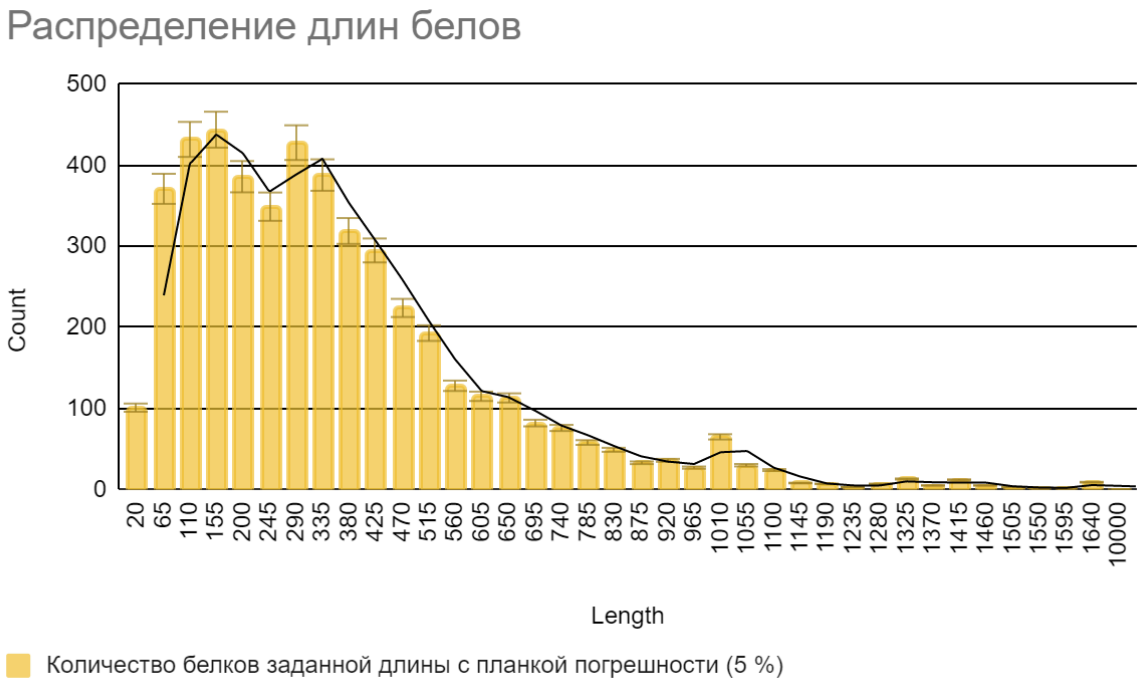
Тогда $\chi^2 = 1229,4$ по (*). Значение превышает критическое, гипотеза отвергнута.

3.2 Описание генов

В хромосоме *B. faeces* содержится 859 гипотетических белков (17,38 %), 58 генов рибосомальных белков, 212 белков-транспортеров и 15 транспортных белков. Генов РНК 88 штук, из них 15 рРНК, 71 тРНК, 1 некодирующая РНК и 1 транспортно-матричная.

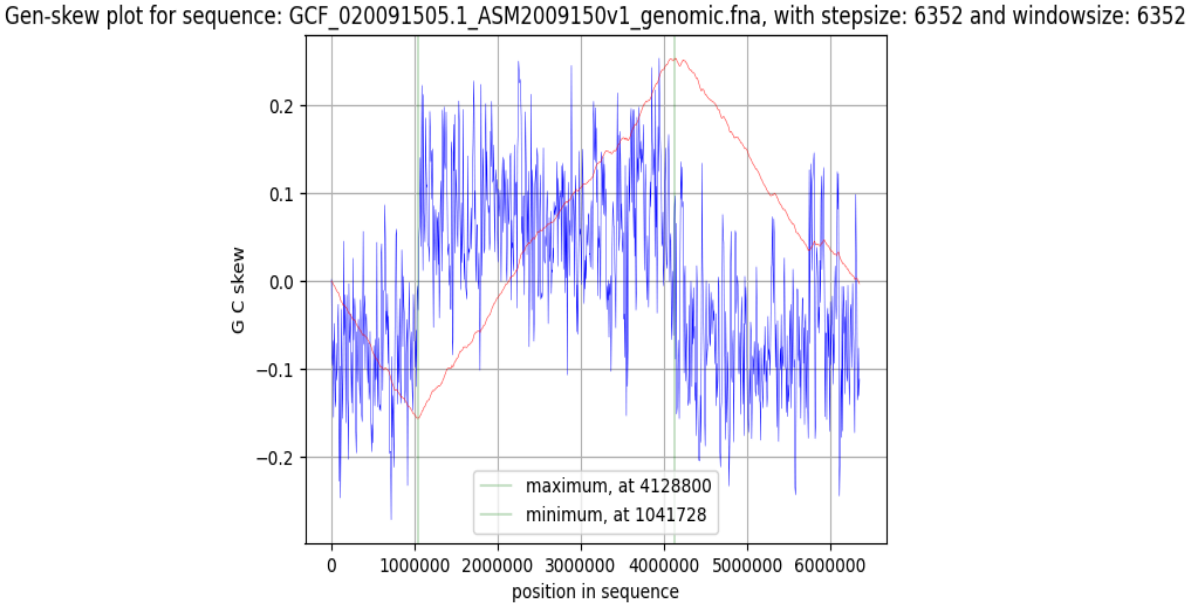
Выдвинем гипотезу, что каждый ген находится на прямой или обратной цепи ДНК равновероятно. Количество степеней свободы равно 1. Критическое значение равно 3,8. На прямой цепи хромосомы 2453 генов, на обратной - 2489. Тогда ожидаемое равно среднему арифметическому O = 2471. по формуле (*) получим значение $\chi^2 = 0,262$, что ниже критического значения. Следовательно, гипотеза не отвергнута.

Для построения диаграммы длин белков был подобран карман 45 а/к. Распределение имеет 2 максимума с длиной 110-145 и 245-290 аминокислот.



3.3 Участки ter и oriC

При помощи GC-skew определены возможные координаты участков ter - 4128800 и oriC - 1041728. Алгоритм строит график зависимости величины GC-skew = $\frac{\text{кол-во } C - \text{кол-во } G}{\text{кол-во } C + \text{кол-во } G}$, где количество берется в окне фиксированного размера.



3.4 Распределение нуклеотидов

Выдвинем гипотезу, что на обеих цепях хромосомы встречаемость нуклеотидов G равна встречаемости C, встречаемость A равна встречаемости T. Количество степеней свободы равно 1 в обоих случаях, критическое значение 3,8. Содержание нуклеотидов на "+" цепи: A - 1806757, C - 1336731, G - 1339027, T - 1799907. Для G и C: O = 1337879, тогда $\chi^2 = 1,97$ - не отвергаем гипотезу. Для A и T: O = 1803332, тогда $\chi^2 = 13,01$ - отвергаем гипотезу. Результат - количество G примерно равно количеству C, но для T и A это не верно.

3.5 Описание перекрываний

3.5.1 Общая характеристика перекрываний

В хромосоме *B. faeces* 408 перекрытий генов, из них 187 длиной 4 (45,83 %). У 166 длина имеет остаток 2 при делении на 3 (40,67 %), у 46 длина имеет остаток 1 при делении на 3, но есть лишь 9 перекрываний, длина которых нацело делится на 3. Самое длинное перекрывание имеет длину 108, медиана - 8. Стоп-кодон TAA встречается в областях перекрывания генов 150 раз, TAG - 46, TGA - 194. В местах перекрывания чаще всего встречается TGA, а TAA, который является самым частым в генах бактерии, только на 2 месте. Нуклеотидный состав: 1876 A, 1853 T, 1010 G, 824 C (GC состав 32,97%).

Последовательность перекрывания встречается ATGA 62 раза, TTAT 20, TCAT 14 раз, остальные еще реже. Последовательности длиннее 20 нуклеотидов встречаются только 1 раз. В таблице [S1] на листе sequence представлены последовательности перекрываний, их длина, молекула ДНК и тип. Тип записан 2 буквами, E - значит наличие стоп-кодона (на прямой или обратной цепи), S - наличие старт-кодона, A - отсутствие конца и начала. Например, запись AE означает, что слева в последовательности нет старт-/стоп-кодона (такое может быть в случае псевдогена или гена РНК[1] или в таблице особенностей указана не рамка считывания) и есть стоп-кодон на прямой цепи справа.

В табл 2 показана встречаемость типов перекрывания генов в хромосоме. SE и ES, EA и AE а также SA и AS - один и тот же тип, но первый ген лежит на разных цепочках (т.е. если цепь ДНК заменить на комплементарную, то последовательности перекрываний поменяют тип). Поэтому далее обозначаются SE*, EA* и SA* соответственно (для обозначения выбирается наиболее часто встречающийся из пары тип).

Табл 2. Типы перекрытий генов

тип перекрытия	SS	SE	SA	ES	EE	EA	AS	AE	AA
количество	3	73	145	17	22	133	0	3	12

[1] Ген, не кодирующий белок, не обязательно не имеет старт- или стоп-кодон

Тип определяется относительно “+” цепи. Первым указывается участок гена, который встречается в таблице особенностей раньше. Усл обозначения: S - start, E - end, A - another (любой не S и не E).

Видно, что тип SS почти не встречаются. Это скорее всего является следствием давления естественного отбора. Удивительно, но SE*, EA* и SA* расположены преимущественно в одной ориентации относительно прямой цепи, объяснить это не представляется возможным.

3.5.2 Поиск видоспецифичных перекрываний

5 самых длинных последовательностей перекрываний были использованы для выравнивания при помощи NCBI blast. Результат представлен в табл 3.

Табл 3. Выравнивание последовательностей перекрываний

последовательность	родственна
CTACTTAACAATTTCTAAGTTTTCAACTAAATT TACATTCATCTTTATTTTTTTCTGCTCCTCTTC TGTTATAGGAGGAGGGTCTGTTATTAATTTTA ACAATTCATT	только 2 штамма *B. thetaiotaomicron* (но не *B. faeces*) и *Parabacteroides distasonis*, причем все со 100 % сходством
ATGAAAAAGAAACTGAAAGCCGTCCTGTTCG ATATGGACGGCGTACTCTTTAATTCCATGCCC TACCACTC	*Bacteroides* и CrAss-like virus sp. (84 %)
TCATTCTCTTCTTACTTCACATTCAAACCCTA GGTTTCTCAATTCATTCATCCGGTATTCTTGT AAAGGTCTGGGCTTTTCTTTCG	*Bacteroides* и *Siphoviridae sp*. (95 %)
TTAATAATCCTTGCTCCGTGTTGACTGACGC TTTCCTTTGTTGCCGAAGGTGTTGAACCGAT AGGCAATATATGCCATAAAGTATACG	*B. faeces* и *B. thetaiotaomicron* (89,66 - 93,1 %)
TCACTCCCTGTCGCATGAAGGCAGCGGGTG GTGTGGGTGTGATAATTCGTCCTCATTACTT GTTATAATTATGTCCATGCAGATATCGA	*Bacteroides* и Uncultured bacterium (100 %)

Самая длинная (108 пн) последовательность перекрываний генов оказалась настолько специфична, что не встречается у всех штаммов. Перекрывания длины, делящейся нацело на 3, встречаются в геноме бактерии очень редко. Это объясняет отсутствие данной последовательности у некоторых штаммов *B. faeces.* Возможно, такие редкие последовательности будут найдены у других бактерий. Эти последовательности можно будет в будущем использовать для обнаружения отдельных видов бактерий и штаммов при помощи ПЦР.

4. ССЫЛКИ

[1] Min-Soo Kim, Seong Woon Roh, Jin-Woo Bae (2010) Bacteroide faeces sp. nov., isolated from human faeces. International journal of systematic and evolutionary microbiology 60 (11), 2572-2576

[2] Yangsoon Lee, Hyun Soo Kim, Dongeun Yong, Seok Hoon Jeong, Kyungwon Lee, Yunsop Chong (2015) Bacteroides faeces and Bacteroides intestinalis recovered from clinical specimens of human intestinal origin. Yonsei medical journal 56 (1), 292-294

[3] Charlotte Kaeuffer, Tiffany Ruge, Laure Diancourt, Benoit Romain, Yvon Ruch, Benoit Jaulhac, Pierre H Boyer (2021) First case of Bacteraemia Due to Carbapenem-Resistant Bacteroides faeces. Antibiotics 10 (3), 319

[4] Ссылка на геном и таблицу особенностей бактерии на NCBI - B. faeces

[5] Ссылка на сайт для расчета GC-skew - Webskew

[6] Математическая статистика: Учеб. для вузов / В.Б. Горяинов, И.В. Павлов, Г.М. Цветкова, и др.; Под ред. В.С. Зарубина, А.П. Крищенко. - М.: Изд-во МГТУ им. В.Э. Баумана, 2001. -424с.

5. СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

S1) Таблица перекрываний генов Overlaps. На листе sequence представлены последовательности перекрываний, их длина, молекула ДНК, на которой расположены перекрывающиеся гены, тип перекрывания (см п. 3.5.1). На остальных листах находятся части таблицы особенностей для каждой молекулы, расчет длины перекрытий и статистических данных о них.