

Обзор протеома бактерии *Halocynthiibacter arcticus* PAMC 20958^T

Мыларщиков Д.Е.¹

¹Факультет биоинженерии и биоинформатики МГУ им. М.В.Ломоносова

Резюме

В данной работе представлены результаты анализа генома и протеома бактерии *Halocynthiibacter arcticus* с помощью табличного процессора. Представлены статистические данные по различным типам генов и их продуктам, проверена независимость размещения генов на цепях ДНК, а также определены возможные эволюционные тенденции развития данного штамма относительно параметров его генома.

1 Введение

Halocynthiibacter arcticus – грам-отрицательная аэробная бактерия класса Alphaproteobacteria, порядок Rhodobacterales, выделенная в морских осадках близ Шпицбергена, Норвегия. Бактерии палочкообразной формы, неподвижны, способны образовывать колонии белого цвета. Условия роста для штамма PAMC 20958^T таковы: температура 10-27°C, pH 5,5-9,5 и наличие 0,5-7,5% раствора хлорида натрия^[1].

Геном бактерии представлен одной кольцевой хромосомой (RefSeq: NZ_CP014327.1) и одной

кольцевой плазмидой (RefSeq: NZ_CP014328.1) и содержит 4297 генов^[2] при общей длине генома в 3,467 мегабаз. Изучение генома важно с практической точки зрения, так как данный штамм продуцирует несколько важных ферментов с оптимумом активности при температуре 21°C, таким образом, их термочувствительность может найти применение в биотехнологии^[1].

2 Материалы и методы

Для анализа мной были взяты два исходных файла:

GCF_000812665.2_ASM81266v2_feature_table.txt^[3] с сервера NCBI и таблица со сведениями о достоверности существования белков *H. arcticus* с сервера Uniprot. Обработка данных производилась мной в табличном редакторе LibreOffice Calc 5.2.0.4. В ходе работы были применены статистические и логические функции, различные типы фильтров, мастера создания сводных таблиц и построения диаграмм.

3 Результаты

Исходный файл с результатами анализа называется H_arcticus_description.ods (см. раздел «Сопроводительные материалы»).

3.1 Число генов белков и РНК

Суммарно обнаружен 4001 ген белков и РНК, а также 291 псевдоген (один из которых расположен на плазмиде). В сумме это даёт 4292 гена против сведений GenBank (4297 генов). Важно обратить внимание на то, что число тРНК (45) меньше числа кодонов (61). Среднее число генов на 1 млн пар нуклеотидов (п.н.) составляет 981 ген.

Таблица 1. Число генов белков и РНК *H. arcticus*

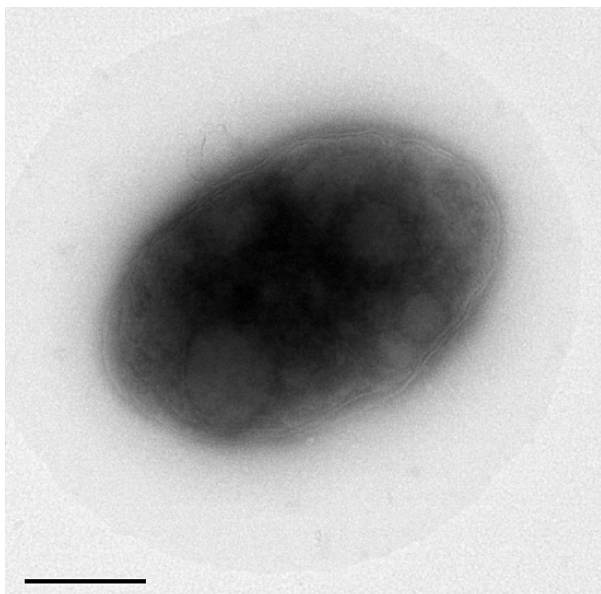


Рис. 1. Трансмиссионная электронная микрофотография клетки бактерии штамма PAMC 20958^T. Полоска длиной 0,5 мкм.

Тип гена	Количество	Проценты
CDS	3944	98,58%
ncRNA	2	0,05%
rRNA	9	0,22%
tmRNA	1	0,02%
tRNA	45	1,12%
Сумма	4001	100%

3.2 Длины белков

Самый короткий белок состоит из 37 аминокислотных остатков (а.к.), а самый большой — из 2470 а.к. Обнаружено, что 92% белков имеют длину меньше 550 а.к.

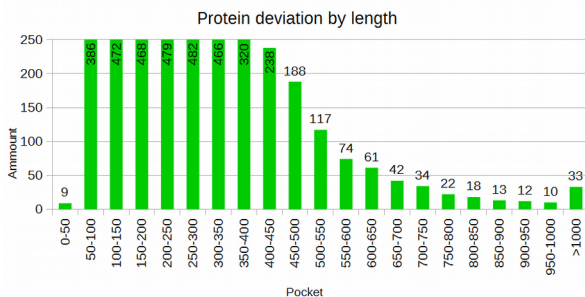


Рис. 2. Гистограмма длин белков (карман - 50 а.к.)

3.3 Распределение генов белков и РНК на прямой и комплементарной цепях

Гены распределены более-менее равномерно на разных цепях. Значение интегральной функции для биномиального распределения в случае, когда 1981 (или меньше) из 4001 генов оказались на одной цепи, составляет 0,274.

Таблица 2. Распределение генов белков и РНК по цепям хромосомы и плазмиды

Тип гена	+	-	Сумма	Проценты
CDS	1990	1954	3944	98,58%
RNA	30	27	57	1,42%
Сумма	2020	1981	4001	100,00%
Отклонение	19,5	-19,5		

3.4 Количество квазиоперонов в геноме

Было подсчитано число квазиоперонов в геноме в зависимости от максимально допустимого расстояния между генами в одном опероне.

Таблица 3. Количество квазиоперонов на обеих цепях хромосомы и плазмиды в зависимости от

максимально допустимого расстояния между генами в пределах одного оперона

Цепь\Шаг	50	100	200
хромосома +	1241	1033	860
-	1246	1032	841
плаزمида +	16	15	14
-	12	12	10
Сумма	2515	2092	1725
Ёмкость оперона	1,59	1,91	2,32

3.5 Статистические данные о пересечении генов

Обнаружено большое число пересечений генов в геноме бактерии (17% от общего числа функциональных генов).

Таблица 4. Данные о пересечениях генов на хромосоме и плазмиде

	Хромосома	Плаزمида	Итого
Полное количество	661	6	667
На одной цепи	543	6	549
На противоположных цепях	118	0	118
Минимальная величина, п.н.	1	3	1
Максимальная величина, п.н.	78	21	78

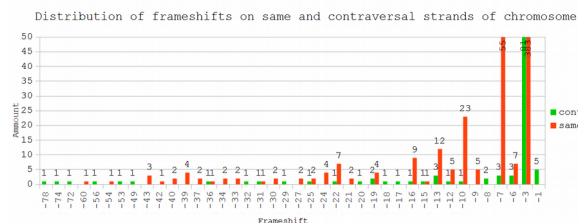


Рис. 3. Распределение сдвигов (п.н.) на одной и разноимённых цепях хромосомы

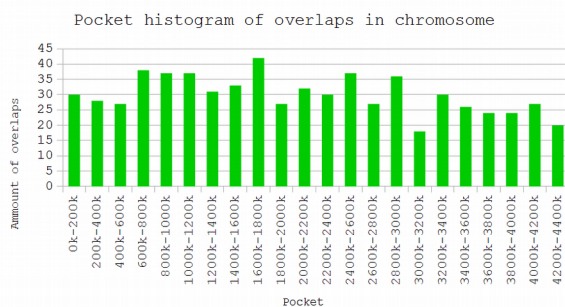


Рис. 4. Гистограмма количества пересечений генов на хромосоме

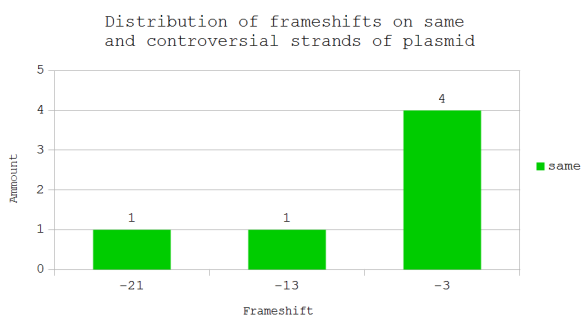


Рис. 5. Распределение длин пересечений генов на плазмиде

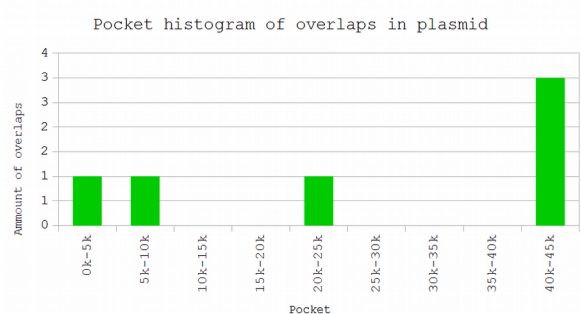


Рис. 6. Гистограмма пересечений генов на плазмиде

3.6 Статистика белков по достоверности их существования

С помощью сервиса перекодировки идентификаторов белков из формата GenBank в формат Uniprot была получена таблица с данными о существовании того или иного белка. Расшифровка таблицы:

транскрипция — нет очевидных доказательств существования белка, при этом показана активность его гена на уровне транскрипции;

гомология — нет свидетельств существования белка, однако обнаружены его ортологи;

предсказаны — нет никаких доказательств существования белка, определены с помощью специальных алгоритмов, выделяющих открытые рамки считывания и мотивы ДНК.

Таблица 5. Краткая статистическая таблица с данными о достоверности существования белков

Существование белка	Количество
Транскрипция	2
Гомология	1063
Предсказаны	2817

При совмещении идентификаторов GenBank и Uniprot оказалось возможным определить достоверность существования тех или иных белков на соответствующих цепях генома.

Таблица 6. Расширенная статистическая таблица с данными о достоверности существования белков

	Хромосома		Плаزمида		X	Сумма
	-	+	-	+		
Транскрипция	2					2
Гомология	581	481			1	1063
Предсказаны	1306	1454	17	18	22	2817
Сумма	1889	1935	17	18	23	3882

X - белки с одинаковыми названиями

3.7 Белки с повторяющимися названиями

Было обнаружено, что 1482 белка имеют одинаковые названия (всего 338 групп). Наибольшее число «копий» имеют интегразы и транспозазы.

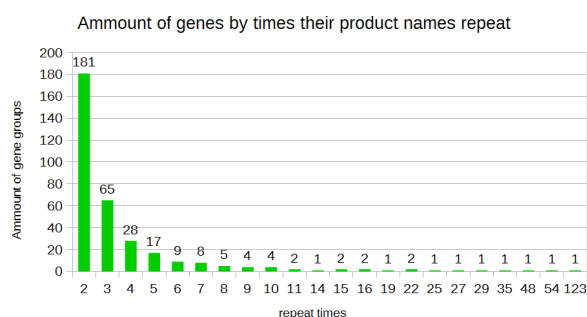


Рис. 7. Распределение количества повторяющихся белков по количеству повторений

4. Обсуждение

4.1 Число генов белков и РНК

Подавляющее большинство генов — белковые, что неудивительно, так как бактериям нужно небольшое количество «сервисных» аминокислот. Расхождение в значении числа генов с базой данных GenBank может быть вызвано некорректной первоначальной обработкой исходной таблицы. Меньшее число тРНК по сравнению с необходимым может быть вызвано компактизацией генома в ходе исторического развития вида благодаря вырожденности генетического кода. Наличие большого числа псевдогенов (около 5%) также позволяет говорить о богатой на генетические изменения истории этого штамма.

4.2 Длины белков

Бактериям нет необходимости в большом количестве огромных белков, поэтому у этой бактерии большинство белков относительно короткие. Несмотря на это, самый большой белок (2470 а.к.) - гипотетический белок AML51732.1, относится к семейству гистидиновых киназ-подобных АТФаз^[4].

4.3 Распределение генов белков и РНК на прямой и комплементарной цепях

Значение интегральной функции для биномиального распределения в случае, когда не более 1981 гена из 4001 расположены на одной цепи, составляет 0,274, в то время как для пары чисел (2000; 4001) значение составляет 0,5, поэтому реальное распределение генов можно назвать случайным (вероятность того, что не более 1981 гена окажутся на одной цепи, больше вероятности того, что от 1982 до 2000 генов включительно окажутся на одной цепи).

4.4 Количество квазиоперонов в геноме

При изменении максимально возможного расстояния между генами в одном опероне оказалось, что количество оперонов на плазмиде практически не меняется по сравнению с хромосомой, что можно объяснить большей вовлечённостью генов в опероны на плазмиде из-за особенностей процессов встраивания на неё генов.

4.5 Статистические данные о пересечении генов

Оказалось, что около 17% генов пересекаются между собой. Вероятно, это может быть связано с высокой мутационной активностью геномов прокариот и стремлению к компактизации генома. Эта гипотеза может быть подтверждена более-менее равномерным распределением пересечений по хромосоме, в то время как на плазмиде пересечения малораспространены.

4.6 Статистика белков по достоверности их существования

Показано, что огромное число белков лишь предсказано; удачно найдены ортологи лишь для части хромосомальных генов. Это позволяет говорить о том, что на настоящий момент для плазмидных белков не найдены ортологи, поэтому их можно считать в некоторой степени уникальными для данной бактерии.

4.7. Белки с повторяющимися названиями

Большое количество транспозаз и интеграз может свидетельствовать о широкой вовлечённости генетических механизмов в процессы изменения генома.

5. Заключение

В ходе анализа генома бактерии *Halocynthiibacter arcticus* PAMC 20958^T была выявлена возможная тенденция к компактизации генетического материала в ходе эволюции.

Сопроводительные материалы

1. Файл [H_arcticus_description.ods](#) с результатами табличной обработки исходных файлов.

Список литературы

1. Baek K, Lee YM, Shin SC, Hwang K, Hwang CY, Hong SG, Lee HK, *Halocynthiibacter arcticus* sp. nov., isolated from Arctic marine sediment, *International Journal of Systematic and Evolutionary Microbiology*, November 2015 65: 3861-3865, doi: [10.1099/ijsem.0.000507](#);
2. <https://www.ncbi.nlm.nih.gov/genome/?term=Halocynthiibacter+arcticus>
3. ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/00/812/665/GCA_000812665.2_ASM81266v2/GCA_000812665.2_ASM81266v2_feature_table.txt.gz
4. <https://www.ncbi.nlm.nih.gov/protein/AML51732.1>