Биоинформатика

Краткий обзор генома и протеома бактерии *Halocynthiibacter arcticus* PAMC 20958^{T}

Коннов С.И. [1]

[1]МГУ им М.В.Ломномова, факультет биоинженерии и биоинформатики, 1 курс

РЕЗЮМЕ

Работа посвящена анализу генома и протеома бактерии *Halocynthiibacter arcticus* штамма 20958^{T}**.** Рассчеты проводились с использование Microsoft Excel 2011 и Google Sheets. В данной работе было подсчитано число генов по типам продуктов, представлены статистические данные по распределению генов на прямой и комплиментарной цепях ДНК, проверена гипотеза о случайности распределения генов по прямой и комплиментарной цепями ДНК. Также в работе приводятся данные по количеству и длине "квазиоперонов" в геноме бактерии, пересечении генов и статистика белков по критериям достоверности их существования.

1 ВВЕДЕНИЕ

Род Halocynthiibacter (Kim et al., 2014) принадлежит к порядку Rhodobacterales в классе Alphaproteobacteria, основной филогенетической группе в глобальных океанах, и предполагается, что он оказывает значительное влияние на различные биогеохимические циклы (Giovannoni and Rappeı, 2000).

Halocynthiibacter arcticus - это грамотрицательная, оксидазоположительная, каталозоположительная бактерия, которая была выделена из арктического морского осадка. Данная бактерия неподвижна, имеет стержневидную форму (размер 0.8-1.5X1.5-3.0 μm) и образует круглые, выпуклые, блестящие колонии белого цвета диаметром 0.5-1.0 mm. Основными полярными липидами являются фосфатидилхолин, фосфатидилглицерин, неизвестный аминолипид и два неизвестных липида.

Оптимальными условиями роста штамма 20958^{T} считается температура 21^{O}C, pH 7,0-7,5 и присутствие 2 %(Об.) NaCl.

Геном бактерии представлен одной кольцевой хромосомой и одной кольцевой плазмидой. На данный момент данная бактерия не имеет практического применения в биотехнологии.(1)

2 МАТЕРИАЛЫ И МЕТОДЫ

Для анализа с сервера NCBI был взят файл GCA_000812665.2_ASM81266v2_feature_table.txt (2) и таблица со сведениями о достоверности существоания белков с сервера Uniprot. Обрботка данных проводилась в Google Sheets и Microsoft Excel 2011.

2.1 Для подсчёта числа генов транспортных и рибосомальных белков с помощью фильтра были вынесены на отдельные листы CDS, кодирующие рибосомальные белки и транспортные белки, после чего к столбцам на этих листах была применена функция СЧЁТЕСЛИ. Для подсчёта числа гипотетических белков и РНК разных видов к столбцам листа, содержащего исходную таблицу с аннотациями, была применена функция СЧЁТЕСЛИМН с различными условиями.

2.2 Для подсчёта длины белков были взяты данные столбца product_length исходной таблицы. Минимальное, максимальное, среднее и медианное значение длины, а также среднее отклонение были посчитаны с помощью соответствующих формул Microsoft Excel.

2.3 Для получения статистики распределения генов и псевдогенов на прямой и комплиментарной цепочках ДНК к столбцам листа, содержащего исходную таблицу с аннотациями, была применена функция СЧЁТЕСЛИМН с различными условиями.

2.4 Для проверки гипотезы о случайом распределении генов на прямой и комплиментарной цепочках ДНК на отдельном листе с помощью функции СЛУЧМЕЖДУ был создан столбец из случайным образом полученных нулей и единиц, равный по длине общему числу генов. С помощью функции СЧЁТЕСЛИ было подсчитано суммарное количество нулей и единиц, подсчитано отклонение от разделения общего числа генов на две равные части (1 были приняты за гены на прямой цепи, 0 за гены на обратной) и сопоставлено с наблюдаемым в геноме бактерии отклонением.

2.5 Для нахождения числа "квазиоперонов" в геноме бактерии на отдельном листе по порядку были вынесены начала и концы кодирующих последовательностей, по-

сле чего с помощью функции ЕСЛИ были определены пересечения начала последущего гена с концом предыдущего. Число “квазиоперонов” было подсчитано для различных значений максимального расстояния между генами.

2.6 Для подсчёта числа генов суммировались пересечения конца гена с началом предыдущего. Полученные данные были статистически обработаны соответсвующими функциями Microsoft Excel.

2.7 Для подсчета статистики белков по категориям достоверности существования к столбцу листа, содержащего таблицу со сведениями о достоверности существования белков, была применена функция СЧЁТЕСЛИ.

3 РЕЗУЛЬТАТЫ

В ходе анализа данных были получены следущие результаты

3.1 Число генов белков и генов РНК по категориям

Всего было обнаружено 4291 ген, из которых 291 псевдогена(один из которых расположен в плазмиде).

Тип	Генов белков	Генов на миллион азотистых оснований
Рибосомальных	54	12,34
Транспортных	385	87,99
Гипотетических	1 534	350,58
Прочих	1 971	450,46
Всего	3 944	

Таблица 1.Число генов белков по категориям

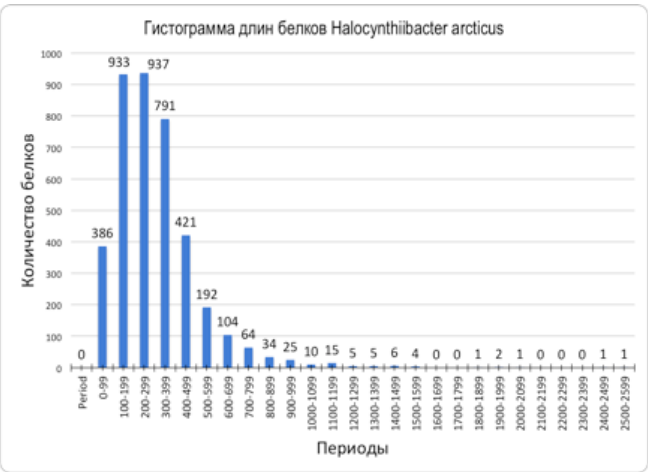
Тип	Гены РНК	Генов на миллион азотистых оснований
тРНК	45	10,28
рРНК	9	2,06
Прочих	2	0,46
Всего	56	

Таблица 2.Число генов РНК по категориям

3.2 Длины белков

Максимальная длина белка составила 2565 аминокислотных остатков (а.о.), минимальная длина белка составила 37 а.о.. Средняя длина белка 305 а.о. , медиана длин белков 267, стандартное отклонение 210 а.о.Длина большинства белков не превышает 700 а.о.

Рис.1 Гистограмма длин белков



3.3 Распределение генов, псевдогенов на прямой и комплиментарной цепочках ДНК

В целом число генов на прямой цепи оказалось несколько больше числа генов на комплиментарной цепи

Цепь	CDS	Гены РНК	Псевдогены
+	1 990	30	160
-	1 954	26	131

Таблица 3.Распределние генов белков, генов РНК и псевдогенов на прямой и комплиментарной цепи ДНК

3.4 Проверка гипотезы о случайном распределнии генов по цепочкам ДНК

При случайном распределении отклонение от ожидаемого составило 93 для прямой цепи и 50 для комплиментарной. Реально наблюдаемое отклонение

как для прямой цепи, так и для комплиментарной составило 18 (рассматривались CDS).

3.5 “Квазиопероны” в геноме бактерии

Квазиоперономи в данном случая мы считаем совокупности генов, находящихся на одной цепи С промежутками между ними не больше определенного порогового значения. При пороговом значении 100 нуклеотидов наибольшая длина квазиоперона составила 18 генов, большинство квазиоперонов имеют длину 2 или 3 гена. При увеличении порогового значения уменьшается число квазиоперонов.

Порог	Количество квазиоперонов
25	2 881
50	2 617
100	2 132
200	1 702

Таблица 4.Число квазиоперонов в зависимости от порогового значения

3.6 Статистика пересечения генов

В геноме бактерии *Halocynthiibacter arcticus* имеется 764 пересечения генов. Наибольшее пересечение имеет длину 78 п.н., средняя длина пересечения 7 п.н., медиана длин пересечения 3.

Количество пересечений	На одной цепи	На разных
Всего	624	140
В одной рамке считывания	484	105
Со сдвигом на один нуклеотид	139	23
Со сдвигом на два нуклеотида	1	12

Таблица 5 .Распределение длин пересечения генов на одной и на разных цепях с учетом сдвига рамки считывания

3.7 Статистика белков по категориям достоверности их существования

Существование большинства (71.17%) белков было предсказано на основании анализа последовательности ДНК. Существование 28.77% белков было установлено сопоставлением с известными белками.О существовании 2 белков имеются свидетельства на основании транскриптов.

	Предположение на основании гомологии	Предсказано	Свидетельства на уровне транскриптов
Число белков	1 117	2 763	2
Процент от общего числа	28,77%	71,17%	0,05%

Таблица 6 .Число белков по категориям достоверности существования

4 ОБСУЖДЕНИЕ И ЗАКЛЮЧЕНИЕ

4.1 Число генов белков и РНК

Ожидаемо большинство генов кодируют белки.Большое число гипотетических белков свидетельствует о недостаточной изученности данного организма.Значительное количество псевдогенов может быть связано с ошибками при секвенировании или долгим процессом адаптации бактерии с изменением генома.

4.2 Длины белков

Можно увидеть что наибольшее число белков имеет длину 100-500 а.о. , что может быть связано с тем, что данная длина белка яляется наилучшим соотношением между функциональностью и затратами на синтез. Возможно,что у данной бактерии нет необходимости в больших белках.

4.4 Проверка гипотезы о случайном распределении

Можно предположить,что гены распределены по цепям не случайным образом, поскольку отклонение от ожидаемого при случайном распределении больше и число отклонений от ожидаемого равно для каждой из цепей.

4.6 Пересечения генов

Большое число пересечений генов может быть связано с компактизацией генома, о чем можно судить по равномерному распределению пересечений генов в хромосоме.

4.7 Категории достоверности существования белков

Существование большинства белков предсказано на основании анализа ДНК, что может быть связано с недостаточной изученностью родственнх связей бактерии и сопосталением предположенных белков с белками бактерий из рода Halocynthiibacter.

БЛАГОДАРНОСТИ

Автор выражает благодарность преподавательскому составу факультета биоинженерии и биоинформатики МГУ им.М.В. Ломоносова.

5 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Файл со статистическими данными http://kodomo.fbb.msu.ru/~simon_konnov/term1/pr_13.xlsx

6 ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

(1) Kiwoon Baek, Yung Mi Lee, Seung Chul Shin, Kyuin Hwang, Chung Yeon Hwang, Soon Gyu Hong, Hong Kum Lee - Halocynthiibacter arcticus sp. nov., isolated from Arctic marine sediment http://www.microbiologyresearch.org/docserver/fulltext/ijsem/65/11/3861_ijsem000507.pdf

(2)ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/812/665/GCA_000812665.2_ASM81266v2//GCA_000812665.2_ASM81266v2_feature_table.txt.gz

(3)https://www.ncbi.nlm.nih.gov/genome/?term=Halocynthiibacter+arcticus