Практическая биоинформатика

Обзор протеома бактерии Candidatus Liberibacter asiaticus str. psy62

Исаев Сергей[1]

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

АННОТАЦИЯ

Данная работа посвящена исследованию протеома бактерии Candidatus Liberibacter asiaticus, штамма psy62. В её ходе рассмотрены особенности распределения длин белков, закономерности распределения белок-кодирующих, рРНК-кодирующих и тРНК-кодирующих генов в зависимости от направленности той цепочки ДНК, на которой они находятся (проверена гипотеза случайного распределения генов на прямой и обратной цепях). Были проанализированы закономерности объединения белков в квазиоперoны.

1 ВВЕДЕНИЕ

Candidatus Liberibacter – род грамотрицательных бактерий семейства Rhizobiaceae. Первая часть родового названия – Candidatus – означает, что бактерии этого рода нельзя вырастить, используя методику клеточных культур [1]. Candidatus Liberibacter asiaticus – это вид вышеназванного рода.

Candidatus Liberibacter asiaticus распространён в Азии. Там он приносит немало бед сельскому хозяйству: эта бактерия вызывает Huanglongbing (HLB), или озеленение цитрусовых, – летальную для растений инфекцию, которая передаётся через сокососущее насекомое Diaphorina citri. Побег заражённого дерева желтеет («Huanglongbing» дословно переводится с китайского как «заболевание жёлтого дракона»), а его плоды – зеленеют (полностью или частично) [2].

Геном бактерии представлен 1100 генами, которые кодируют 1016 белков. Общая длина генома – 1227328 пар нуклеотидов. Он был собран 11 июня 2013 года [3]. Для рода Liberibacter неизвестны какие-либо плазмиды, весь его геном представлен 1109 белок-кодирующими генами и 53 белок-некодирующими генами (т. е. не проходящими стадию трансляции после транскрипции).

В настоящей работе был изучены особенности протеома и генома Candidatus Liberibacter asiaticus str. psy62.

2 МЕТОДЫ

Данные о протеоме бактерии были использованы с сайта NCBI [3]. Были взяты файлы NC_012985.ptt [4] и NC_012985.rnt [5]. Для анализа данных была использована программа Microsoft Office Excel 2013.

После успешного импортирования файлов [4] и [5] к таблицам, которые там содержались, были добавлены несколько колонок:

1) **Type**, которая содержит информацию о том, что кодирует данный ген: белок или РНК (тРНК или рРНК);

2) **Start** и **Stop**, которые являются трансформированной колонкой **Location** и отражают позицию начала гена и конца гена в геноме;

3) **Gene length**, содержащую информацию о длине генов. Значения элементов столбца вычислялись по формуле $Gene\_length_n = Stop_n – Start_n + 1$;

4) **Len%3**, которая содержала в себе информацию о кратности длины гена трём: $Len\%3_n = Gene\_length_n \bmod 3$.

Следующим шагом были созданные дополнительные таблицы, помогающие при выполнении задач работы. Так, с помощью фильтра данных для каждого из изначальных файлов были составлены две таблицы, содержащие информацию только о тех генах, которые были закодированы на определённой цепи: прямой или обратной. После формировались сводные таблицы всех генов (и кодирующих белок, и кодирующих тРНК и рРНК), находящихся на (+)- или (–)-цепочке. Для того, чтобы посмотреть на закономерности распределения именно РНК-кодирующих генов, подготовили информацию о генах некодирующих РНК для каждой из исходных цепочек ДНК: прямой и обратной.

Дальше в каждой из таблиц гены упорядочили по порядку слева направо по порядку возрастания положения первого нуклеотида гена на геноме. Это позволило удобно найти дистанции между соседними генами: расстояние между генами n и (n + 1): $Distance_{n+1} = Start_{n+1} – Stop_n$. Для полноты данных возьмём $Distance_1 = 0$. Был создан соответствующий столбец **Distance** во всех таблицах, в которых имело смысл искать дистанцию.

После обрабатывания данных вышеуказанными методами можно приступить непосредственно к анализу квазиоперонов. Номер вхождения n-ного члена ряда генов – это номер вхождения предыдущего гена с прибавлением единицы, если $Distance_n = Start_n – Stop_{n–1} > C$ (C – коэффициент вхождения в квазиоперон); иначе же номер вхождения – нулевой.

На протяжении всей работы для подсчёта элементов с интересующими нас свойствами (например, при создании гистограммы или описания числа перекрывающихся генов) использовалась функция «СЧЁТЕСЛИ()», для создания простейших логических условий – функция «ЕСЛИ()». Также использовались простейшие математические функции – «МАКС()», «МИН()» и т. д.

Проверка случайности распределения генов на цепях ДНК была проверена с применением критерия согласия Пирсона по формуле:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e},$$

где f_0 – предварительно вычисленное наблюдаемое число генов; f_e – предполагаемое число генов.

Гистограмма строилась при помощи задачи карманов (с шагом в 50, например), а потом поиском элементов, подходящих под границы карманов, функцией «СЧЁТЕСЛИМН()».

3 РЕЗУЛЬТАТЫ

3.1 Анализ протеома бактерии

В геноме Candidatus Liberibacter asiaticus str. psy62 закодировано 1109 белков, статистические данные о которых представлены в таблице 1.

Таблица 1. Основные характеристики протеома Candidatus Liberibacter asiaticus str. psy62

Характеристика	Значение характеристики
Максимальная длина белка, а.к.	2210
Минимальная длина белка, а.к.	30
Медиана длин белков, а.к.	231,5
Среднее значение длин белков, а.к.	286,8

Распределение длин белков по длине отражено на гистограмме (диаграмма 1):

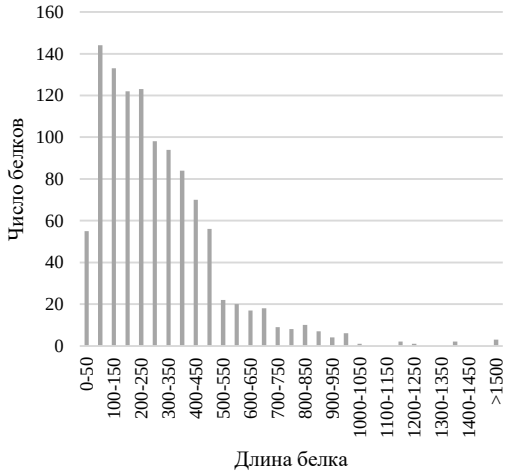

Диаграмма 1. Распределение белков бактерии Candidatus Liberibacter asiaticus, штамма psy62 по длине

3.2 Анализ генома бактерии

Был проделан анализ распределения генов в геноме и численной характеристики их разнообразия, результаты которого можно представить в виде таблицы 2.

Таблица 2. Распределение генов белков и некодирующей РНК по прямой и обратной цепям в протеоме Candidatus Liberibacter asiaticus str. psy62

Цепь\Тип гена	CDS	RNA	Всего
(+)-цепь	554	30	584
(–)-цепь	555	23	578
Всего	1109	53	1162

В процессе проверки гипотезы о том, что гены распределены по обеим цепям случайно было получено значение критерия χ^2 = 0,030981, что соответствует вероятности в том, что это совпадение, 0,05. Это означает, что гипотеза с высокой долей вероятности верна.

Некоторые числовые характеристики генома бактерии представлены в таблице 3.

Таблица 3. Статистические данные для генома бактерии Candidatus Liberibacter asiaticus, штамм psy62

Характеристика	Значение характеристики
Максимальная длина гена, п.н.	5487
Минимальная длина гена	74
Медиана длин генов	697,5
Среднее значение длин генов	837,6

В геноме Candidatus Liberibacter asiaticus str. psy62 имеются повторы, причём характерно, что ни перекрываются только гены белков: ни один ген РНК не перекрывается ни с самим собой, ни с каким-либо белком. На обратной цепи наблюдаются 57 перекрытий, на прямой – 56.

Интересно также и то, что обнаружены 45 кодирующих последовательностей, не делящихся на 3. И все эти последовательности – это гены, кодирующие т- и рРНК. Это можно объяснить в первую очередь тем, что принцип триплетности генетического кода имеет смысл только в контексте тех процессов, при которых идёт трансляция с синтезировавшейся на матрице ДНК РНК. Если ДНК превращается в некодирующую РНК, то есть в рРНК, миРНК, тРНК и проч., то принцип триплетности может не соблюдаться.

3.3 Квазиоперо ны

Пусть квазиоперон – это совокупность генов, лежащих на одной цепи ДНК, при этом расстояние между любыми двумя соседними генами не может превышать какой-то пороговой величины C (расстоянием между генами будем называть

число нуклеотидов между ближайшими крайними точками соседних генов).

Очевидно, что в зависимости от выбора пороговой величины C будет меняться число белков, входящих в квазиоперон, следует, квазиоперонный состав генома тоже будет непостоянен и зависеть от C. Наглядно эта зависимость представлена в таблице 4.

Таблица 4. Распределение генов бактерии Candidatus Liberibacter asiaticus str. psy62 по квазиоперонам в зависимости от параметра C

Пороговое расстояние	50 bp	100 bp	200 bp
Максимальное число генов	15	17	34
Максимальная длина	20025	21561	21561
(+)-цепь	94	111	123
(–)-цепь	90	111	122
Всего	184	222	245

Видно, что распределение квазиоперонов по прямой и обратной цепям примерно одинаково.

Также понятие квазиоперона позволяет нам лучше понять некоторые аспекты устройства работы генома. Так, было замечено, что у Candidatus Liberibacter asiaticus str. psy62 есть сцепленная структура генов, напоминающая настоящий оперон, которая существует в геноме бактерии в трёх копиях и обязательно в одинаковом составе: 16S rRNA, Ile tRNA, Ala tRNA, 23S rRNA, 5S rRNA, Met tRNA. Достаточно очевидно, что все эти гены находятся вместе не просто так: они синтезируются вместе, а, следовательно, вовлечены в какие-то жизненно важные метаболические пути бактерии вместе. Месторасположения сцеплений генов указано в таблице 5.

Таблица 5. Координаты повторяющихся последовательностей генов (по предположению образующих истинный оперон) у Candidatus Liberibacter asiaticus str. psy62

Gene	(+)-цепь	(+)-цепь	(–)-цепь
16S rRNA	[854295..855801]	[786255..787761]	[416812..418322]
Ile tRNA	[855982..856058]	[787942..788018]	[416555..416631]
Ala tRNA	[856071..856146]	[788031..788106]	[416467..416542]
hypothetical protein	[856371..856493]	[788331..788453]	[416120..416242]
23S rRNA	[856972..859181]	[788932..791141]	[413432..415641]
5S rRNA	[859242..859356]	[791202..791316]	[413257..413371]
Met tRNA	[859402..859478]	[791362..791438]	[413135..413211]
Mrp protein	[859502..859807]		[412806..413111]

4 ОБСУЖДЕНИЕ

В целом выполнено правило, что гены распределены случайно по двум цепочкам ДНК (гипотеза подтверждена статистически). Мало отличаются и количества квазиоперонов, и числа пересечений на соответствующих цепях. Полученные в ходе работы данные с избыточностью свидетельствуют о том, что у генов нет предрасположенности выбирать какую-то определённую цепь ДНК.

Также было замечено, что какие-то гены в геноме существуют функциональной группой – опероном, что типично для представителей прокариот. Типичен и протеом: небольшие полипептидные цепи, смещённая гистограмма распределения длин белков. Геном тоже небольшой, что является характерной чертой бактерий.

В целом в ходе работы было найдено избыточное количество черт, присущих исключительно прокариотам, что позволяет сказать, что Candidatus Liberibacter asiaticus – типичный представитель прокариот.

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Таблицы и расчёты по ссылке:
http://kodomo.fbb.msu.ru/~s.isaev/term1/3.xlsx

ССЫЛКИ

1. Wikipedia: Candidatus Liberibacter;
2. NCBI Genome: Candidatus Liberibacter asiaticus
3. NCBI Genome: Candidatus Liberibacter asiaticus str. psy62
4. NC_012985.ptt
5. NC_012985.rnt