Практикум по биоинформатике

Анализ генома и протеома бактерии *Bacillus amyloliquefaciens*, штамм MBE1283.

Белоусова Е.[1,*]

[1]Факультет биоинженерии и биоинформатики МГУ им. Ломоносова, 119991, Москва, ГСП-1, Ленинские горы, д. 1, стр. 73

РЕЗЮМЕ

В данной работе я исследовала некоторые характеристики протеома и генома данной бактерии, такие как распределение белков по функциям и по длинам, распределение генов по прямой и обратной цепям, и выявила закономерности, некоторые из которых удалось объяснить.

1 ВВЕДЕНИЕ

Bacillus amyloliquefaciens – вид свободноживущих грамположительных почвенных бактерий, имеющий важное значение как в природе, так и для человека. Филогенетическое дерево штаммов этого вида представлено на рис. 1.

В природе некоторые штаммы являются клубеньковыми бактериями, способными жить в симбиозе с бобовыми и крестоцветными растениями. *B. amyloliquefaciens* метаболизирует фитат и помогает растениям усваивать фосфор [2]. Данный вид также активирует защитные реакции растений на патогенов и синтезирует бактерицидные, фунгицидные и даже нематоцидные вещества (например, бацилломицин D, сурфактин) [3], [4].

В индустрии *B. amyloliquefaciens* используется для наработки некоторых протеаз и амилаз, данный вид также используют для наработки широко используемых в научной деятельности рестриктаз (BamH1) и рибонуклеаз (barnase) [1]. Интересная перспектива для использования *B. amyloliquefaciens* открывается в агрикультуре и медицине за счет того, что данный вид способен синтезировать антибиотики против конкретных патогенов.

Полный геном *B. amyloliquefaciens* был отсеквенирован в 2011 году, его размер составляет 3.956 Mb, содержание GC пар – около 46 процентов [5]. В данной работе мы провели анализ протеома данной бактерии и выяснили, как белки распределены в геноме по длинам, по функциям, а также, как кодирующие их гены распределены по прямой и обратной цепям.

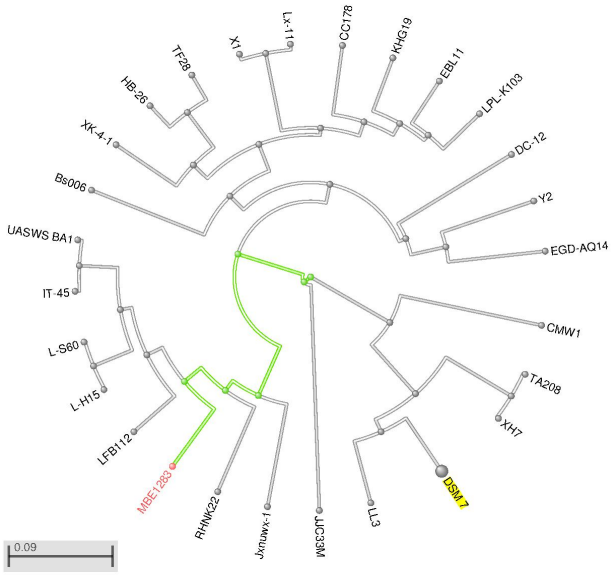

Рисунок 1. Филогенез штаммов B. amyloliquefaciens.

2 МЕТОДЫ

Данные о всех генах бактерии были взяты с сайта NCBI. Адрес файла, данные которого использовались в работе:
GCA_001483885.1_ASM148388v1_feature_table.txt.gz.

С помощью электронной таблицы Microsoft Excel я построила гистограмму распределения длин белков, с помошью фильтров было посчитано количество экспрессируемых генов белков с различными функциями, число РНК различных видов и число псевдогенов. Использовались функции МИН(), МАКС(), МЕДИАНА(), СТАНДОТКЛОН(), СРЗНАЧ(), СЧЁТЕСЛИМН(), СЦЕПИТЬ(). Под экспрессируемыми генами белков понимаются гены, с которых считываются продукты, для которых указана длина. Под псевдогенами понимаются гены, у которых в колонке class – значение pseudogene. Я также посчитала количество генов белков, РНК и псевдогенов на прямой и об

ратной цепях. Далее проверила, отличается ли вероятность распределения генов белков, РНК и псевдогенов по прямой и обратной цепям от 0,5 с помощью Python 2.7.

3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Анализ распределения длин белков

Не для каждого белкового продукта понятно, функционален ли он. Для некоторых кодирующих последовательностей в колонке class указано значение without_protein, но указано название считываемого белка. То есть, видимо, не для всех экспрессируемых в различных условиях белков клетки определена длина. Но, поскольку, таких случаев не очень много, можно принять, что проведенное статистическое исследование длин белков отражает реальность. Стандартные статистические показатели распределения длин белков приведены в таблице 1 (длины указаны в аминокислотных остатках).

Таблица 1. Стандартные статистические показатели длин белков.

Минимальная длина	26
Максимальная длина	5432
Средняя длина	309,74
Стандартное отклонение	330,57
Медиана	257

Из таблицы 1 видно, что длины белков варьируют в очень широком диапазоне (26 – 5432 а. о.), стандартное отклонение больше среднего, но средняя длина и медиана длин сильно сдвинуты к минимальной длине. Чаще всего встречаются белки с длиной в диапазоне от 25 до 731 а. о. Минимальными длинами, как правило, обладают белки-факторы споруляции, а максимальными – синтазы полимеров.

Также я построила диаграмму распределения длин белков (рисунок 2). Согласно ей большинство белков (около 97%) имеют длину в диапазоне 26 – 833 аминокислот.

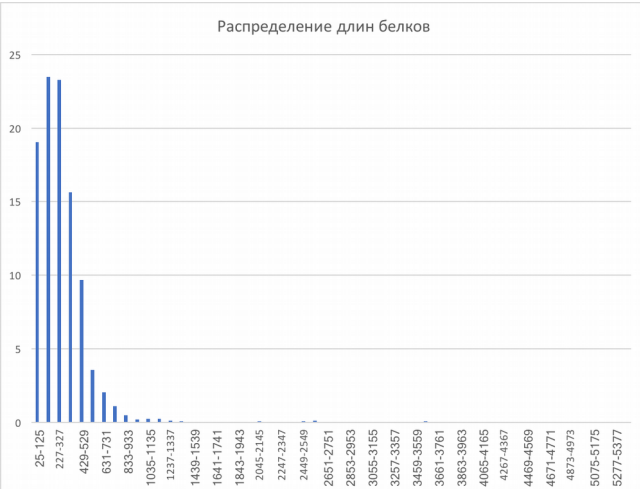

Рисунок 2. Распределение длин белков.

3.2 Распределения генов по категориям.

Как видно из таблицы 2, распределение генов белков и РНК с разными функциями очень неоднородно. То есть белков некоторых классов (трансферразы, транспортеры) значительно больше, чем других. К сожалению, для 14 процентов генов не выяснена функция. Для РНК видно, что генов тРНК больше, чем всех остальных РНК.

Таблица 2. Распределение генов по категориям.

Категория	Число генов	В процентах
Белки (экспрессируемые)	**3734**	91,54
Киназы	*105*	2,57
Трансферазы	*257*	6,3
Транспортеры	*330*	8,09
Синтазы	*151*	3,7
Редуктазы	*118*	2,89
Рибосомальные	*79*	1,94
Дегидрогеназы	*101*	2,48
Гипотетические	*573*	14,05
Остальные	*2020*	49,52
Псевдогены	**109**	2,67
РНК	**118**	2,89
рРНК	*27*	0,66
тРНК	*86*	2,11
Остальные	*5*	0,12
Всего генов	4079	100
Примерная оценка числа генов на 1 млн п. н.	102,5	2,51

3.3 Распределения генов по прямой и обратной цепям.

Также я выяснила, как гены распределены по прямой и обратной цепям в геноме *B. amyloliquefaciens*. Из таблицы 3 не понятно, являются ли данные отклонения случайными, или они играют биологическую роль. Тогда я провела эксперимент – «подкидывание монетки» по числу генов на обеих цепях 1000 раз с помощью программы в python 2.7 для генов белков, РНК, и псевдогенов. Удивительно то, что вероятность случайного распределения генов по прямой и обратной цепям очень мала. Для белков она оказалась примерно

6%, для РНК – 0%, для псевдогенов – около 50%. Конечно, для РНК и псевдогенов имеющаяся выборка невелика, возможно, этим и объясняются крайние результаты. Но для белков большинство генов находятся на обратной цепи не случайно. Но, в чем суть данного явления, остается неясным. Возможно, такая картина достигается за счет направленно происходящих инверсий

Категория	На прямой	На обратной	Всего
Белки	1808	1926	3734
РНК	81	37	118
Псевдогены	51	58	109

Таблица 3. Распределение по прямой и обратной цепям.

4 ЗАКЛЮЧЕНИЕ

Так как *B. amyloliquefaciens* исключительно важна и как лабораторный, и как фармакологический объект, можно сказать, что ее геном изучен недостаточно: для 14% белков не изучены функции, не выяснена причина, по которой большинство генов располагаются на (-) цепи.

5 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

1. Ссылка на Microsoft Excel файл: http://kodomo.fbb.msu.ru/~zhenia/term1/Belousova14.xlsx.filepart
2. Ссылка на вспомогательную программу: http://kodomo.fbb.msu.ru/~zhenia/term1/Belousova_script.png

6 БЛАГОДАРНОСТИ

Хотелось бы поблагодарить преподавателей кафедры биоинформатики за помощь и полученные мной знания и умения, однокурсников и соседей по комнате за помощь и увлекательное обсуждение результатов.

7 СПИСОК ЛИТЕРАТУРЫ

[1] Статья о *Bacillus amyloliquefaciens*: https://en.wikipedia.org/wiki/Bacillus_amyloliquefaciens_-_Status_as_a_species
[2] Статья об участии *Bacillus amyloliquefaciens* в защитных реакциях растений: https://www.ncbi.nlm.nih.gov/pubmed/29090851
[3] Статья о выработке нематоцидных веществ: https://www.ncbi.nlm.nih.gov/pubmed/29077011
[4] Статья об антибактериальных веществах: https://www.ncbi.nlm.nih.gov/pubmed/?term=Antimicrobial+and+bactericidal+impacts+of+Bacillus+amyloliquefaciens+CECT+5940+on+fecal+shedding+of+pathogenic+bacteria+in+dairy+calves+and+adult+dogs
[5] Описание генома на сайте NCBI: https://www.ncbi.nlm.nih.gov/genome/?term=Bacillus+amiloliquefaciens+MBE1283
[6] Адреса на скачивание геномов разных штаммов: https://www.ncbi.nlm.nih.gov/genome/genomes/848