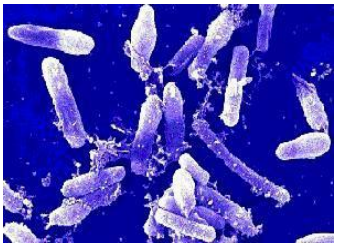
Обзор генома бактерии *Bacillus Cereus* strain A1

Беляева Юлия[1*]

[1]Факультет Биоинженерии и Биоинформатики МГУ им. М. В. Ломоносова

АННОТАЦИЯ



Данная работа посвящена исследованию генома бактерии *Bacillus Cereus* штамм A1, изучению статистических характеристик протеома бактерии, проверке гипотезы о случайном распределении генов по прямой и обратной цепям ДНК. В ходе работы на Python был написан скрипт, который может быть полезен для моделирования опыта «бросания монетки», в него могут вноситься необходимые числовые корректировки.

1 ВВЕДЕНИЕ

Bacillus cereus –вид грамположительных почвенных бактерий. Некоторые штаммы вредны для людей и вызывают болезни пищевого происхождения, в то время как другие могут быть полезны в качестве пробиотиков для животных. [1]

B. cereus является хемоорганогетеротрофом, факультативным анаэробом, способным к нитроредукции; растет на простых питательных средах, на плотных образует плоские, мелкобугристые матовые колонии. Клетки 1 × 3—4 мкм, эндоспоры расположены центрально. Жгутики расположены перитрихиально. [1], [4]

Штамм A1 – недавно выделенный производитель водорода, способный утилизировать биоотходы (к примеру, сточные воды. [2]

B. Cereus A1, демонстрирует удивительную способность превращать аммоний в газообразные азотные соединения в полностью аэробных условиях при автотрофном или гетеротрофном образе жизни.

Размер генома составляет 5352307 пар оснований. Имеется три плазмиды – pBCA1, pBCA2, pBCA3. [3]

2 МАТЕРИАЛЫ И МЕТОДЫ

2.1. Используемые ресурсы и программы

Поиск и загрузка файла для работы с геномом исследуемой бактерии проводился с помощью NCBI.

Для составления таблиц, построения диаграмм и анализа данных был использован Microsoft Excel (русскоязычная версия).

Опыты проверки гипотезы о случайном распределении генов по цепям выполнялись скриптом, который был написан на Python 2.7.

Для работы были использованы данные, полученные в ходе предыдущей работы (сопроводительные материалы [3])

2.2. Список основных используемых функций при работе в Excel

=СЧЁТЕСЛИМН()
=СУММ()
{=ЧАСТОТА()}
=СЦЕПИТЬ()
=ЕСЛИ()
=ABS()
=ЕСЛИ()
=СЧЁТЕСЛИ()

2.2. Определение числа генов белков и генов РНК по категориям

Для определения числа генов белков каждой категории был использован фильтр по значению “CDS” (столбец A:A). Полученная таблица была скопирована на другой лист рабочего файла.

Далее для поиска генов белков категории “гипотетические” применялась функция [=СЧЁТЕСЛИМН()], которая может осуществлять поиск соответствий по нескольким условиям. В столбце N:N – “name” указывается название белка, определяющее его функцию. Для подсчёта числа генов транспортных белков ставим фильтр на слово “transport” по этому столбцу; исключим из полученной выборки пермеазы (являются ферментоподобными белкам), транспортные белковые субъединицы оставим в составе выборки. Посмотрим на количество оставшихся белков и занесем данные в таблицу.

Поставим новый фильтр на колонку N:N – “ribosomal” для подсчета генов рибосомальных белков. Просмотрим нашу выборку, чтобы исключить белки, имеющие, по предположениям, иную функцию (например, ABC-F type ribosomal protection protein, ribosomal subunit interface protein или синтетазы, которые являются ферментами, то-есть, несут функцию, отличную от структурной). Таким образом, в нашей выборке осталось 58 генов рибосомальных белков.

Остальные белки считаем следующим образом: из общего числа белковых структур (“CDS” в столбце A:A выберем те, что несут во второй колонке значение “with_protein”, и вычтем из полученного числа сумму количеств белков, входящих в состав предыдущих трех групп). Получилось, что общее число белков равно 5418.

Для подсчета генов РНК, относящихся к трем различным группам, будем использовать фильтр по столбцу A:A – “tRNA”, “rRNA”. Количество РНК, относящихся к группе «остальные» определим следующим образом: из общего числа

(фильтр на “tRNA”, “rRNA”, “ncRNA”, “tnRNA”, то-есть, на все “РНК-группы”) вычтем уже известное нам количество транспортных рибосомальных РНК (а точнее, количество их генов).
Результаты представлены в Таблице 1, в сопроводительных материалах.

2.3. Распределение длин белков *B.Cereus* strain A1

Для построения диаграммы длин белков воспользуемся данными из таблиц, которые были сделаны в ходе 13 практикума (три колонки – верхняя граница диапазона, диапазон, частота встречаемости). На основе последних двух колонок построим диаграмму длин белков из протеома бактерии Bacillus Cereus A1.

2.4. Определение числа генов белков, РНК, псевдогенов на прямой и обратной цепях ДНК

Для построения таблицы числа генов белков, псевдогенов и генов РНК на прямой и комплементарной цепях ДНК применим к исходной таблице(“first_table”) фильтры по “CDS” и “RNA” (всех типов), заведем для них отдельные листы в том же excel – файле. Применим несколько видоизменную функцию [=СЧЁТЕСЛИМН()] с добавлением условия о присутствии + или – в столбце J:J (для подсчёта генов белков и РНК; подсчет псевдогенов проведем с помощью фильтров по “gene” в столбце A:A исходной таблицы совместно с условиями о “pseudogene ” и +/- в столбцах B:B и J:J соответственно.

2.5. Проверка гипотезы о распределении генов по цепям случайно, с вероятностью 0,5

Штамм A1 содержит 5648 генов, из которых 77 - псевдогены; исключим их из рассматриваемой выборки. Тогда, согласно Таблице 2, в стартовой выборке для проверки гипотезы о случайном распределении участвует **5571** генов, из которых 2749 находится на прямой цепи ДНК, 2822 – на обратной. При случайном распределении на каждой цепи должно было быть 2785,5 генов (в данной случае воспользуемся нецелым числом генов для проверки вероятности). Таким образом, отклонение составляет 36,5 генов.
С помощью скрипта, написанного на Python (сопроводительные материалы), проведем эксперимент с подбрасыванием монетки (7000 раз для 5571 гена).
Пусть знак "+" - это 1 (прямая цепь), знак "-" - это 0 (обратная цепь) в проводимом эксперименте. Скрипт записывает результаты работы в файл.
Данные из файла выведем на лист файла excel и обработаем (посчитаем отклонение от 0,5 – распределения, проверим на критерий «больше или равно» отклонения, полученного из табличных данных). Далее проставим «+» или «-» в колонке D:D в случаях соответствия или несоответствия критерию; посчитаем количество “+” и найдем отношение их количества к 7000. Сравним с 0,5 (будем считать, что при значении отношения >0,5 предполагаемая гипотеза верна).

3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1. Геномный состав

Штамм содержит 5648 генов (с учетом псевдогенов). Количество генов белков, генов РНК, псевдогенов, а также их распределение по цепям ДНК представлено в Таблице 1.
В геноме *B. Cereus* A1 49,43% генов расположены на прямой, 50,57% - на обратной цепях ДНК.
По данным NCBI, геном *B. Cereus* A1 содержит 5352307 пар нуклеотидов. Количество генов на 1 млн. пар нуклеотидов равно 1055 (с учетом псевдогенов), 1041 (за исключением псевдогенов).

3.2. Длины белков из протеома *Bacillus Cereus* A1

На рисунке 1 представлена диаграмма распределения длин белков по различным диапазонам. Большинство белков имеют длину, попадающую по значению в диапазон 26 – 400 аминокислотных остатков (78,11%). Количество белков, попадающих в диапазоны 201 - 1300 резко уменьшается на протяжении этого интервала длин (рисунок 2).
Наибольшая длина – 5010 аминокислотных остатков (“cell surface protein”, белок клеточной поверхности), минимальная – 26 (“K+-transporting ATPase subunit F”, белок субъединицы F K+ - транспортирующей АТФ-азы; “stage V sporulation protein M”, белок M V стадии споруляции).
Средняя длина белка, синтезируемого исследуемой бактерией составляет около 302 пар аминокислотных остатков.
Результаты статистического обзора приведены в таблице 2.

Таблица 1. Категории генов, распределение по цепям ДНК

	Белки	
	+	-
рибосомальные	40	18
транспортные	153	192
гипотетические	726	723
остальные	1712	1854
общее количество	5418	
	РНК	
	+	-
транспортные	75	31
рибосомальные	39	3
остальные (ncRNA, tmRNA)	4	1
общее количество	153	
	Псевдогены	
	+	-
Псевдогены	43	34
Общее число	77	

Таблица 2. Статистика по длине белков

Параметр	Значение
Минимальная длина	26
Максимальная длина	5010
Средняя длина	302,116
Стандартное отклонение	269,936
Медиана	248

Рисунок 1. Диаграмма длин белков

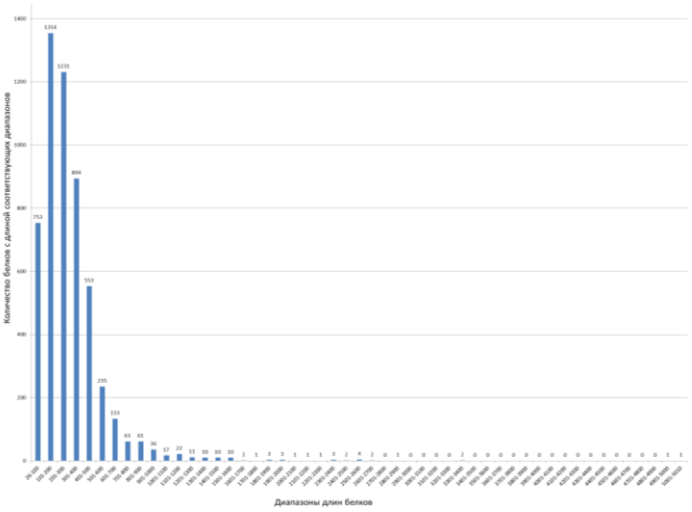
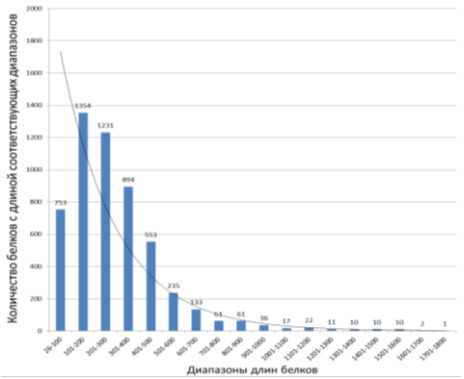


Рисунок 2. Диаграмма белков длины 26 – 1800 с экспоненциальной линией тренда



3.3. Распределение генов белков и РНК по прямой и обратной цепям ДНК

В результате опытов с «подбрасыванием монетки», проведенных Python было выяснено, что отклонение от 0,5 – вероятностного распределения наблюдается в 2320 случаях из 7000 (33,1%), 0,331 - в доле, таким образом, мы можем сказать, что гипотеза о случайном распределении генов (с вероятностью 0,5) не верна, согласно установленному нами значению истинности. То-есть, в 66,9% случаях гены распределены по цепям не случайно. Стоит заметить, что % выполнения условия гипотезы может различаться, в зависимости от количества проведенных опытов.

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

1. http://kodomo.fbb.msu.ru/~belyaevajuly/term1/genom.xlsx

2. http://kodomo.fbb.msu.ru/~belyaevajuly/term1/exp.py

СПИСОК ЛИТЕРАТУРЫ

1. https://ru.wikipedia.org/wiki/bacillus_cereus
2. https://www.ncbi.nlm.nih.gov/genome/157
3. ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/bacillus_cereus/latest_assembly_versions/gca_000635895.2_asm63589v2/gca_000635895.2_asm63589v2_assembly_report.txt
4. Микробиология. Гусев М.В., Минеева Л.А. 4-Е Изд., Стер. - М.: Академия, 2003. — 464 С. В Учебнике (1-Е Изд. - 1978 Г.)