Биоинформатика, микробиология

Мини-обзор генома и протеома бактерии *Salmonella enterica* subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1735

Титова Алена Олеговна[1,*],
[1]ФГБОУ ВО МГУ им. М.В. Ломоносова, факультет биоинженерии и биоинформатики

Ключевые слова: Salmonella enterica, Excel, геном, протеом
Дата публикации: 21.12.2017

РЕЗЮМЕ
В данной работе был исследован геном и протеом бактерии Salmonella enterica: проанализировано количество различных по функциям белков и РНК, посчитано распределение генов по прямой и обратной цепи ДНК и примерное число квазиоперонов, построена гистограмма длин белков и исследованы пересечения генов.

1 ВВЕДЕНИЕ

Таблица 1. Таксономическое положение данной бактерии [5]

Домен	Prokaryota
Царство	Bacteria
Тип	Proteobacteria
Класс	Gammaproteobacteria
Порядок	Enterobacteriales
Семейство	Enterobacteriaceae
Род, вид	*Salmonella enterica*

S. enterica является грамотрицательной палочковидной подвижной бактерией (рис.1, рис.2) из семейства Enterobacteriaceae (см. таблицу 1), которая представляет особый интерес для медицинского и биологического сообщества ввиду ее способности вызывать инфекционные заболевания у людей и животных – сальмонеллезы (острые кишечные инфекции). Бактерия была названа в честь доктора Дэниела Э. Сэлмона, который впервые выделил и идентифицировал этот вид более 100 лет назад.[3], [5]

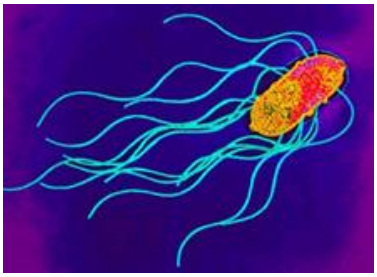


Рис.1. Фото S.enterica, полученное с помощью трансмиссионного электронного микроскопа [3]

Геном изучаемого штамма состоит из одной кольцевой хромосомы (идентификатор GenBank: CP007584.2) и одной кольцевой плазмиды (идентификатор GenBank: CP014707.1).[4] Всего генов, включая псевдогены, 5087, а число пар оснований 4945533 (~5 мегабаз).[1]

Несмотря на то, что геном *S. Enterica* был секвенирован еще в 2001 году[7] (хотя геном данного конкретного штамма – лишь в 2016[9]) и к настоящему моменту сальмонелла уже довольно хорошо изучена, авторам было интересно внести свою лепту в анализ информации об этом биологическом объекте. В данной работе был применен биоинформатический подход к исследованию генома и протеома бактерии. Полученные результаты позволяют выдвинуть некоторые гипотезы и еще раз подтвердить полученные ранее данные.

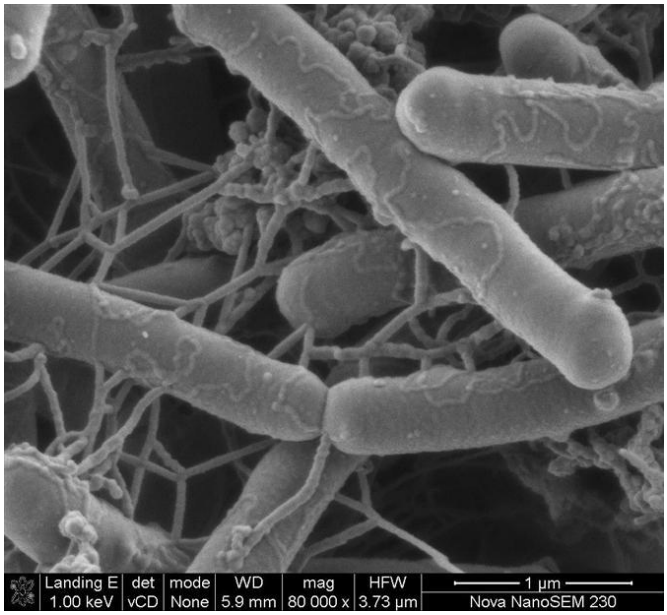


Рис.2. Фото *Salmonella* sp., полученное с помощью электронного микроскопа [8]

*По всем вопросам пишите: alena.titova.8m@mail.ru

2 МАТЕРИАЛЫ И МЕТОДЫ

Данные о геноме и протеоме бактерии были взяты с сайта NCBI.[2] Обработка данных производилась в программе Microsoft Office Excel 2013, были использованы встроенные функции.

При подсчете количества генов транспортных, рибосомальных, гипотетических и других белков, а также генов тРНК, рРНК и других типов РНК применялся фильтр по значениям ячеек в столбцах. При этом велся поиск не по точному совпадению, а по наличию заданной подстроки: так, для транспортных белков в колонке “name” искалась подстрока “transport”, для рибосомальных - “ribosomal”, для гипотетических - “hypothetical”, а остальные получались путем вычитания уже посчитанных белков из общего количества белок-кодирующих генов, для которых обязательным и достаточным условием считалось значение “protein_coding” в графе “class”; для рРНК искалось точное совпадение с “rRNA” в столбце “feature”, для тРНК - “tRNA”, остальные же РНК были получены вычитанием уже посчитанных РНК из всех имеющихся РНК, которые в свою очередь искались по наличию подстроки “RNA” в колонке “feature”.

Подсчет количества квазиоперонов выполнялся с тем допущением, что квазиопероном считался набор генов на одной цепи, расстояние между которыми не превышает 100 нуклеотидов.

Для проверки гипотезы о случайном распределении генов по прямой и обратной цепи ДНК была проведена серия экспериментов, симулированных с помощью скрипта на языке программирования Python версии 3.4.

Наконец, для перекодировки ID белков использовался инструмент Retrieve/ID mapping на сайте UniProt.[11]

3 РЕЗУЛЬТАТЫ

3.1 Гены функционально различных белков

Из таблицы 2 видно, что количество закодированных рибосомальных белков сильно уступает числу транспортных и гипотетических. Между тем, в геноме есть целых 1059 генов неаннотированных белков, чья функция не определена. Остальные 3245 белков, закодированных в геноме, выполняют иные функции.

Таблица 2. Количество генов функционально разных белков

количество генов белков			
рибосомальные	транспортные	гипотетические	остальные
99	418	1059	3245

3.2 Гены различных типов РНК

Таблица 3. Количество генов разных типов РНК

количество генов РНК		
тРНК	рРНК	остальные
87	22	9

Таблица 3 демонстрирует, что в геноме *S.enterica* закодировано количественно больше транспортных РНК, чем каких бы то ни было других (их 87). Гены рибосомальных РНК встречаются гораздо реже, их всего 22. Также в геноме присутствует совсем небольшое количество некодирующих, антисмысловых, сигнал-распознающих РНК и РНК рибонуклеазы Р.

3.3 Генно-нуклеотидный состав

В геноме изучаемой бактерии всего 4945533 пар нуклеотидов (считается и хромосома, и плазмида) и 5087 генов (включая и псевдогены) [1]. Таким образом, получается, что число генов в пересчете на 1 миллион пар нуклеотидов составляет около 1029.

Таблица 4. Генно-нуклеотидный состав

всего генов в геноме	5087
всего нуклеотидов в геноме	4945533
число генов на 1 млн bp	1028,61

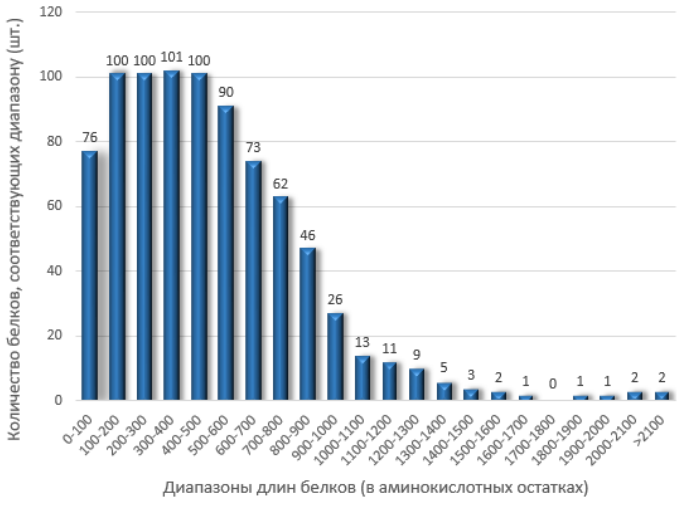
3.4 Распределение генов по прямой и обратной цепям

В таблице 5 представлено распределение генов белков, генов РНК и псевдогенов по прямой и обратной цепям ДНК. Нетрудно заметить, что на обеих цепях находится примерно одинаковое количество генов.

Таблица 5. Распределение генов по прямой и обратной цепям

	число генов белков	**число псевдогенов**	**число генов РНК**
на прямой цепи ДНК	2462	67	45
на комплементарной цепи ДНК	2359	81	73

3.5 Распределение белков по их длинам



Гистограмма 1. Распределение белков по их длинам

На гистограмме 1 по горизонтальной оси отложены диапазоны длин белков (в аминокислотных остатках), каждый диа-

пазон по 100 аминокислотных остатков, кроме последнего (он задан просто как “более, чем 2100”). Большинство белков попадает в длину от 100 до 600 аминокислот, а далее, чем больше длина, тем меньше белков ей соответствуют.

Таблица 6 указывает на то, что минимальная длина белка, закодированного в геноме бактерии – 14 аминокислотных остатков (это так называемый “trp operon leader peptide”, очень маленький пептид, входящий в систему регуляции экспрессии триптофанового оперона), максимальная – 5559 (огромный мембранный белок), а в среднем длина белков составляет около 450 аминокислот.

Таблица 6. Статистические данные по белкам

min	max	среднее арифметическое	медиана	среднее квадратичное отклонение
14	5559	488,20995	434,5	384,17499

3.6 Квазиопероны

Квазиопероном назовем максимальную последовательность генов, закодированных на одной цепи ДНК с промежутками между генами не более порога 100 пар нуклеотидов.[6] Получилось, что всего в геноме 2617 квазиоперонов, 1314 на «+» цепи и 1303 на «-». Таким образом, по цепям ДНК они распределяются равномерно, независимо. Более того, на гистограммах 2 и 3, а также по значению среднего арифметического числа генов в квазиопероне можно сделать вывод, что более многочисленны небольшие опероны, содержащие в себе всего 1-2 гена. Впрочем, количество генов вплоть до 6 в одном квазиопероне также встречается довольно часто, а вот больше – уже редкость. Удивительно, с какой точностью эти результаты сходятся для “+” и для “-“ цепей: распределение почти одинаковое.

Гистограмма 2. Распределение квазиоперонов «+» цепи по числу генов в них

Гистограмма 3. Распределение квазиоперонов «-» цепи по числу генов в них

3.7 Пересечения генов

Пересечениями генов назывались те случаи, когда старт-кодон следующего гена лежит до стоп-кодона предыдущего гена. Всего таких пересечений в геноме бактерии насчиталось 1072 (из 5087 генов всего), то есть около 20%. Более того, из таблицы 7 заметно, что пересекающихся генов на одной цепи вдвое больше, чем таковых на разных цепях.

Таблица 7. Распределение пересекающихся генов и их % от общего числа генов

всего пересекающихся генов	на одной цепи	на разных цепях
1072	774	298

% пересекающихся генов
20,59%

Таблица 8 демонстрирует сдвиги рамки считывания при пересечении генов. Видно, что если гены пересекаются на одной цепи, то рамка считывания обычно сдвигается либо на 1, либо на 2 нуклеотида, а если пересекаются гены разных цепей, то рамка может сдвинуться на 1 или на 2 нуклеотида, а может и не сдвигаться с примерно равной вероятностью.

Таблица 8. Сдвиг рамки считывания у пересекающихся генов

сдвиг рамки считывания	на одной цепи	на разных цепях
0	29	92
1	289	101
2	456	105

3.8 Гипотеза о случайности распределения генов по цепям ДНК

В геноме 4821 генов белков[1], из них 2462 на одной цепи, 2359 - на другой. Отклонение от ожидаемого числа 2410.5 равно 51.5. Чтобы ответить на вопрос, возможно ли получить такое или большее отклонение при независимом случайном выборе цепочки для каждого гена, устроим серию экспериментов: выберем рандомное число из (0, 1) 4821 раз, посчитаем, сколько раз выпал ноль (ген на + цепочке). Вычислим, каково отклонение этого числа от ожидаемого 2410.5 и сравним с наблюденным нами отклонением 51.5. Если больше или равно - ставим условный «плюс», если меньше – ставим условный «минус». Повторим эксперимент 1000 раз для верности, а затем посчитаем число полученных «плюсов» и «минусов».[6]

Рис.3. Запуск скрипта python

Скрипт выдает в консоль два числа: первое – количество «плюсов», а второе – количество «минусов». На рис.3 показаны 5 независимых запусков программы. Видим, что из 1000 экспериментов лишь примерно 14% показывают отклонение, большее или равное наблюдаемому нами.

3.9 Статистика белков по категориям достоверности их существования

В геноме изучаемого организма 4969 белок-кодирующих последовательностей (CDS), но 148 из них относятся к псевдогенам и находятся в нерабочем состоянии. Еще для 8 CDS не указаны идентификаторы соответствующих белков. Из оставшихся 4813 CDS для 3378 кодируемых ими продуктов были найдены соответствующие белки в базе данных UniProt (найдено порядка 70%)[11]. Результаты их сортировки по категориям достоверности существования представлены в таблице 9.

Таблица 9. Статистика белков по категориям достоверности их существования

Категории достоверности существования белков	количество белков	% белков
доказательство на уровне белка	1	0,03%
доказательство на уровне транскрипта	10	0,30%
предсказаны по гомологии	1393	41,24%
предсказаны	1974	58,44%

Категории достоверности существования означают следующее:

- доказательство на уровне белка – есть экспериментальное подтверждение существования белка (масс-спектрометрия, рентгеноструктурный анализ, ядерно-магнитно-резонансная спектроскопия, обнаружение белка антителами и др.);
- доказательство на уровне транскрипта – существование белка строго не доказано, но есть подтверждение существование его транскрипта (наличие комплементарной ДНК, результаты ПЦР с обратной транскрипцией или Нозерн-блоттинга);
- предсказаны по гомологии – существование белка вероятно, поскольку у близкородственных видов есть его ортологи;
- предсказаны – нет доказательств ни на уровне белка, ни на уровне транскрипта, ни на уровне гомологии.[10]

3.10 Сравнение генов на хромосоме и на плазмиде

Была выдвинута гипотеза, что на плазмиде гены расположены компактнее, чем на хромосоме, и соответственно, пересечений среди них больше. Также мы предположили, что плазмидные гены в среднем короче, поскольку сама плазмида гораздо меньше хромосомы по количеству нуклеотидов (это не очень логичное предположение, скорее на интуитивном уровне).

Выяснилось, что действительно, процент пересекающихся генов на хромосоме равен ~20%, а на плазмиде он составляет 62%.

Таблица10. Проценты пересекающихся генов на хромосоме и на плазмиде

% пересечений на хромосоме	% пересечений на плазмиде
19,57%	61,60%

Длины белков, хоть и незначительно, но в среднем меньше у соответствующих генов, расположенных на плазмиде (среднее арифметическое длины хромосомных белков составляет 300 аминокислотных остатков, в то время как плазмидных – 242; но в данном случае гораздо лучше ориентироваться на медиану – медианный хромосомный белок имеет длину 258, а плазмидный - всего 188). На хромосоме гораздо больше разброс значений.

Таблица 11. Статистика длин белков, закодированных на хромосоме и на плазмиде

	min	max	среднее	медиана
хромосомные белки	14	5559	300,16	258
плазмидные белки	53	1543	241,56	188

4 ОБСУЖДЕНИЕ

1. Рибосомальных белков меньше всего, поскольку рибосомы – небольшие немембранные органеллы, состоящие только из нескольких десятков консервативных белков и рРНК. Тем временем транспортных белков в разы больше, потому что почти для каждого соединения нужен свой уникальный специфичный транспортер, и таковых получается много. Более того, у транспортеров бывает еще и не одна субъединица, а несколько разных, и каждая является полноценным пептидом. Также отмечено удивительно много «неопознанных», неаннотированных гипотетических белков. Видимо, дело в том, что биоинформатики пока не могут опознать их функции.
2. С рРНК, судя по всему, действует такая же логика, как и с рибосомальными белками. Их мало и они консервативны, поэтому никакой вариативности у них не наблюдается. Транспортные РНК присутствуют в количестве 87 штук. Каждому кодону соответствует своя тРНК, и если кодонов всего около 60, то, наверное, есть дублирующиеся тРНК. Оставшие малочисленные типы РНК – некодирующие РНК, антисмысловые РНК, сигнал-распознающие РНК и РНК рибонуклеазы Р – вообще звучат очень загадочно.
3. Генно-нуклеотидный состав показывает достаточно высокую плотность генов: более 1000 генов на миллион пар нуклеотидов. По сравнению с эукариотами это очень высокая частота покрытия генами. Скорее всего, тут также играют роль перекрытия между генами. И, безусловно, у прокариот имеет место более жесткий естественный отбор, который заставляет их экономить ресурсы на репликацию, и соответственно, место в геноме, поэтому они не могут «позволить себе» пустые места, не занятые генами, в отличие от эукариот, у которых полно «мусорной ДНК» и некодирующих участков.
4. Проверяя гипотезу про случайность распределения генов по цепям, мы получили следующее. Из 1000 экспериментов лишь примерно 14% показывают отклонение, большее или равное наблюдаемому нами. Событие, происходящее с вероятностью 14% (примерно 1 раз из 7) - не настолько редкое, чтобы считать его маловероятным, поэтому мы все же делаем вывод, что гипотеза верна и гены распределены случайно. Однако интересно, почему же наблюдается подобный сдвиг вероятности.
5. Что касается распределения белков по их длинам, то здесь результаты вполне предсказуемы. Большинство белков находятся в диапазоне от 100 до 600 аминокислот, и лишь несколько являются очень длинными или очень короткими.
6. У данной бактерии наблюдается 20% пересечений генов, что является довольно высоким результатом. В большинстве случаев рамка считывания сдвигается на 1 или 2 нуклеотида, если пересекаются гены на одной цепи ДНК, и очень редко остается такой же. Вообще непонятно, как рамка в таком случае может не измениться, ведь тогда стоп-кодон предыдущего гена станет и стоп-кодоном для следующего. Наверное, имеют место какие-то механизмы проскакивания стоп-кодона, либо же это ошибки предсказания старта транскрипции. В случае же, когда гены пересекаются на разных цепях, все логично: рамка считывания может остаться прежней, а может сдвинуться на 1 или 2 нуклеотида с примерно равной вероятностью.
7. Лишь для одного белка существует явное доказательство его существования, для 10 белков имеются доказательства существования соответствующих транскриптов, 41% белков предсказаны по гомологии, и оставшиеся 58% не имеют четкого доказательства существования.
8. Гипотеза, предполагающая более компактное расположение генов на плазмиде, нежели на хромосоме, может считаться доказанной.

5 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Со всеми представленными расчетами, таблицами и диаграммами можно ознакомиться по предоставленным ниже ссылкам.

Ссылка на файл xlsx:
http://kodomo.fbb.msu.ru/~titova.alena/documents/materials.xlsx

Ссылка на скрипт python:
http://kodomo.fbb.msu.ru/~titova.alena/documents/excel.py

6 БЛАГОДАРНОСТИ

Коллектив авторов выражает свою глубокую признательность преподавателям по биоинформатике ФББ МГУ, ученым, секвенировавшим геном *S.enterica* и исследовавшим этот организм, ребятам из NCBI и UniProt, сотрудникам всеми любимого Google, преподавателям по микробиологии и основам молекулярной генетики из старого доброго РНИМУ им. Пирогова, а также учителю биологии школы №1253 Кохову А.В. Наконец, передаем привет любимым одногруппникам.

7 СПИСОК ЛИТЕРАТУРЫ

[1] Обзор генома организма данного штамма (genome assembly and annotation report)
https://www.ncbi.nlm.nih.gov/genome/152?genome_assembly_id=275381

[2] Файлы RefSeq по данному организму
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/988/525/GCF_000988525.3_ASM98852v3

[3] Статья на MicrobeWiki о Salmonella enterica, фото и описание
https://microbewiki.kenyon.edu/index.php/Salmonella_enterica_NEU2011

[4] Сборная информация по геному данного организма (assembly)
https://www.ncbi.nlm.nih.gov/assembly/GCA_000988525.3

[5] Информация о *Salmonella enterica*, усредненные значения для всех сероваров и штаммов (organism overview) https://www.ncbi.nlm.nih.gov/genome/152

[6] Образовательный портал факультета биоинженерии и биоинформатики МГУ https://kodomo.fbb.msu.ru/wiki/2017/1/pr13

[7] Журнал Innovations report: новость о секвенировании генома *S.enterica* serovar Typhi и Typhimurium со ссылками на оригинальные статьи в Nature http://www.innovations-report.com/html/reports/life-sciences/report-5573.html

[8] Сайт производителя электронных микроскопов, фото сальмонеллы https://www.fei.com/image-gallery/salmonella-bacteria/

[9] Публикация Complete and Closed Genome Sequences of 10 Salmonella enterica subsp. enterica Serovar Anatum Isolates from Human and Bovine Sources. DOI: 10.1128/genomeA.00447-16 https://www.ncbi.nlm.nih.gov/pubmed/27257192

[10] UniProt, категории достоверности существования белков http://www.uniprot.org/help/protein_existence

[11] UniProt, Retrieve/ID mapping - инструмент для перекодировки идентификаторов белков http://www.uniprot.org/uploadlists/