Биоинформатика

Краткий анализ генома и протеома бактерии Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1735

Галкин С.О.[1]

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

РЕЗЮМЕ

В данном обзоре посредством возможностей программы Microsoft Office Excel был произведен краткий анализ генома бактерии Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1735 с целью наглядно представить уже существующую информацию об этой бактерии, и, на основе этого представления, сделать выводы. В процессе работы были выполнены следующие действия: построение гистограммы длин белков, изучение распределения генов по + и – цепям ДНК, подсчет числа генов белков и генов РНК по категориям (транспортные, рибосомальные, некодирующие и транспортно-матричные), подсчет количества квазиоперонов в геноме, была собрана статистика о пересечениях генов и достоверности существования белков бактерии (правда, данных о белках конкретно данного штамма не нашлось, для анализа использовался протеом близкого штамма Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1736).

1 ВВЕДЕНИЕ

Объект обзора, бактерия Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1735, как следует из её названия, относится к роду Salmonella, палочкообразных Грам-негативных неспорообразующих энтеробактерий, которые инфицируют миллионы людей ежегодно. Salmonella anatum – распространённая причина сальмонеллеза, опасной инфекции человека и животных. Изучаемая бактерия может быть выделена из крупного рогатого скота, телят, свиней, лошадей, собак, куриц, гусей, индейки, яиц, сухого молока, рыбы. Этот микроорганизм также имеет важное экономическое значение: например, в 2013 году власти США понесли 3.6 млрд. $ убытка из-за его деятельности. Геном бактерии состоит из одной хромосомы длиной в 4844415 п.н. и плазмиды размером 101118 п.н. Предполагается, что геном содержит 4844 гена (конечно же, большая часть из них – лишь предсказанные)[1]. В виду важного социально-экономического значения этого организма, необходимость его изучения очевидна. Цель же этой работы – систематизация и интерпретация уже накопленных знаний об этой бактерии.

2 МЕТОДЫ

Как уже упоминалось, большая часть работы была выполнена с использованием инструментария программы Microsoft Office Excel (фильтры, сортировка, базовые функции в роде «СЧЕТЕСЛИ», «СЧЕТЕСЛИМН», «СЛУЧМЕЖДУ» и т.д.). При работе с большим количеством данных использовались сводные таблицы. Все данные для работы были получены из открытой базы данных NCBI[2] (а именно был скачан файл-описание генома «GCA_000988525.3_ASM98852v3_feature_table.txt» , который впоследствии использовался как основной источник данных для работы. Для получения данных о протеоме использовался сайт uniprot.org[3].

3 РЕЗУЛЬТАТЫ

Все полученные данные, которые используются в данной работе, находятся в сопроводительном файле «Galkin_pr13.xlsx» в разделе «Сопроводительные материалы». Далее в подразделах будут подробно рассмотрены результаты проделанного анализа исходной информации.

3.1 Гистограмма длин белков из протеома Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1735

Гистограмма представлена на Рисунке 1. При построении диаграммы псевдогены не учитывались. Как видно из рисунка,

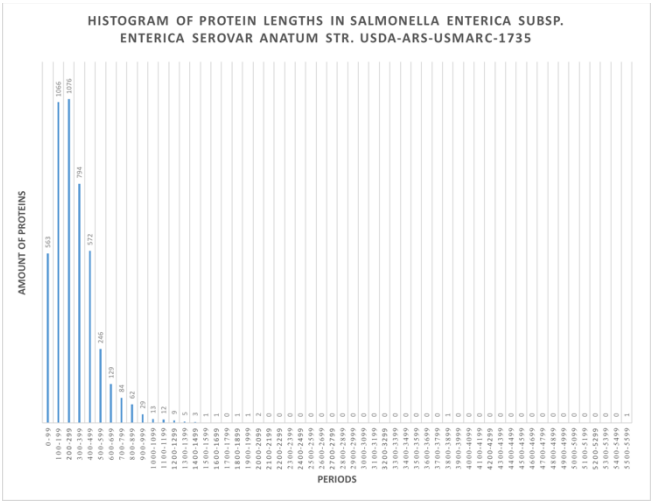


Рис. 1. Гистограмма длин белков в Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1735

большая часть белков имеет длину от 100 до 300 аминокислот.

Самый длинный белок имеет длину в 5559 аминокислот и возможно участвует в процессах узнавания подходящего субстрата, а самый короткий состоит из 21 аминокислоты и является лидерным пептидом треонинового оперона.

3.2 Распределение генов по цепи ДНК, их категории и проверка гипотезы об их случайном распределении по + и - цепи

То, как распределены гены по + и – цепям ДНК, отражено в Таблице 1, а количество генов, принадлежащих каждой категории, отражено с помощью Таблицы 2.

Таблица 1. Распределение генов белков и РНК по цепям ДНК

Strand	Protein genes	RNA genes
+	2391	45
-	2280	77

Таблица 2. Число генов белков и РНК по категориям, количество генов каждой категории в расчете на 1 млн. пар нуклеотидов

Feature	Amount of genes	Amount of genes per 1 million of nucleobases
CDS	4671	944,489
ncRNA	12	2,426
rRNA	22	4,448
tmRNA	1	0,202
tRNA	87	17,592

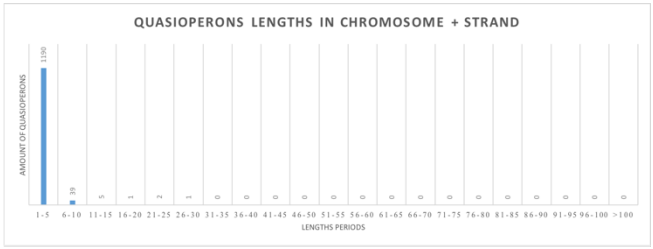
Вероятность, того, что 4793 гена распределились по двум цепям ДНК случайно составляет около 0.26, что не опровергает гипотезу о случайном распределении генов при пороге, равном 0.05. Однако это истинно для хромосомы и целого генома, для плазмиды же получаются значения вероятности существенно меньше 0.05. Попытка объяснения данного явления дана в соответствующем разделе обсуждений.

3.3 Квазиоперoны в геноме бактерии

Все результаты для данного раздела были получены из допущения, что квазиопероном будет считаться совокупность генов, располагающихся на одной цепи и на расстоянии не более ста нуклеотидов друг от друга. При желании, для своих целей, в сопроводительном файле можно изменить принимаемый размер шага. Для каждой цепи хромосомы и плазмиды было найдено количество квазиоперонов (Таблица 3). Также были построены гистограммы длин (в генах) квазиоперонов. Ознакомиться с ними можно, посмотрев на рисунки 2 и 3.

Таблица 3. Количество квазиоперонов в геноме

Sequence type	Strand	Amount of quasioperones
Chromosome	+	1190
Chromosome	-	1140
Plasmid	+	9
Plasmid	-	28
		Total: 2367



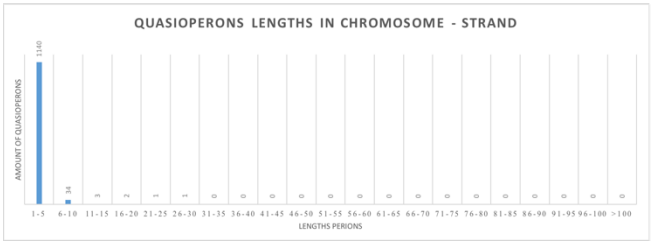
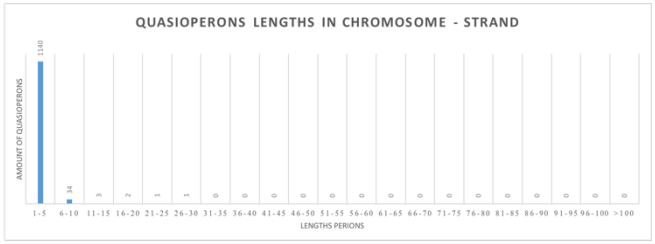


Рис. 2. Гистограмма длин квазиоперонов для бактериальной хромосомы



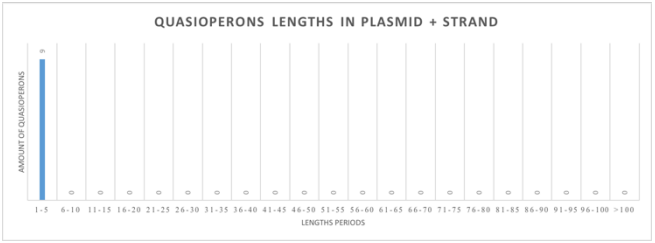


Рис. 3. Гистограмма длин квазиоперонов для плазмиды

3.4 Пересечения генов

Вся собранная информация о пересечении генов собрана в таблице 4. Интересно, что процент пересекающихся генов имеет значение около 14%.

Таблица 4. Данные о пересечении генов

Intersection	Amount
Total amount of intersections	671
Intersections on one strand	588
Intersections on different strands	83
frameshift = 0	407
frameshift = 1	243
frameshift = 2	21

3.5 Достоверность существования белков

В таблице 5 отражены существующие данные о белках довольно близкого штамма Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1736.

Таблица 5. Данные о существовании белков

Protein existance	Amount
Evidence at transcript level	13
Inferred from homology	1591
Predicted	2407

4 ОБСУЖДЕНИЕ

В этом разделе будут проанализированы некоторые полученные данные.

4.1 Гистограмма длин бактериальных белков

Интересным наблюдением является то, что существенная часть самых больших и, соответственно, сложно устроенных белков, отвечает за узнавание клеткой специфического субстрата и связь с ним. В том числе, два самых больших белка в клетке связаны с такими функциями. Такое явление, возможно, связано с тем, что исследуемый организм в силу своего способа существования (он вызывает сальмонеллез, кишечную инфекцию), должен надежно узнавать «искомую» поверхность и прикрепляться к поверхности ЖКТ.

4.2 Распределение генов по цепям ДНК

В изначальном плане проводимого анализа предполагалось проверить гипотезу о независимом распределении генов сразу для всего генома, однако в ходе работы было замечено, что существенный вклад в отклонение расположения генов от равномерного вносит бактериальная плазмида. В связи с этим было решено провести анализ распределения генов по + и – цепям ДНК плазмиды. Выяснилось, что для неё гипотеза о случайном распределении не выполняется. Изучение продуктов плазмиды показало, что она содержит довольно большое количество генов, принадлежащих бактериофагу, причем все его гены, кроме одного, расположены на - цепи. Так как обычно бактериофаг при встраивании в клетку переносит не только свой наследственный материал, но и часть последовательностей ДНК предыдущего хозяина (это явление называется трансдукцией), причины «дисбаланса» в распределении генов по цепям становятся очевидны.

4.3 Квазиопероны

Очевидным образом, самый большой интерес вызвали самые длинные квазиопероны. Подробной информации про самый большой из них, длиной в 28 генов, не нашлось, но, судя по наличию в нем белков, связанных с бактериофагами, можно сделать предположение о том, что такой квазиоперон – результат жизнедеятельности какой-то вирусной частицы. А вот второй по размеру квазиоперон длиной в 26 генов и располагающийся на - цепи оказался интереснее: он содержит гены, ответственные за секрецию патогенных белков. Изучение статей показало, что этот оперон называется «Type III secretion system» и в свою очередь принадлежит системе, ёмко названной «pathogenicity island», что дословно переводится как «остров патогенности»[4], то есть участок генома, ответственный за патогенность бактерии. Также примечательно, что довольно велико количество квазиоперонов небольшой длины, в которых гены располагаются парами или тройками. Скорее всего это связано с тем, что мало какой процесс в клетке возможно выполнить с использованием одного белка, а выполняющие комплементарные функции гены белков «выгодно» объединять в ДНК просто с точки зрения энергоэффективности и регуляции их синтеза в нужном соотношении.

4.4 Перекрывание генов

Согласно полученным результатам, в геноме перекрываются около 14% процентов генов, что не так уж мало. Это явление можно попытаться объяснить рядом причин:

1. Первая причина и, возможно, самая существенная – это отсутствие точно подтвержденных данных просто о существовании многих генов
2. Заражение клетки бактериофагом также может исказить данные, так как у вирусов всегда идет сильный отбор в сторону уменьшения размера генома, и они часто «прибегают» к использованию перекрывающихся генов.
3. С некоторой натяжкой можно рассмотреть это явление и как объяснение давления отбора уже на саму бактерию. Меньше геном, есть возможность быстрее поделиться, а значит вероятность удачно инфицировать хозяина повышается. Но данная гипотеза, например, не может ответить на вопрос, почему сперва просто не уменьшилось количество некодирующих последовательностей.
4. Некоторые перекрывания, особенно со смещенной рамкой считывания можно объяснить ошибками, возникающими при репликации и репарации ДНК.

5 ЗАКЛЮЧЕНИЕ

При подведении итогов очень полезной оказывается полученная статистика достоверности существования белков: на уровне транскриптов мы знаем только о 13 белках из целого генома. Существование большей части белков просто предсказано и не имеет даже прогнозов на основе гомологии другим белкам. Учитывая огромное социально-экономическое значение данного микроорганизма, становится понятно, что, несмотря на большое количество имеющегося материала, бактерия всё ещё остается недостаточно изученной. Данная работа сделала пусть и не существенный, но всё же вклад в знание об этой бактерии, представив в наглядном виде и систематизировав уже существующие данные.

6 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Все результаты и рассчеты, использованные в обзоре, содержатся в файле Galkin_pr13.xlsx, который можно загрузить по адресу
https://kodomo.fbb.msu.ru/~simon/term1/excelprc/Galkin_pr13.xlsx

БЛАГОДАРНОСТИ

Автор работы выражает искреннюю благодарность к.ф.-м.н., ведущему научному сотруднику НИИ ФХБ им. А.Н.Белозерского, Алексеевскому Андрею Владимировичу за оперативную помощь в разрешении затруднительных ситуаций, возникавших в процессе написания работы.

СПИСОК ИСТОЧНИКОВ

[1] https://kodomo.fbb.msu.ru/~simon/term1/pr6_genome.html
[2] https://www.ncbi.nlm.nih.gov/nuccore/1026303264?report=graph
[3] http://www.uniprot.org/uploadlists/
[4] https://www.ncbi.nlm.nih.gov/pubmed/9140973/