Обзор бактерии Thiomicrorhabdus aquaedulcis

Артём Севастьянов (https://kodomo.fbb.msu.ru/wiki/Users/sevartem)

Факультет биоинженерии и биоинформатики, МГУ им. М.В. Ломоносова, Москва, Россия

Аннотация — Исследование посвящено анаэробной серной бактерий Thiomicrorhabdus aquaedulcis. Будут рассматриваться особенности бактериального генома.

Ключевые слова: Thiomicrorhabdus aquaedulcis, серобактерия, kodomo

ВВЕДЕНИЕ

Таксономическое положение бактерии: …/84/96/63/Thiomicrorhabdus aquaedulcis HaS4^T; Bacteria, Pseudomonadota, Gammaproteobacteria. Рассматриваемый организм – аэробная сероокисляющая бактерия, изолированная из воды озера Харутори в Японии. Точный таксономический статус ещё не был определен. Является облигатным хемолитоавтотрофом, который растет при температуре от 0 до 25 °C (оптимум — 22 °C) и pH от 6.2 до 8.8 (оптимум — pH 6.6-7.4). Бактерии имеютт форму палочек, длиной 1.6–2.5 мкм, шириной 0.7–0.9 мкм и негативны по Граму. Филогенетический анализ, основанный на гене 16S рРНК, показал, что штамм связан с родом Thiomicrorhabdus, но филогенетически отличается от типовых штаммов существующих видов в этом роду. На основе его филогенетических и фенотипических свойств штамм HaS4T (=NBRC 112315T=BCRC 81110T) предлагается как типовой штамм нового неморского вида рода Thiomicrorhabdus с именем Thiomicrorhabdus aquaedulcis sp. nov. [1]

В исследовании будут рассмотрены количество встречаемых в геноме белков в зависимости от того, какому диапазону принадлежат их длины, количество пересечений кодирующих последовательностей (CDS) на плюс-цепи кольцевой молекулы ДНК и анализ встречаемости кодирующих последовательностей белков Thiomicrorhabdus aquaedulcis у других бактерий таксона Bacteria, Pseudomonadota, Gammaproteobacteria (далее – Сравниваемые бактерии).

МЕТОДЫ

Встречаемость белков в зависимости от их длин. В качестве источника информации используется таблица особенностей генома Thiomicrorhabdus aquaedulcis, размещенная на сайте Национального института здоровья. Анализ проведен с использованием возможностей сервиса Google sheets.

Распределение длин пересечений кодирующих последовательностей. Использованы, методы, идентичные методам, использованным в предыдущей задаче.

Встречаемость белков определенных назначений у бактерий одного и того же таксона. Рассмотрены таблицы особенности генома других бактерий, имеющих таксон Bacteria, Pseudomonadota, Gammaproteobacteria. Работа с таблицами производилась с помощью Microsoft Excel. Данные в таблицах были отфильтрованы и систематизированы с помощью программ, написанных на языке программирования Python. Для упрощения работы с таблицами использовались библиотеки csv (для записи содержимого csv-файлов в двумерные списки или списки из словарей вида {название столбца: значение i-й строки рассматриваемого столбца}), fnmatch (для работы с масками файлов и папок) и модуль listdir библиотеки os (для просмотра содержимого папок).

Для выполнения задачи было написано 3 программы:

1) filter.py – создает копию таблицы, в которой присутствуют только CDS, расположенные на кольцевой молекуле ДНК и кодирующие белок. Копия таблицы содержит только столбцы ID белка и его названия. Предполагается, что полученная таблица может использоваться для других задач, а текущая задача может быть изменена или расширена, поэтому был добавлен столбец ID белка, позволяющий получить кодирующую последовательность из fna-файлов бактериального генома.
2) name_comparator.py – сравнивает таблицы, созданные скриптом filter.py для Thiomicrorhabdus aquaedulcis и Сравниваемой бактерии, выводит результат в файл result.txt, содержащий отношения количеств названий белков, имеющихся у обеих бактерий, к общему количеству белков у основной и рассматриваемой бактерий. Общим считается название белка, полностью совпадающее у Thiomicrorhabdus aqyaedulcins и Сравниваемой бактерии.
3) common_hist_script.py – собирает информацию о встречаемости белков с определенными названиями у Сравниваемых бактерий и преобразует информацию в таблицу, используемую для получения конечных результатов.

РЕЗУЛЬТАТЫ

Гистограмма длин белков. Рассмотрены длины 2395 возможных продуктов трансляции CDS. На гистограмме (Рис. 1) изображено количество встречаемых белков в зависимости от того, в какой диапазон длины они входят.

Максимум количества белков приходится на диапазон 90-140 (295 единиц), помимо которого существуют локальные максимумы на диапазонах 290-340 (233 единицы), 590-640 (45 единиц) и 690-740 (41 единица). Количество возможных продуктов трансляции, превосходящих по длине 1040 (23 единицы), существенно меньше потенциальных белков, не превосходящих по длине 1040 (2383).

Наименьшая длина возможного полипептида – 21, белки такой длины встречаются 4 раза. Наибольшая длина – 3176 аминокислот, белок такой длины встречается единожды. Второй по величине полипептид имеет длину 1496 аминокислот, то есть его длина отличается от максимальной более чем в 2 раза. Белок максимальной длины не изображен на гистограмме, поскольку его размещение привело бы к затруднению анализа рисунка.

Рис. 1. Количество белков, попадающих в определенные диапазоны длин

Определение длин пересечения кодирующих последовательностей. Рассмотрено 1255 кодирующих участков, расположенных на кольцевой бактериальной молекуле ДНК на плюс-цепи, среди них 122 последовательности пересекаются со следующей, что составляет около 9.72% от их общего количества. Большая доля пересекающихся последовательностей дает основание предположить, что в ходе эволюции бактерии уменьшение размера генома благоприятно влияло на распространение бактерии, из-за чего Thiomicrorhabdus aquaedulcis приобрела такую особенность.

На гистограмме (Рис. 2) изображено количество пересекающихся кодирующих последовательностей в зависимости от того, в какой диапазон попадают их длины пересечения. Наиболее распространенными длинами пересечений оказались значения от 1 до 10 нуклеотидов. За исключением локальных максимумов в диапазонах 26-30 и 51-55, зависимость можно считать убывающей. Интересно, что в геноме имеется одно пересечение соседних последовательностей длиной 91 нуклеотид. Его название – VWA domain-containing protein, однако связь с фактором фон Виллебранда мне пока что не понятна.

Рис. 2. Гистограмма количества пересекающихся кодирующих последовательностей в зависимости от длины пересечения

Отдельно рассмотрены длины пересечений в диапазоне от 1 до 15 нуклеотидов (Рис. 3). Чаще всего встречаются пересечения длиной 3 нуклеотида (57 единиц) и 7 нуклеотидов (16 единиц). Почти половина (7 единиц) длин пересечений не встречается ни разу, то есть пересечения “сгруппированы” по длинам. Такое распределение длин пересечений может говорить о том, что, с одной стороны, чем меньше длина пересечения, тем чаще она встречается, однако, с другой стороны, пересечения именно определенных длин могут иметь значение для бактерии.

Рис. 3. Гистограмма количества пересекающихся кодирующих последовательностей в зависимости от длины пересечения, выбраны длины пересечений от 1 до 15 нуклеотидов

Схожие кодирующие последовательности у других видов таксона Bacteria, Pseudomonadota, Gammaproteobacteria. Помимо таблицы особенностей генома исследуемой бактерии, рассмотрены 16 таблиц особенностей генома бактерий, предложенных другим студентам моего курса. Выбранные бактерии принадлежат тому же таксону, что и *Thiomicrorhabdus aquaedulcis*, то есть *Bacteria, Pseudomonadota, Gammaproteobacteria*. Проведен анализ встречаемости названий белков *Thiomicrorhabdus aquaedulcis* у Сравниваемых бактерий, а также анализ встречаемости названий белков Сравниваемых бактерий у *Thiomicrorhabdus aquaedulcis* (Рис. 4).

Наибольшая доля совпадений с названиями белков *Thiomicrorhabdus aquaedulcis* наблюдается у *Pseudomonas frederiksbergensis* (48,91%), что дает основание предполагать их возможное родство и/или схожие условия обитания. Наименьшая доля – у *Glaesserella parasuis* (32,79%), что составляет почти треть от общего набора названий белков исследуемой бактерии и также дает почву для предположений того, что бактерии принадлежат относительно близким таксонам.

Отдельно рассмотрена встречаемость названий белков Сравниваемых бактерий у исследуемой бактерии. Наибольшая доля наблюдается у *Moraxella catarrhalis* (42,29%), наименьшая – у *Klebsiella pasteurii* (16,28%). И среднее арифметическое, (24,92%) и медианное (23,14%) значения оказались ниже, чем соответствующие значения при измерении совпадений названий белков Сравниваемых бактерий с названиями белков *Thiomicrorhabdus aquaedulcis* (40,28% и 38,37% соответственно). Такой результат позволяет предположить, что геном рассматриваемой бактерии менее разнообразен, чем геном бактерий таксона *Bacteria, Pseudomonadota,*

Gammaproteobacteria. Предположение о менее разнообразном геноме *Thiomicrorhabdus aquaedulcis* также подкрепляется результатом определения длин пересечений кодирующих последовательностей, доля которых оказалась аномально большой.

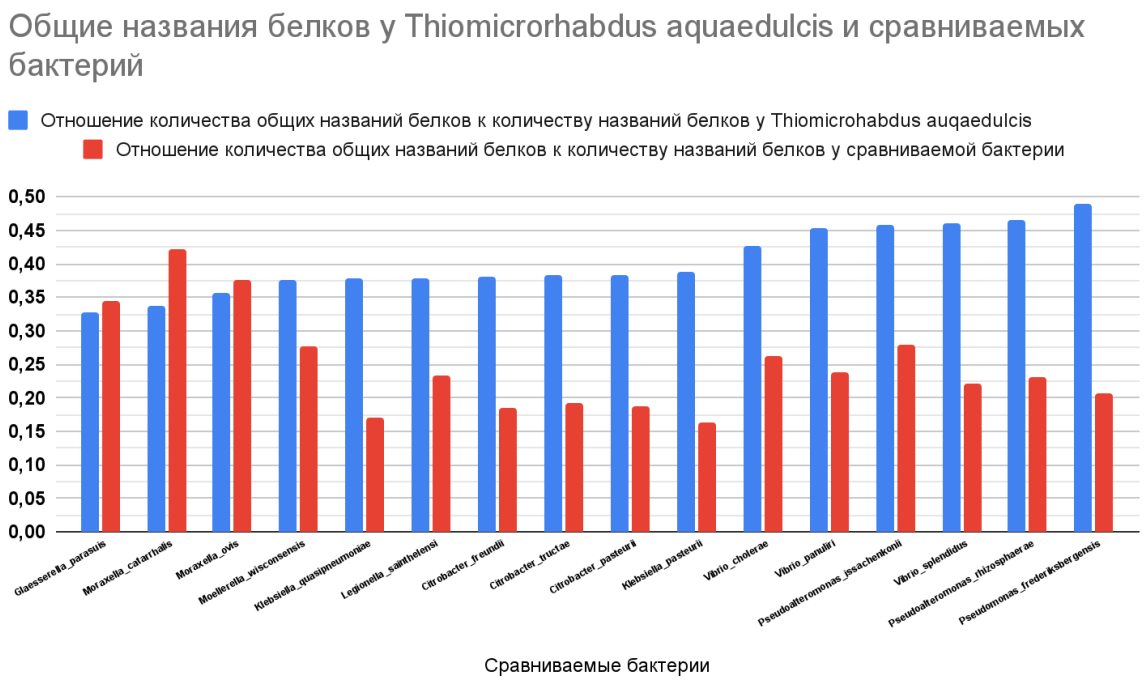

Рис. 4. Гистограмма общих названий белков Thiomicrohabdus aquaedulcis и сравниваемых бактерий

БЛАГОДАРНОСТИ

1) Преподавательскому составу, ведущему информатику на факультете биоинженерии и биоинформатики, за предоставление возможности освоения необходимых навыков для реализации мини-обзора и проведение уникального для МГУ курса, ориентированного на реальные и практические, а не абстрактные задачи. Серьёзно.
2) Командной строке Bash, работа с которой позволяет понять, что существуют более простые, быстрые и удобные методы работы с файлами: (например: Python)
3) Языку программирования Python, без него работа заняла бы в сотни раз большее количество времени
4) Google Sheets за возможность (пока что бесплатно) пользоваться облачными таблицами и совершение невозможного – перевода формул на английский язык. Больше никаких нечитаемых СРЗНАЧЕСЛИМН(БДДИСПП(ВПР(ЕНЕТЕКСТ(МЕДИАНА(...

ИСТОЧНИКИ

1. Hisaya Kojima, Manabu Fukui. Thiomicrorhabdus aquaedulcis sp. nov., a sulfur-oxidizing bacterium isolated from lake water. Int J Syst Evol Microbiol

2019;69:2849–2853
https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003567
2. Xuewen Liu, Baoping Chen, Qiliang Lai, Zongze Shao and Lijing Jiang. Thiomicrorhabdus sediminis sp. nov. and Thiomicrorhabdus xiamenensis sp. nov., novel sulfur-oxidizing bacteria isolated from coastal sediments and an emended description of the genus Thiomicrorhabdus. Int. J. Syst. Evol. Microbiol. 2021;71:004660.
https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.004660

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

1. Таблица особенностей генома *Thiomicrorhabdus aquaedulcis* с листами, соотвествующими выполненным задачам:
https://docs.google.com/spreadsheets/d/17-x7QThsKnZky-FRZ8gxQc6cXt5jrMF33Yc0USyMKgA/edit?usp=drive_link
2. Архив с программами и файлами, использованными для изучения схожих кодирующие последовательности у видов таксона Bacteria, Pseudomonadota, Gammaproteobacteria, помимо *Thiomicrorhabdus aquaedulcis*:
https://drive.google.com/file/d/1aG7EMGI1PSDT7dS1pWYXqQOfdx4QqYH0/view?usp=drive_link

ИСПОЛЬЗУЕМЫЕ РЕСУРСЫ

1. Сведения о *Thiomicrorhabdus aquaedulcis* на сайте Национального института здоровья (NIH):
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/004/001/325/GCF_004001325.1_ASM400132v1