Биоинформатика, 14 практикум. Дата публикации - 7 декабря 2018 года.

Обзор генома бактерии *Kitasatospora setae* KM-6054

Грызунов Никита Сергеевич

студент Факультета биоинженерии и биоинформатики, Московский Государственный университет имени М. В. Ломоносова, г. Москва, Российская Федерация

АННОТАЦИЯ

Главной целью данной работы является изучение генома бактерии *Kitasatospora setae* KM-6054 на основе особенностей этой бактерии. В данной работе исследуется распределение кодирующих и некодирующих генов бактерии по выполняемым функциям, их перекрытие, закономерности распределения генов по длине. Также, при изучении литературы, была проверена поставленная гипотеза о зависимости расположения генов от полярности цепей ДНК, на которой они расположены. Помимо этого была выявлена разница в количестве тРНК аминокислот данного типа между данным организмом и остальными бактериями. Одним из результатов работы также является приблизительная оценка количества оперонов на хромосоме бактерии. Дополнительно были составлены общие картины о характере расположения генов на хромосоме и о продуктах этих генов.

ВВЕДЕНИЕ

Рисунок 1. Таксономия *Kitasatospora setae*

>Bacteria
>Terrabacteria group
>Actinobacteria
>Streptomycetales
>Streptomycetaceae
>Kitasatospora
>Kitasatospora setae

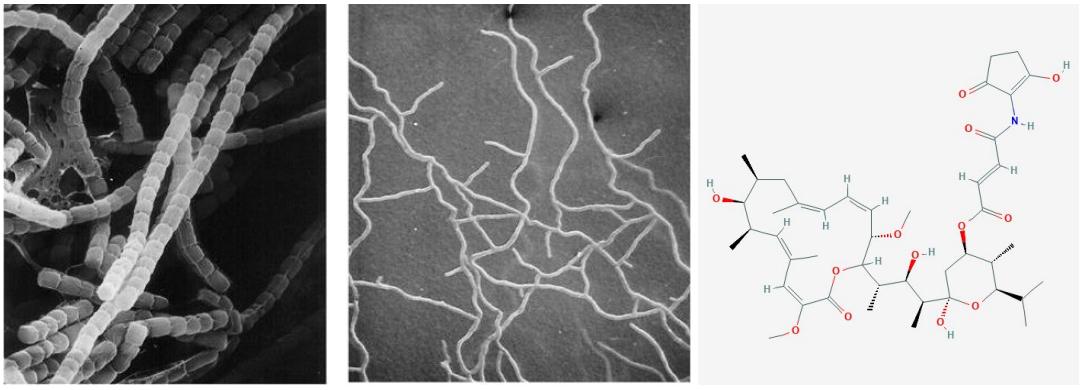
Kitasatospora setae - аэробная непатогенная мезофильная бактерия, обитающая у поверхности почвы и образующая мицелий актинобактерия (1). Колония *K. setae* образует дифференцированный мицелий, состоящий из двух частей Как у всех актинобактерий, большая часть нуклеотидов ДНК приходится на гуанин и цитозин (в сумме 73,1% от общего количества нуклеотидов). Бактерия содержит одну линейную хромосому из 8 783 278 пар нуклеотидов с терминальными инвертированными повторами. Плазмид не содержит (2).

Многие актинобактерии, в особенности Streptomycetaceae, отличаются способностью производить различные вторичные метаболиты, полезные для человека. Около двух третей известных науке антибиотиков встречаются в качестве продуктов у представителей данной группы бактерий. Известно, что *K. setae* вырабатывает антибиотик сетамицин (setamycin, bafilomycin B1), подавляющий развитие трихомонад и оказывающий слабое воздействие на некоторых грамположительных бактерий (3).

Бактерии рода Kitasatospora довольно схожи с распространёнными Streptomyces из-за их близкого родства. Когда открыли первые штаммы Kitasatospora, их сразу же начали относить к Streptomyces, но различия в строении 16S рРНК, клеточной стенки вегетативного мицелия (рисунок 2), а также малая степень синтении не позволили причислить их к Streptomyces (4). Вообще, геном *K. setae* - первый секвенированный геном из Streptomycetaceae, отличный от Streptomyces.

Цель данной работы - изучить геном *K. setae*, выявить самые интересные моменты, связанные с данным геномом, и исследовать их. Немаловажным является и получение опыта в написании подобного рода научных статей. Ведь правильная постановка важной проблемы и ее наиболее простое, логичное и в то же время интересное решение и есть искусство и красота науки.

Рисунок 2. Слева - воздушная цепочка спор (часть воздушного мицелия), **посередине** - вегетативный мицелий *Kitasatospora setae* KM-6054 (сканирующая электронная микроскопия), **справа -** антибиотик сетамицин (setamycin, bafilomycin B1) .



можно найти во вкладке “Вставка”.

МАТЕРИАЛЫ И МЕТОДЫ

Ввод данных генома бактерии
В процессе исследования для обработки и анализа данных о геноме организма было использовано следующее программное обеспечение: Google Таблицы (электронная таблица), Far Manager 3 (файловый менеджер). В качестве источника данных для анализа генома бактерии была использована база данных NCBI RefSeq, откуда был взят файл GCF_000269985.1_ASM26998v1_feature_table.txt, содержавший информацию о кодирующих и некодирующих последовательностях ДНК и РНК *K. setae*. Геном бактерии был загружен в Google Таблицы и представлял из себя плоскую таблицу с информацией о последовательностях ДНК.

Анализ последовательностей ДНК
На основе плоской таблицы генов бактерии, для подсчета количества генов белков, псевдогенов, генов рРНК, тРНК и яРНК была использована сводная таблица. Распределение белков по длинам и распределение пересечений последовательностей было реализовано с помощью логической функции СЧЁТЕСЛИМН(). Максимальная и минимальная длины продуктов, среднее значение, стандартное отклонение от среднего и медиана длин кодирующих последовательностей были подсчитаны с помощью статистических функций МАКС(), МИН(), СРЗНАЧ(), СТАНДОТКЛОН() и МЕДИАНА() соответственно.

Визуальное представление данных
Google Таблицы предоставляют возможность создания наглядных средств представления и обработки информации. В данной работе были использованы гистограммы, столбчатые диаграммы и таблицы. Их

Проверка гипотезы о зависимости количества генов от полярности цепочки-носителя проводилась посредством биномиального распределения, наиболее распространенного вида дискретного распределения. Гипотеза отвергалась тогда, когда интегральное биномиальное распределение наблюдаемой величины возвращало значение большее, чем 0,95 (95%). Функция биномиального распределения - БИНОМРАСП(). Также дополнительная проверка проводилась по Критерию согласия Пирсона (критерий хи-квадрат), функция CHITEST(). В качестве аргументов эта функция принимает два аргумента - наблюдаемый и ожидаемый результаты. Функция возвращала значение хи-квадрат для заданных аргументов, и, сверяя это значение с заранее изготовленными таблицами вероятности результата эксперимента быть случайным отклонением при заданных значении хи-квадрат и количестве степеней свободы, выносился вердикт предложенным гипотезам.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Число белок-кодирующих генов по категориям
Целью данной части работы была классификация продуктов белок-кодирующих генов на четыре категории, результат которой представлен в таблице 1.

Таблица 1. Количество белок-кодирующих генов

Тип белков	Количество белков	%
Рибосомальные	64	0,9
Транспортные	331	4,5
Гипотетические	2163	29,6
Другие	4742	65
Всего	7300	100

Из таблицы можно заключить, что рибосомальных и транспортных белков довольно много по сравнению с остальными белками. Это довольно логично, ведь эти две категории белков очень востребованы и относятся к продуктам генов “домашнего хозяйства”. В дополнение можно сказать, что геном *K. setae* нуждается в дальнейшем исследовании, так как количество гипотетических белков, чьи функции еще не установлены, оценивается в 30%.

Число всех генов по категориям

При анализе генома бактерии было проведено распределение большинства генов на четыре следующих категории: белок-кодирующие гены, псевдогены, гены рРНК и тРНК (таблица 2).

Таблица 2. Количество генов по классам

Группы генов	Количество генов в группе
Гены белков	**7300**
Псевдогены	**302**
Гены тРНК	**72**
Гены рРНК	**27**

Анализируя результаты, можно заметить, что отношение количества псевдогенов к количеству генов белков гораздо меньше, чем у эукариот. Это дополнительно свидетельствует о том, что геном K. setae, как и у большинства бактерий, компактный.

Также были выявлены некоторые другие последовательности, не вписывающиеся в данные категории. Среди них - ген SRP RNA (1 шт.), который входит в состав SRP (частица узнавания сигнала) - рибонуклеиновой частицы, локализованной на мембране. Она обеспечивает прохождение через мембрану секретируемых полипептидов, содержащих сигнальную последовательность (5, 6). Также был выявлен ген tmRNA (транспортно-матричная РНК, участвующая в механизме транс-трансляции (7)), 1 шт. И, в конце концов, обнаружена RNase P (эндорибонуклеаза P, рибозим, функция - расщепление РНК (8)).

Число продуктов белок-кодирующих генов по длине

С целью выявления общей картины о белках бактерии было проведено распределение протеинов по их длине. Результатом является гистограмма 2. Также были выявлены следующие моменты. Белок с минимальной длиной среди всех продуктов трансляции - АДФ-рибоза дифосфатаза (ADP-ribose diphosphatase), длина этого белка составляет 27 аминокислотных остатков. Белок с максимальной длиной - KR domain-containing protein, состоящий из 6512 аминокислотных остатков.

Анализируя данные, можно заключить, что пик длин пришелся на значения между 200 и 300 аминокислотными остатками. Однако средняя длина всех протеинов бактерии - 336 аминокислотных остатков, стандартное отклонение от этой величины - 278 аминокислотных остатков. Можно сказать, что наблюдается умеренный разброс длин протеинов. Дополнительно была посчитана медиана длин белков - 285 аминокислотных остатков.

Анализ промежутков между генами на хромосоме

В данном разделе рассматривается статистика межгенных промежутков. Как видно из гистограммы 3, пик длин промежутков пришелся на значения от 0 до 100 нуклеотидов. Уже на данном этапе можно дать верхнюю оценку количества квазиоперонов - 1697, о количестве которых будем говорить в следующем подразделе. Пристальное рассмотрение этого участка длин выявило, что количество белков везде на этом интервале примерно одинаковое (гистограмма 1).

Гистограмма 1. Распределение белков по их длинам (только участок длин 0-100 нуклеотидов). На горизонтальной оси расположены группы протеинов с длинами из заданных промежутков длин. Вертикальная ось демонстрирует количество протеинов в группах. Жёлтые числа у оснований столбцов - точное количество протеинов для данной группы протеинов.

Также, исходя из таблицы 3 можно сделать вывод, что гены на хромосоме расположены довольно тесно.

Таблица 3. Статистические величины для множества межгенных расстояний.

Статистическая величина	Значение
Среднее арифметическое	**1268**
Медиана	**190**
Стандартное отклонение	**2789**
Минимальное значение	**-3 133**
Максимальное значение	**63 569**

Плотность расположения генов на хромосоме

Для того, чтобы оценить компактность генома бактерии,

было произведено вычисление среднего количества генов на миллион пар нуклеотидов ДНК. Оно составило 889 генов/Mb.p. И это довольно странно для бактерии, ведь плотность генов, характерная для бактерий, колеблется от 800 до 1000 генов/Mb.p. Но всё же, этот геном остается более компактным, чем у большинства эукариот и сопоставим с геномом *Saccharomyces cerevisiae* (9).

Квазиопероны в геноме и их количество

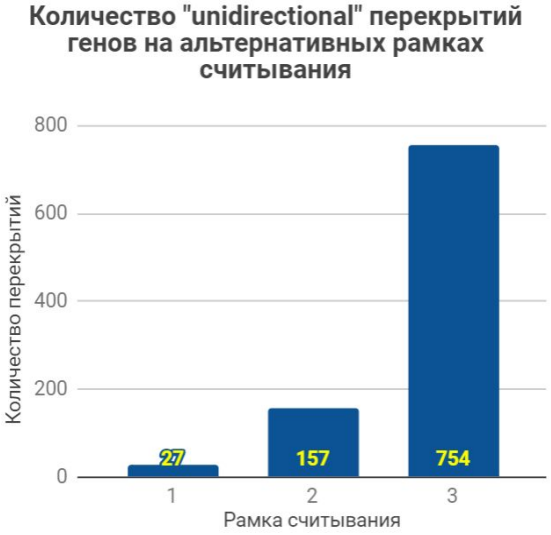
Квазиопероны - гипотетические опероны, которые выделяются только по межгенному растоянию и не формируются на основе реальных наблюдений и экспериментов. Тем не менее, с помощью них можно дать приблизительную оценку количества оперонов на хромосоме бактерии. В вычислениях за квазиопероны принимались те последовательности генов, соседние элементы последовательности которых отстояли не более, чем на определенную величину. Эта величина варьировалась от 10 до 100 нуклеотидов в максимально возможном промежутке. Также количество квазиоперонов было посчитано отдельно на двух цепях. Здесь присутствует интересная закономерность: количество квазиоперонов на прямой цепи немного больше, чем на комлементарной, кроме случая, когда максимально возможное расстояние между генами в квазиопероне принималось за 10 нуклеотидов (столбчатая диаграмма 1).

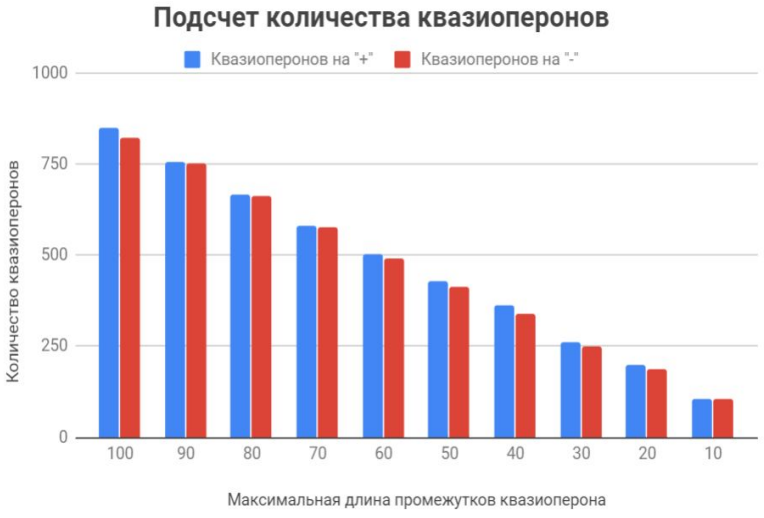
Гистограммы 2 и 3. 2 - распределение белков по их длинам. На горизонтальной оси расположены группы протеинов с длинами из заданных промежутков длин. Вертикальная ось демонстрирует количество протеинов в группах. Жёлтые числа у оснований столбцов - точное количество протеинов для данной группы протеинов. **3** - распределение промежутков между генами на хромосоме. По горизонтальной оси отложены группы по длинам перекрытий в нуклеотидах. По вертикальной - количество промежутков между генами, входящих в данную группу по длине. Жёлтые числа у оснований столбцов - мощность группы протеинов данной длины.

Количество пересечений генов на хромосоме

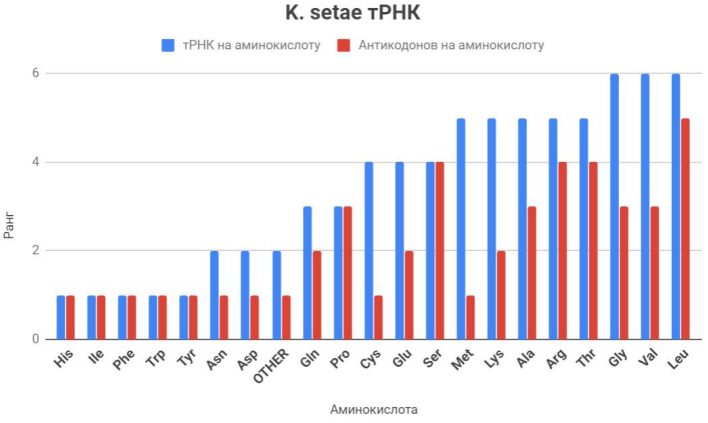
В процессе работы было подсчитано количество и величина перекрытий соседних генов на хромосоме *K. setae*. Данная величина была подсчитана для всего генома в целом и для прямой и обратной цепи отдельно, но подсчет по цепям не выявил каких-либо интересных моментов и количество перекрытий не зависело от цепи. Зато в сумме с двух цепей было произведено очень интересное наблюдение, зафиксированное на гистограмме 3. Основное количество перекрываний было таково, что длина перекрытий была кратна трем. Существует классификация перекрытий генов (10), и самый распространенный тип перекрытий для бактерий по этой классификации - “unidirectional overlap” - перекрытие, длина которого кратна трем и считывание перекрывающихся генов при котором происходит в одном направлении. Объясняется такая частота данного типа перекрытий тем, что для бактерий как для быстро эволюционирующих организмов свойственно иметь много точечных мутаций. Такие мутации могут происходить в стоп-кодонах, что часто и происходит на практике. Эти теоретические выкладки подтверждает наблюдаемая статистика перекрытий генов *K. setae* (гистограмма 3). Такие перекрытия могут быть полезны для бактерии, так как они обеспечивают совместную регуляцию транскрипции и трансляции (11).

Гистограмма 3 и столбчатая диаграмма 1. *Гистограмма 3*. Количество перекрытий генов по рамкам считывания. На горизонтальной оси отложены рамки считывания (здесь они определены как остаток деления длины пересечения на 3), на вертикальной оси - количество пересечений генов для данной рамки считывания. Желтые ярлыки - мощность множества перекрытий с данной рамкой считывания. *Столбчатая диаграмма 1*. Зависимость количества квазиоперонов от максимальной длины межгенных промежутков в квазиопероне. На горизонтальной оси - максимальная длина межгенных промежутков в квазиопероне, по вертикальной оси - количество квазиоперонов с заданной максимальной величиной промежутков. Синим - на прямой цепи ДНК, красным - на обратной.





Гистограмма 4 и столбчатая диаграмма 2. *Гистограмма 4*. Процентное содержание аминокислот в белках бактерий (усредненно). *Столбчатая диаграмма 2*. Выстраивание аминокислот в порядке возрастания количества тРНК и антикодонов для данной аминокислоты. Синим - количество тРНК для данной аминокислоты, красным - количество антикодонов для данной аминокислоты.



Метаболизм бактерии и тРНК

Этот подраздел работы посвящен выявлению влияния метаболизма бактерии на отношения тРНК на аминокислоту и антикодон на аминокислоту. Проанализировав тРНК *K. setae* и сравних их с тРНК других бактерий (12), можно выделить несколько отличий. Во-первых, цистеин и метионин - полярные незаряженные алифатические серосодержащие аминокислоты, их роль в *K. setae,* предполагается, выше, чем у среднестатистической бактерии. Влияние же аспарагиновой кислоты (отрицательно заряженная, алифатическая), напротив, полагается ниже. Объяснение этим наблюдениям пока не было придумано.

Распределение генов по цепям. Проверка гипотезы

Применив метод проверки гипотез, использующий функцию БИНОМРАСП(), выяснили, что в маленьких выборках из таблицы 2 распределение по цепям не было случайным. Это же подтвердилось и для белков рибосом. В случае с транспортными белками, псевдогенами, генами тРНК, рРНК и генами белков в целом опровергнуть гипотезу о неслучайности распределения не удалось.

ЗАКЛЮЧЕНИЕ

В заключение хочется сказать, что хоть мы и провели исследование генома, многие вещи так и остались во мраке неведения. Поэтому научный поиск никогда не прекратится, а если он не прекратится - то на планете Земля всегда будет интересно. Поэтому ищите и не сдавайтесь!

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Файл с выполненной работой в формате .xlsx можно скачать на учебном сайте автора: https://kodomo.fbb.msu.ru/~nikit00000s/term1/pr13/pr13.html

БЛАГОДАРНОСТИ

Автору хочется от всего сердца поблагодарить всех преподавателей биоинформатики 1 курса ФББ МГУ за проделанный труд и бесконечное терпение.

СПИСОК ЛИТЕРАТУРЫ

1. Genome Passport of Kitasatospora setae KM-6054. StrainInfo bioportal. http://www.straininfo.net/genomes/19951

2. ASM26998v1. Complete Genome of Kitasatospora setae. Otoguro, M. is at NITE Biological Resource Center (NBRC). Ikeda, H., Miura, H. and Takahashi Y. are at Kitasato Institute for Life Science. Ishikawa, J. is at National Institute of Infectious Diseases. Horinouchi, S., Ohnishi, Y. and Kuzuyama, T. are at The University of Tokyo. Hayakawa, M. is at University of Yamanashi. Nihira, T. and Kitani, S. are at Osaka University. Arisawa A. is at Mercian Corp. Nomoto F. is at Nagase & Co. Ltd. The other authors are at NITE Genome Analysis Center (NGAC). https://www.ncbi.nlm.nih.gov/assembly/GCF_000269985.1/

3. Genome Sequence of Kitasatospora setae NBRC 14216T: An Evolutionary Snapshot of the Family Streptomycetaceae. DNA Res. 2010 Dec;17(6):393-406. doi: 10.1093/dnares/dsq026. Epub 2010 Nov 8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2993542/

4. Genus Kitasatospora, taxonomic features and diversity of secondary metabolites. Yōko Takahashi
The Journal of Antibiotics volume 70, pages 506–513 (2017). https://www.nature.com/articles/ja20178

5. Signal recognition particle. Wikipedia
https://en.wikipedia.org/wiki/Signal_recognition_particle_RNA

6. Англо-русский толковый словарь генетических терминов 1995 407с. Арефьев В.А., Лисовенко Л.А.

7. Biology of trans-translation. Keiler KC. Annu Rev Microbiol. 2008;62:133-51. doi: 10.1146/annurev.micro.62.081307.162948. https://www.ncbi.nlm.nih.gov/pubmed/18557701

8. Ribonuclease P. Sidney Altman. Philos Trans R Soc Lond B Biol Sci. 2011 Oct 27; 366(1580): 2936–2941.
doi: 10.1098/rstb.2011.0142.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158923/

9. The size of the genome and the complexity of living beings. Amparo Latorre, Francisco J. Silva. 28/02/2013. MÈTODE.
https://metode.org/issues/monographs/the-size-of-the-genome-and-the-complexity-of-living-beings.html

10. Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae Yoko Fukuda, Takanori Washio and Masaru Tomita. Nucleic Acids Research, 1999, Vol. 27, No. 8. https://academic.oup.com/nar/article/27/8/1847/2848231

11. Properties of overlapping genes are conserved across microbial genomes. Zackary I. Johnson and Sallie W. Chisholm. Genome Res. 2004 Nov; 14(11): 2268–2272. doi: 10.1101/gr.2433104. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC525685/

12. Proteome-pI: proteome isoelectric point database. Lukasz P. Kozlowski. Nucleic Acids Res. 2017 Jan 4; 45(Database issue): D1112–D1116. Published online 2016 Oct 26. doi: 10.1093/nar/gkw978. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210655/