Мини-обзор

Анализ геномных и протеомных данных батерии *Clostridium saccharobutylicum DSM 13864*, имеющей практическое значение для производства ряда растворителей

Малышев Андрей Дмитриевич[1]

[1]Факультет биоинженерии и биоинформатики, Московский Государственный Университет имени М.В.Ломоносова, Москва, Россия

АННОТАЦИЯ

Контакты: malyshev.andrey@kodomo.fbb.msu.ru

1 ВВЕДЕНИЕ

Для бактерии *Clostridium saccharobutylicum*, как и для многих других клостридий, характерно маслянокислое брожение, отличающееся высоким разнообразием продуктов, состав которых зависит от условий среды и стадии роста колонии бактерий (Maczulak, 2011). Существует несколько исследований, связанных со способностью этих бактерий перерабатывать крахмал и другие углеводы в ряд соединений, используемых человеком в качестве растворителей, например, так могут быть получены ацетон, бутанол и этанол (Liew et al., 2006).

Для этих анаэробных бактерий, способных образовывать споры, в 2013 году был расшифрован полный геном. Выяснилось, что весь генетический материал *C. saccharobutylicum* находится на единственной хромосоме размером 5,107,814 пар оснований (Poehlein et al., 2013). И среди всех белок-кодирующих последовательностей бактерии были идентифицированы гены ферментов, необходимых для производства растворителей. Например, оказалось, что гены альдегиддегидрогеназы, CoA-трансферазы и ацетоацетатдекарбоксилазы входят в состав того же оперона *sol*, что и у некоторых родственных видов (Poehlein et al., 2013). Хотя нельзя не отметить, что и до этого предпринимались попытки построить генетические карты для этого организма (Keis et al., 2001). Всё это указывает на высокую практическую значимость *C. saccharobutylicum* и открывает простор для будущих исследований.

В данном мини-обзоре рассматриваются и анализируются базовые особенности генома и протеома бактерии с помощью простых биоинформатических методов. Исследуются кодирующие белок последовательности и составляющие их триплеты нуклеотидов, а также особенности распределения последовательностей по кольцевой ДНК бактерии.

2 МАТЕРИАЛЫ И МЕТОДЫ

Последовательность генома бактерии и другая информация, например, файл с хромосомной таблице, скачивались со следующего сайта: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/473/995/GCF_000473995.1_ASM47399v1/

Для работы с последовательностью генома бактерии, главным образом, использовались скрипты, написанные на языке программирования Python, которые можно найти по ссылке, указанной в разделе “Сопроводительные материалы”. Для многих программ предполагается единственность последовательности ДНК в геноме бактерии, что установлено в результате работы первых скриптов. Нумерация скриптов соответствует нумерации пунктов в разделе “Результаты и обсуждение”. Для удобства использования файлы скриптов содержат в начале сопроводительные комментарии, кратко поясняющие, что делает программа.

Для анализа встречаемости длинных повторов в разделе 3 использовалась программа BLAST, куда вводились интересующие последовательности.

Для построения графиков и диаграмм в пунктах 5 и 7 использовались соответствующие инструменты Google Sheets. Исходные данные для построения диаграмм можно найти по ссылке в разделе “Сопроводительные материалы”.

В части 5 при расчете GC-skew каждая рассматриваемая последовательность (окно) начиналась с того нуклеотида, который получался при суммировании параметра “step”.

В части 7 использовался ряд функций описательной статистики: AVERAGE, MEDIAN, STDEV.P - это функции Google Sheets.

Для оценки случайности распределения генов по цепям хромосомы в разделе 8 использовалась функция BINOM.DIST всё тех же электронных таблиц. При вычислении вероятности производилось умножение результата на 2, поскольку изначально было неизвестно, на какой из цепей генов больше.

Классы белков в разделе 9 определялись по названиям, а классы молекул РНК - по типам последовательностей в хромосомных таблицах (1 и 2 столбцы).

3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

1. Исследование геномных последовательностей бактерии. С помощью первого скрипта можно узнать количество последовательностей в составе генома организма, основываясь на FASTA-файле. У *Clostridium saccharobutylicum* последовательность в геноме единственная, её название

"NC_022571.1". Длина данной кольцевой хромосомы составляет 5107814 пар нуклеотидов, а содержание G-C пар равно 28,7%. Пониженный G+C-состав используется как систематический признак, на основе которого выделяют группу Firmicutes (Briggs et al., 2012), к которой и относится род *Clostridium* (Ciccarelli et al., 2006).

2. Исследование нуклеотидного состава геномной ДНК. В состав единственной геномной последовательности *Clostridium saccharobutylicum* входят 4 нуклеотида, распределение которых на "+"-цепи отражено в таблице 1.

Таблица 1. Распределние нуклеотидов на "+"-цепи.

Нуклеотид	Число вхождений
A	1825229
C	728309
G	735712
T	1818564

Данные таблицы показывают, что для отдельной цепи хромосомы *Clostridium saccharobutylicum* выполняется второе правило Чаргаффа, то есть содержание гуанина приблизительно равно содержанию цитозина, а содержание аденина - содержанию тимина (Rudner et al, 1968).

3. Поиск повторов большой длины в геноме бактерии. Скрипт №3 ищет повторы указываемой длины и выводит те из них, которые встречаются чаще минимального числа раз. Интересными оказываются 2 типа повторов. Есть повторы длины 30 следующего вида: ATTTAAATACATCTCATGTTAATGTTAATC. В точности такая последовательность встречается в геноме 32 раза, причем все они умещаются в промежутке 2818220..2820245 на "+"-цепи. Важно отметить, что сразу после них идут последовательности, кодирующие белки Cas. Всё это указывает на то, что обнаруженные повторы разделяют спейсеры в последовательности CRISPR (Barrangou, R., & Marraffini, L. A, 2014). Это предположение подтверждается данными со страницы GenBank, на которой для этой области хромосомы определена область повторов, относящихся к семейству CRISPR.
Другой тип повторов имеет длину в 37 нуклеотидов и последовательность следующего вида: AAAATAATCATGGCATTTTGGACTTGTTATTTTCTTT. Такой повтор встречается в геноме 15 раз в разных участках хромосомы. В таблице 2 приведены координаты первого нуклеотида в повторе и цепь, на которой он находится.

Всего 2 повтора такого типа на "–"-цепи имеют перекрывания с кодирующими белок последовательностями, это шикимат-дегидрогеназа (725881..726726) на "+"-цепи и гипотетический белок (884942..885595) тоже на "+"-цепи. Поиск такой последовательности из нуклеотидов с помощью программы BLAST показывает, что она встречается у ряда представителей рода *Clostridium*: *Clostridium diolis, Clostridium beijerinckii, Clostridium saccharoperbutylacetonicum.*

Таблица 2. Расположение повторов второго типа в геноме бактерии.

Положение первого нуклеотида	Цепь
418741	+
759598	+
761978	+
810114	+
2020765	+
2936165	+
3520916	+
4046425	+
4603297	+
4990167	+
726697	–
885566	–
1316413	–
3580260	–
4768166	–

Вполне очевидно, что отношения реальной встречаемости повторов к ожидаемой встречаемости для этих повторов - большие числа, ведь у них относительно большая длина и встречаются они значительное число раз.

4. Частоты встречаемости старт- и стоп-кодонов у *Clostridium saccharobutylicum.* Частоты старт-кодонов приведены в таблице 3, а частоты стоп-кодонов - в таблице 4.

Таблица 3. Встречаемость старт-кодонов в кодирующих белок последовательностях бактерии.

Старт-кодоны	Число вхождений	% от всех старт-кодонов
ATG	3633	86,3%
GTG	195	4,6%
TTG	313	7,4%
ATT	33	0,8%
ATA	36	0,9%
ATC	2	0,05%

Таблица показывает, что у *Clostridium saccharobutylicum* наиболее часто встречающимся старт-кодоном является триплет ATG, но, помимо него, есть и другие варианты, в том числе достаточно распространённые.

Если рассматривать более редкие старт-кодоны: GTG, TTG, ATT, ATA, ATC – то оказывается, что все они встречаются в последовательностях белков, не являющихся псевдогенами. Кодоны GTG и TTG являются стандартными для бактерий, и с них могут начинаться несколько процентов всех кодирующих белок последовательностей (Kozak, M., 1983). Есть статьи,

указывающие на то, что такие варианты кодонов могут использоваться в жизненно важных генах, связанных с транскрипцией, трансляцией и репликацией, что может служить механизмом дополнительной регуляции их экспрессии, что может быть важно в случае, когда клетка находится в состоянии голодания (Gvozdják, A., & Samanta, M.P., 2020).

Таблица 4. Встречаемость стоп-кодонов в кодирующих белок последовательностях бактерии.

Стоп-кодоны	Число вхождений	% от всех стоп-кодонов
TAA	2942	69,8%
TAG	951	22,6%
TGA	319	7,6%

Таблица показывает, что у *Clostridium saccharobutylicum* наиболее часто встречающимся стоп-кодоном является триплет TAA, но, помимо него, есть и другие варианты, в том числе достаточно распространённые.

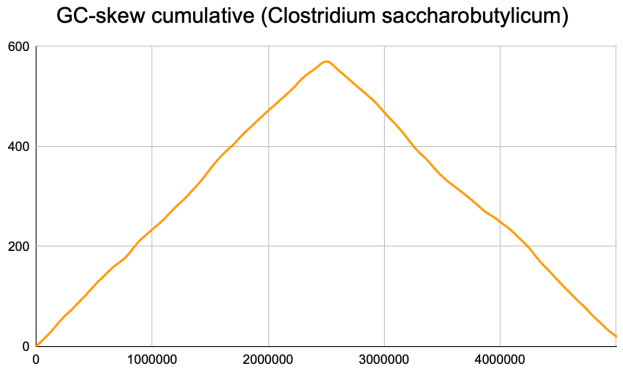
5. Поиск точки начала репликации на основе анализа данных GC-skew. С помощью скрипта №5 и электронных таблиц удалось построить график 1 зависимости GC-skew cumulative от точки на хромосоме для “+”-цепи. При этом размер окна составил 100000 нуклеотидов, а шаг - 100 нуклеотидов.

Формула для вычисления GC-skew в последовательности окна:

$$GC\_skew = \frac{n(G) - n(C)}{n(G) + n(C)}$$

n(G) - количество гуанина в последовательности окна, а n(C) - кол-во цитозина. Для построения графика использовалось интегральное значение GC-skew, то есть для получения значения в определенной точке суммировались все предыдущие значения GC-skew. Как показывает практика, у бактерий минимуму на графике соответствует точка начала репликации (oriC), а максимуму - диаметрально противоположная точка на хромосоме (ter), то есть то место, где после удвоения генетического материала встречаются две репликационные вилки и процесс заканчивается.

График 1. GC-skew cumulative вдоль хромосомы бактерии.



Максимум GC-skew cumulative (570) соответствует точке 2504000, а минимум - точке 0, ведь при картировании циклического генома за начало часто принимают точку начала репликации - oriC.

6. Распределение кодонов, кодирующих разные аминокислотные остатки в последовательностях генов белков. В таблице 5 приведены три наиболее часто встречающиеся аминокислоты и три наиболее редко встречающиеся.

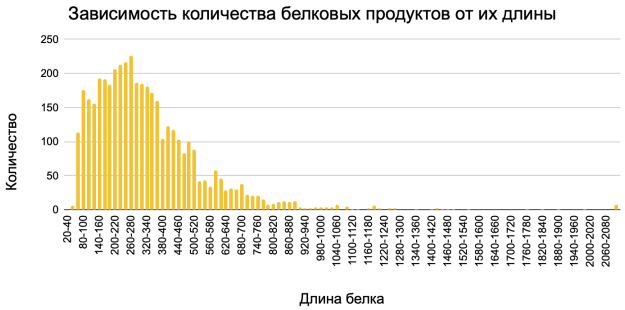
Таблица 5. Встречаемость разных аминокислотных остатков в кодирующих белки последовательностях генома бактерии.

Аминокислота	Число вхождений
Ile	130871
Lys	118525
Leu	115653
His	17402
Cys	16562
Trp	9317

Из данных таблицы следует, что часто в белках бактерии встречаются такие гидрофобные аминокислоты с длинными радикалами, как лейцин и изолейцин. Можно предположить, что они входят в состав α-спиралей мембранных белков, которые служат для заякоривания полипептида в липидном бислое. Лизин - одна из немногих аминокислот с положительно заряженным радикалом, к тому же, эта аминокислота часто входит в состав активных центров ферментов, являясь важным участником катализа. Редко встречаются аминокислотные остатки с гетероциклическими радикалами (триптофан и гистидин), что можно объяснить сложностью их биосинтеза, на который клетка тратит много энергии и углеродных скелетов. Цистеин содержит серу, элемент, который редко содержится в клетках в большом количестве.

7. Гистограмма длин белков *Clostridium saccharobutylicum*. С помощью электронных таблиц и функции COUNTIFS можно построить гистограмму длин белков, кодируемых геномом бактерии. Гистограмма представлена на графике 2.

График 2. Распределение длин белков *C. saccharobutylicum*.



Для построения графика все белки были разбиты по карманам ширины 20 на основе их длины, выраженной в количестве аминокислотных остатков.

Также с помощью электронных таблиц найдены некоторые статистические параметры для распределения длин белков. Средняя длина белка (функция AVERAGE) составила 310,5 аминокислотных остатка, а медиана (функция MEDIAN), то

есть наиболее часто встречающиеся значение - 266. При этом стандартное отклонение (STDEV.P) равно 222,12.

Самым коротким белком всего из 29 остатков оказалась транспозаза из семейства IS3, то есть фермент в составе мобильного генетического элемента, отвечающий за его вырезание и встраивание в разных участках генома. Самый же длинный белок - ndvB, который по данным UniProt является мембранным белком с 7 спиральными трансмембранными доменами, данный фермент способен связывать углеводы и обладает каталитической активностью.

8. Распределение генов РНК и белков по цепям ДНК. Результаты работы скрипта №8 представлены в таблице 6.

Таблица 6. Распределение генов по цепям кольцевой хромосомы бактерии и вероятность получить такое распределение для каждого класса последовательностей.

Класс последовательностей	На “+”-цепи	На “–”-цепи	Вероятность
CDS with protein	2085	2127	52,8%
CDS without protein	71	85	30%
rRNA	14	23	19%
tRNA	73	12	$8 * 10^{-10}$%

Таблица показывает, что у *Clostridium saccharobutylicum* распределение всех генов по цепям ДНК почти равномерное, но более 85% всех транспортных РНК кодируются на прямой цепи.

Использование функции BINOM.DIST электронных таблиц показывает, что вероятность получить такое или более неравномерное распределение составляет 53%, поэтому такую разницу в числе генов белков на разных цепях не стоит считать статистически значимой, чего нельзя сказать о распределении генов тРНК.

Дополнительно можно рассмотреть распределение кодирующих белки последовательностей по половинам хромосомы, считая от точки начала репликации.

9. Анализ встречаемости белков разных классов. Среди белков, закодированных в геноме *Clostridium saccharobutylicum* 693 относятся к “hypothetical protein”, что составляет почти одну шестую (16,45%) всех белков клетки. Также популярным классом являются ABC-транспортеры. Эти белки обладают специальным доменом, позволяющим им связывать молекулы АТФ и осуществлять транспорт веществ между компартментами клетки через мембраны за счет их гидролиза. Всего удалось обнаружить 128 таких белков.

Если рассматривать гены рРНК, то оказывается, что в геноме бактерии их закодировано 3 типа: 5S, 16S и 23S. Последние два типа встречаются по 12 раз, а 5S - 13 раз. Распределение расстояний между молекулами РНК разных типов показывает, что в большинстве случаев расстояние между соседними генами 16S и 23S рРНК равно 242 нуклеотидам (5 раз), а то же число для 5S и 23S рРНК составляет 63 нуклеотида (8 раз). Это указывает на то, что молекулы рРНК закодированы в геноме кластерами. Если считать, что расстояния между соседними рРНК в кластере по порядку величины не больше 1000 нуклеотидов, то в геноме можно насчитать 12 кластеров, в которых последовательно идут гены 16S, 23S и 5S рРНК. Кластеры обнаруживаются как на прямой, так и на обратной цепи, а порядок следования генов сохраняется. Таким образом, лишь один ген 5S рРНК не входит в состав кластера из трёх молекул. Кластеры рРНК соотвествуют опероннам, то есть участкам генома, которые вместе транскрибируются. Наличие оперонов, состоящих из рРНК и тРНК показано для бактерий достаточно давно (Apirion, D., & Miczak, A., 1993).

10. Поиск белков, связанных с образованием спор у *Clostridium saccharobutylicum.* Поиск белков имеющих слово “spore” в названии даёт 31 результат. Это указывает на способность бактерии *Clostridium saccharobutylicum* к спорообразованию.

В результате, в рамках мини-обзора рассмотрены некоторые особенности генома и протеома бактерии *Clostridium saccharobutylicum,* которые могут послужить основой для дальнейших исследований в данной области, в том числе, и имеющих практический характер.

4 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

По данной ссылке Вы сможете найти все скрипты, которые использовались для написания мини-обзора:

https://kodomo.fbb.msu.ru/~malyshev.andrey/term1/mini_review/

Все они по умолчанию работают с файлами, содержащими информацию о геноме, в той же папке, поэтому наиболее простой способ проверить их работу - скачать все файлы сразу. По той же ссылке можно найти и результаты выполнения скриптов, содержащие данные, на основе которых написан мини-обзор.

Остальная часть работы производилась в Google-таблицах. По первой ссылке можно найти данные, на основе которых строился график GC-skew cumulative в разделе 5, а по второй - анализ различных данных протеома:

https://docs.google.com/spreadsheets/d/1G21fXfpTm8B1Yi70FEFrcmhAlGzIXS-kqR3BQ79qKWw/edit?usp=sharing

https://docs.google.com/spreadsheets/d/11XmilRfQxWHMV6JvjjT2HPZfM1Do5dTPufmoQ805evI/edit?usp=sharing

5 БЛАГОДАРНОСТИ

Хотелось бы выразить благодарность моим коллегам, Чекалину Денису и Михаилу Никонову, с которыми у меня была возможность обсудить наиболее интересные концептуальные вопросы, возникавшие по мере написания мини-обзора.

6 ССЫЛКИ НА ИСТОЧНИКИ

Apirion, D., & Miczak, A. (1993). RNA processing in prokaryotic cells. BioEssays, 15(2), 113–120.

Barrangou, R., & Marraffini, L. A. (2014). CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Molecular cell, 54(2), 234–244.

Briggs GS, Smits WK, Soultanas P. Chromosomal replication initiation machinery of low-G+C-content Firmicutes. J Bacteriol. 2012 Oct;194(19):5162-70.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006 Mar 3;311(5765):1283-7.

Gvozdjăk, A., & Samanta, M.P. (2020). Genes Preferring Non-AUG Start Codons in Bacteria. arXiv: Genomics.

Keis S, Sullivan JT, Jones DT. Physical and genetic map of the Clostridium saccharobutylicum (formerly Clostridium acetobutylicum) NCP 262 chromosome. Microbiology (Reading). 2001 Jul;147(Pt 7):1909-1922.

Kozak M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev. 1983 Mar;47(1):1-45.

Lehninger, A.L.; Nelson, D.L.; Cox, M.M. (2005). Lehninger principles of biochemistry. New York: W.H. Freeman.

Liew, S.T. & Ariff, Arbakariya & Mohamad, Rosfarizan & Raha, A.R.. (2006). Production of Solvent (acetone-butanol-ethanol) in Continuous Fermentation by Clostridium saccharobutylicum DSM 13864 Using Gelatinised Sago Starch as a Carbon Source. Malays J Microbiol. 2. 42-50. 10.21161/mjm.220608.

Maczulak A (2011), "Clostridium", Encyclopedia of Microbiology, Facts on File, pp. 168–173, ISBN 978-0-8160-7364-1.

Poehlein A, Hartwich K, Krabben P, Ehrenreich A, Liebl W, Dürre P, Gottschalk G, Daniel R. Complete Genome Sequence of the Solvent Producer Clostridium saccharobutylicum NCP262 (DSM 13864). Genome Announc. 2013 Nov 27;1(6):e00997-13.

Rudner R, Karkas JD, Chargaff E. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci U S A. 1968 Jul;60(3):921-2.