Мини-обзор генома бактерии *Acetobacter pasteurianus subsp. Ascendens*

Бобровский Даниил

Факультет биоинженерии и биоинформатики МГУ им. М.В. Ломоносова
Ленинские горы МГУ 1 стр. 73, г. Москва, 119234, Российская федерация
daniil.bobrovsky@kodomo.fbb.msu.ru

Резюме

Acetobacter pasteurianus – грамотрицательные бактерии, окисляющие этанол до уксусной кислоты. В данной работе был проведен анализ генома и протеома *Acetobacter pasteurianus subsp. ascendens*, штамм LMG 1590, с помощью Excel и программ на Python. В рамках исследования были получены данные по распределению длин и основных типов белков. Было обнаружено, что на кольцевой хромосоме и плазмидах гены распределены по цепям ДНК с разной степенью случайности. Кроме того, были изучены пересечения генов и идущие подряд гены, которые могут оказаться в одном опероне.

Введение

Acetobacter pasteurianus subsp. ascendens – подвид семейства Acetobacteraceae. Это палочковидные облигатно аэробные грамотрицательные бактерии, получающие энергию, окисляя этанол до уксусной кислоты[1]. Род *Acetobacter* отличает способность также окислять ацетат и лактат до углекислого газа и воды[2]. Эти бактерии имеют огромное хозяйственное значение, так как используются для производства уксуса, а также могут портить вина, производя уксусную кислоту или этилацетат. Поэтому исследование генома *Acetobacter* является чрезвычайно важной и актуальной задачей.

Материалы и методы

Описание генома бактерии было получено из базы данных NCBI: https://www.ncbi.nlm.nih.gov/genome/genomes/70944 (штамм LMG 1590, банк GenBank). Для

анализа были использованы скрипты на Python 3.7 и программа MS Excel (таблицы и программы можно найти в сопроводительных материалах). Диаграммы и таблицы были сделаны в MS Excel.

Были созданы сводные таблицы MS Excel по всему геному, кольцевой хромосоме и каждой из трех плазмид. Длины белков были проанализированы с помощью встроенных функций MS Excel. Из сводных таблиц были получены данные по распределению генов по цепям ДНК: белок-кодирующими генами считались гены со значениями «RNase_P_RNA» и «protein_coding» в столбце «class», псевдогенами – со значением «pseudogene», генами РНК – со значениями «ncRNA», «rRNA», «SRP_RNA», «tmRNA», «tRNA». Случайность распределения по цепям ДНК была определена с помощью скрипта на Python, подсчитывающего вероятность такого распределения (неслучайными считаются события с вероятностью меньше 0.05). Программа случайно выбирает одно из двух значений столько раз, сколько генов закодировано на обеих цепях, и затем сравнивает полученную разность числа значений с разностью числа генов на разных цепях. Эти действия программа повторяет 100 тысяч раз, и если разность была больше или равна наблюдаемой в геноме разности меньше, чем в пяти процентах случаев, программа считает такую разность неслучайной.

Также было определено число квазиоперонов - последовательностей генов, закодированных на одной цепочке с промежутками между ними не более 100 пар нуклеотидов (п.н.). Подсчет был выполнен с помощью программы на Python 3.7, статистическая обработка – с помощью встроенных функций MS Excel. Кроме того, были получены данные о попарных пересечениях генов (здесь, как и при подсчете длины квазиоперонов, из рассмотрения были исключены аннотированные гены длиной более 5000 п.н.).

Наконец, с помощью программы на Python 3.7 был проведен анализ распределения белков, закодированных на хромосоме и плазмидах, по различным категориям. К мембранным белкам были отнесены белки с подстрокой «membrane» в столбце «name», к транспортным – белки с подстрокой «transport», к белкам, работающим с РНК, – с подстрокой «RNA». Таким образом, полученные данные о числе белков в этих категориях означают лишь минимальное их число в геноме, в то время как реальное их число больше.

Результаты и обсуждение

Длины белков

Длины белков *Acetobacter ascendens* лежат в интервале от 44 до 1518 аминокислотных остатков (а.о.). Наиболее короткий белок - 50S рибосомальный белок L34, наиболее длинный – альфа-субъединица глутаматсинтазы. Средняя длина (309 а.о.) достаточно близка к медианной (270 а.о.), однако стандартное отклонение сравнительно велико (203,4). Из этого можно сделать

вывод о разнообразии протеома бактерии. С этим согласуется и приведенная на Рис.1 гистограмма длин белков.

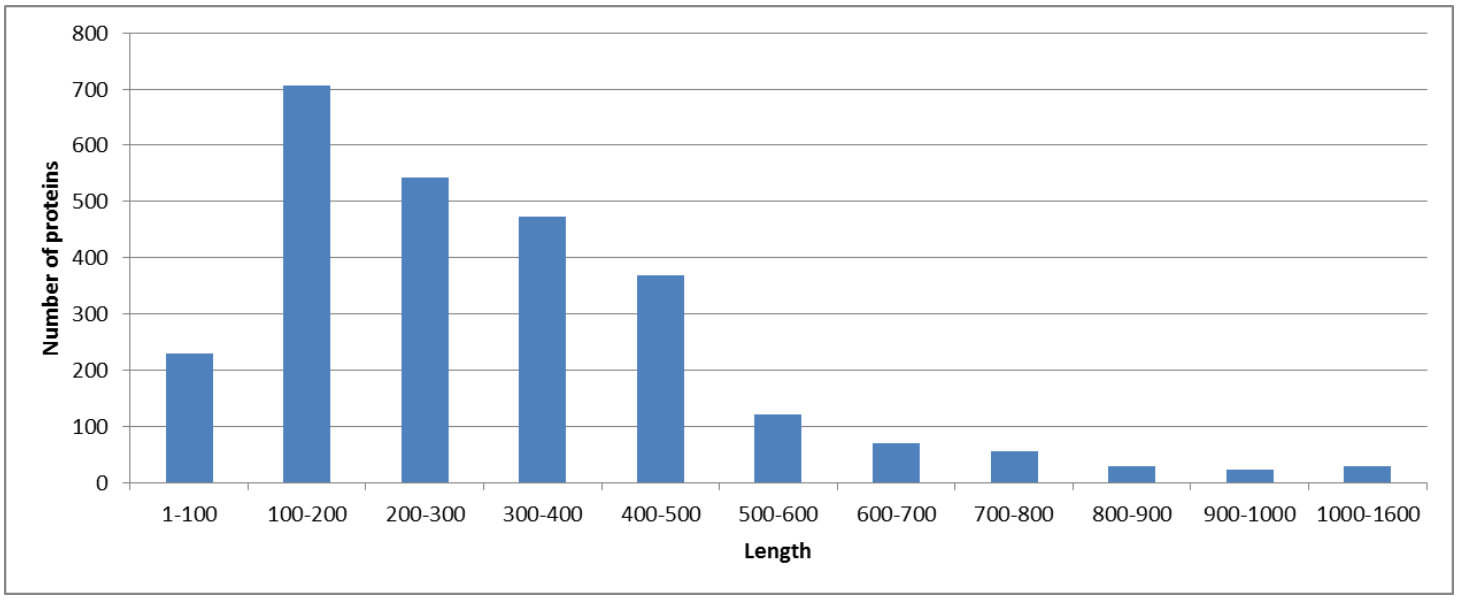

Рис.1. Распределение длин белков. *По вертикали: число белков в геноме. По горизонтали: длина белка в аминокислотных остатках.*

Типы белков

Из общих соображений о типах белков были выделены следующие категории: предполагаемые (гипотетические), мембранные, транспортные и рибосомальные белки. Кроме того, была замечена высокая частота встречаемости транспозаз, была добавлена также категория белков, работающих с РНК. Результаты подсчета представлены в Таблице 1.

Гены гипотетических белков составляют 25% от общего числа, и встречаются в заметных количествах как на хромосоме, так и на плазмидах. Мембранных белков было обнаружено всего 9, и все из них на хромосоме. Вероятно, это связано с несовершенством метода подсчета (не учитывались, например, порины и другие белки, не имеющие в названии слова «membrane»). Гены транспортных белков, в свою очередь, были обнаружены в значительном количестве (6,6% от общего числа белок-кодирующих генов). Причем интересно, что на двух плазмидах их почти нет, а на третьей они составляют 12,5% от общего числа. То есть транспорт веществ может оказаться одной из важных функций, обеспечиваемых этой плазмидой. Наиболее интересным в распределении белков по категориям оказалось то, что 10% генов белков являются генами транспозаз – ферментов, перемещающих в геноме транспозоны. Возможно, с частым перемещением транспозонов может быть связано и большое число псевдогенов. Однако эта гипотеза нуждается в дальнейшей проверке. Наконец, гены рибосомальных и работающих с РНК

белков составляют суммарно 5% от общего числа и встречаются только на хромосоме, что говорит об их исключительной важности в метаболизме бактерии (большинство из них относятся к «генам домашнего хозяйства»).

	Хромосома	Плазмида 1	Плазмида 2	Плазмида 3	Весь геном	Процент
Гипотетические	628	18	14	13	673	25,17%
Мембранные	9	0	0	0	9	0,34%
Транспортные	169	1	0	6	176	6,58%
Транспозазы	251	12	5	6	274	10,25%
Работающие с РНК	81	0	0	0	81	3,03%
Рибосомальные	56	0	0	0	56	2,09%
Другие	1345	16	21	23	1405	52,54%

Таблица 1. Распределение белков по категориям.

Распределение генов по цепям

По распределению генов по цепям были получены особенно интересные данные, представленные в таблице 2 (более подробные таблицы с информацией по разным типам генов есть в сопроводительных материалах). По всему геному в целом и по хромосоме распределение случайное, однако на всех трех плазмидах генов на одной из цепей значительно больше, чем на другой. Результат работы программы на Python показал, что такое распределение можно считать неслучайным ($p<0,05$). Вероятно, неравномерное распределение связано с особенностями репликации плазмид, ведь и на хромосоме гены могут располагаться чаще на одной из цепей по разным сторонам от точки начала репликации.

	Хромосома	Плазмида 1	Плазмида 2	Плазмида 3
"+"-цепь	1384	35	29	40
"-"-цепь	1400	18	15	10

Таблица 2. Распределение генов по цепям.

Квазиоперонs

Стоит еще раз упомянуть, что квазиоперонами считались последовательности идущих подряд на одной цепи генов, расстояние между которыми не превышает 100 пар нуклеотидов. Число квазиоперонов было подсчитано с помощью программы на Python, результаты представлены в таблице 3. Существенных различий в числе или длине квазиоперонов между цепями ДНК не наблюдается (на плазмидах эти различия есть, однако они вызваны, очевидно, разным числом генов). Большая часть генов входит в состав квазиоперонов длиной в один ген, что видно из средней длины квазиоперона (1.88 гена). Стандартное отклонение по сравнению с этой

величиной достаточно велико, что, как и большое максимальное значение длины квазиоперона, равное 21 гену, говорит о том, что длина квазиоперонов достаточно разнообразна. Стоит упомянуть, что различия в суммарном числе генов с Таблицей 2 вызваны исключением из рассмотрения при подсчете квазиоперонов и числа пересечений генов длиной более 5000 пар нуклеотидов.

	"+"-цепь	"-"-цепь	Всего
Число генов	1486	1442	2928
Число квазиоперонов	787	770	1557
Максимальная длина квазиоперона	21,00	16,00	21,00
Средняя длина	1,89	1,87	1,88
Стандартное отклонение	1,64	1,54	1,59

Таблица 3. Число и длина квазиоперонов

Пересечения генов

На одной цепи было обнаружено 224 попарных пересечений генов, на другой – 228. Кроме того, существуют различия в числе пересечений на плазмидах, которые, как и в случае числа квазиоперонов, объясняются разницей в числе самих генов.

Заключение

Геном *Acetobacter pasteurianus subsp. ascendens* может оказаться интересным объектом для дальнейшего изучения. Были сделаны некоторые интересные наблюдения, такие как большое число генов транспозаз и неравномерное распределение генов по цепям в плазмидах. Хотя эти данные и объясняются некоторыми гипотезами, в ходе дальнейших исследований могут возникнуть неожиданные результаты.

Сопроводительные материалы

Все сопроводительные материалы можно найти на моем сайте:
https://kodomo.fbb.msu.ru/~daniil.bobrovsky/term1/excel/block4.html

Благодарности

Я хотел бы поблагодарить Софью Гайдукову за совместное обсуждение и разработку алгоритмов, а также Спирина С.А. за помощь в нахождении способа проверки случайного распределения генов по цепочкам ДНК.

Список литературы

1. Madigan M; Martinko J (editors). Brock Biology of Microorganisms. — 11th ed. — Prentice Hall, 2005. — ISBN 0-13-144329-1.
2. Cleenwerck I; Vandemeulebroecke D; Janssens D; Swings J (2002). "Re-examination of the genus Acetobacter, with descriptions of Acetobacter cerevisiae sp. nov. and Acetobacter malorum sp. nov". International Journal of Systematic and Evolutionary Microbiology. 52: 1551–1558. doi:10.1099/00207713-52-5-1551. PMID 12361257. Retrieved 23 December 2015.