Обзор протеома бактерии Hydrogenobacter thermophilus

Гайдукова Софья

Факультет Биоинженерии и биоинформатики, Московский Государственный Университет им. Ломоносова, Ленинские горы 1-73, Москва, Россия

sofyagdk@gmail.com

Резюме

Мы проанализировали геном бактерии с уже картированными на нем генами белков

Мы анализировала длины белков и их распределение по двум цепям, получив что максимальное число белков – относительно короткие от 50 до 450 аминокислот. Были рассмотрены белки крайней длины. Распределение генов по цепям случайное. Мы также смотрели на пересечения генов на одной цепи, выявив довольно частные пересечения – в ~35% случаев. При этом не было пересечений, длина которых была бы кратна трем.

Ключевые слова: Hydrogenobacter thermophilus, геном, протеом, длина белков, пересечения

Введение

Hydrogenobacter thermophilus - вид грамотрицательных неподвижных бактерий. Бактерия является экстремальным термофилом, оптимум температур: 70-75*C. Штамм был выделен из почвы рядом с горячими источниками на Izu penninsula, Japan.

Обмен хемолитоавтотрофный, молекулярный водород донор электронов, а углекислый газ источник углерода. Хемоорганотрофной активности найдено не было.

При этом окисление происходит на мембране при помощи белка гидрогеназа, катализирующего реакцию H2 -> 2H + 2e

В геноме бактерии нет интронов, поэтому анализ картированных на геном генов белков позволяет с большой точностью сделать выводы о самих белках.

Материалы и методы

Геном был скачан с сайта NCBI из банка RefSeq.
https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/Hydrogenobacter%20thermophilus

Скачанный файл можно посмотреть по ссылке
https://kodomo.fbb.msu.ru/~sofyagdk26/term1/block4.html

Для анализа длин белков использовались средства MS Office16 Excel.
Рассматривались строки со значением “gene” в колонке “feature” и “protein_coding” в колонке “class”.

В колонке “class” встретились значения

Значение	Как интерпретировалось
protein_coding	белки
pseudogene	не экспрессируется
rRNA	РНК
SRP_RNA	
tmRNA	
tRNA	

Для оценки случайности распределения по цепям был написан скрипт для Python 3.5. Если мы считаем, что гены распределены по цепям равновероятно, то тогда математическое ожидание (наиболее вероятный случай) разницы между количеством генов на одной цепи и на другой будет 0. Чем больше разница, тем менее вероятен исход. Программа генерирует случайную последовательности из 1 и -1 такой же длины что и количество генов в геноме бактерии, и считает, в какой доле случаев эта разница больше, чем в настоящем геноме. Порог значимости был взят 5%.

Для анализа пересечения был написан скрипт для Python 3.5. Каждая цепочка рассматривалась в отдельности. Программа считает количество попарных пересечений. Также составляет и печатает в файл массив, где для каждой встреченной длины пересечения хранится количество таких пересечений.

Для подсчета количества квазиоперонов был написан скрипт для Python 3.5. Квазиопероном считалась максимальная последовательность генов, закодированных на одной цепочке с промежутками между генами не более порога 100 п.н.

Результаты и обсуждение

Длины белков

Всего белков в геноме закодировано 1870 белков. Их длина в аминокислотах изменяется, как видно из Таблицы 1, от 38 ак до 1566 ак. Три самых коротких белка – рибосомальные белки 50S (большой) субъединицы. То есть входят в комплекс белков, окружающих рРНК, формируя рибосому. Это высокоспецифичные и высококонсервативные белки, что в принципе согласуется с тем, что они обладают небольшой длиной. Интересно, что в геноме они не расположены рядом, как видно из таблицы 2 (см колонку начало)

start	protein name	product_length
673876	50S ribosomal protein L36	38
1258181	50S ribosomal protein L34	47
806673	50S ribosomal protein L33	52

Таблица 1. *Сводная таблица по самым коротким белкам. В столбце 1 обозначен номер первого нуклеотида кодирующей последовательности*

Самый длинный белок – бэта’-субъединица ДНК-зависимой РНК-полимеразы, которая собственно катализирует присоединение РНК к удлиняющейся цепочке. Бэта’, в отличие от бэта субъединицы, связывется с ДНК неспецифично, поэтому не удивительно, что ее длина больше – должны присутствовать дополнительные процессы регуляции. Заметим, что гены этих субъединиц идут друг за другом в геноме.

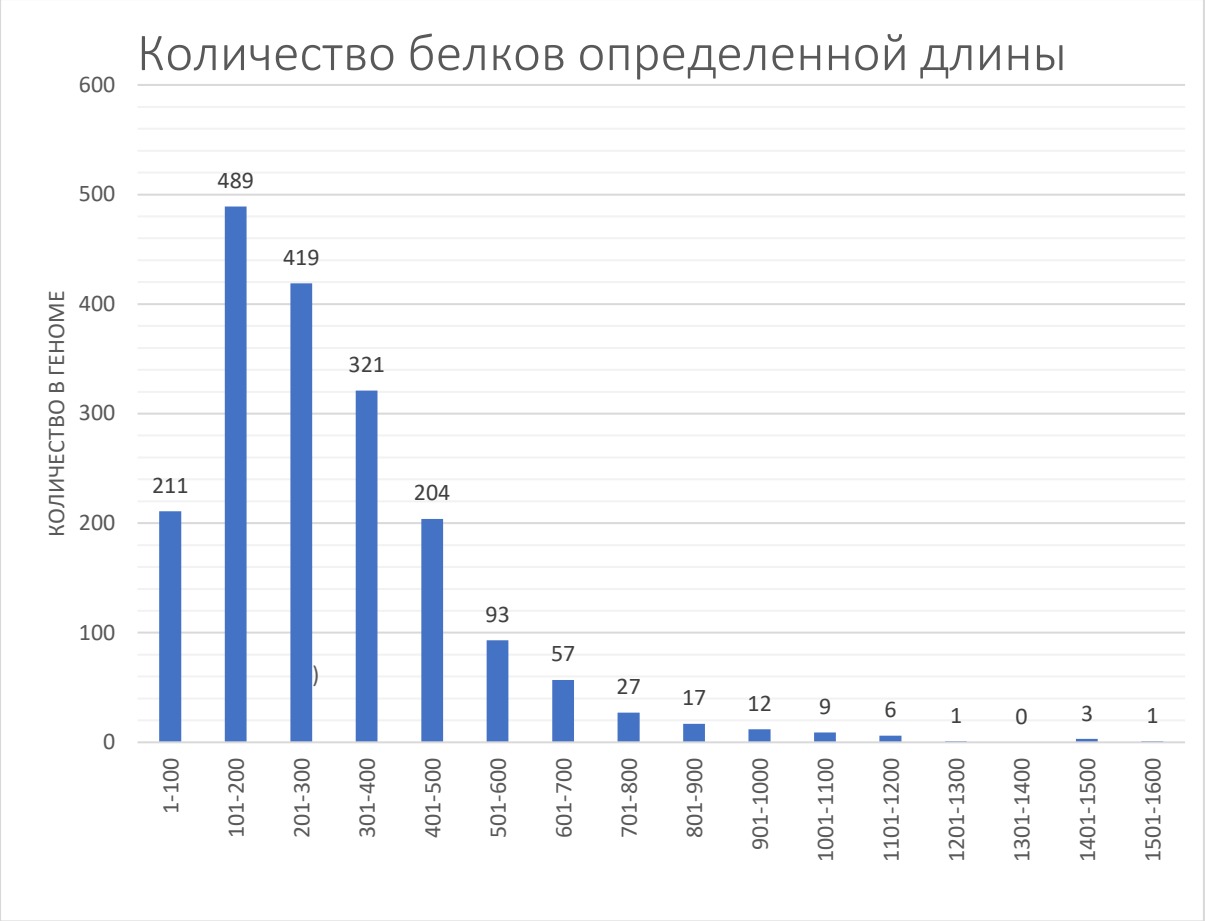
start	name	product_length
794307	DNA-directed RNA polymerase subunit beta'	1566
810636	glutamate synthase large subunit	1500
799004	DNA-directed RNA polymerase subunit beta	1469
290216	restriction endonuclease	1408

Таблица 2. *Сводная таблица по самым длинным белкам. В столбце 1 обозначен номер первого нуклеотида кодирующей последовательности*

Минимальная длина	38
Максимальная длина	1566
Средняя	296.8059
Стандартное отклонение	196.6289
Медиана	253

Таблица 3. *Статистика по длине белков*

Посмотрим на среднее значение длины - ~300ак, что совсем не сильно отличается от медианы, что свидетельствует о том, что действительно подавляющее количество белков приходится на 100-400 аминокислот (65%). Это видно и из гистограммы. Заметим, что стандартное отклонение довольно велико, значит можно сделать вывод, что белковый состав бактерии довольно разнообразен по своему составу.



Гистограмма 1. *Количесво белков определенной длины в геноме*

Пересечения

Пересечения были посчитаны на каждой цепи в отдельности. Все пересечения были поделены на три класса в зависимости от своей длины – по делимости на три. Оказалось, что пересечений, длина которых была бы кратна трем, нет (статистически незначимый единичный случай на одной из цепей относится к пересечению с РНК кодирующим участком). И это логично, ведь последний кодон в кодирующей последовательности – стоп кодон, таким образом пересечение на кратное трем означало бы постоянное наличие преждевременной остановки в транскрибируемой РНК, так что в таком пересечении смысла нет.

Заметим, что пересечений со смещением +2 в два раза больше, чем со смещением +1. Оказалось, что пересечений длины 4 (смещение +2) очень много (144 из 342 и 158 из 342 для + и – цепей соответственно, см Сопровождающий файл 1, лист “intersection”). Последние три буквы в большинстве случаев должен быть стоп-кодон (TAA/TAG/UGA). Обычно наиболее частый старт кодон – AUG. Тогда можно предположить, что эти 4 пересекающиеся буквы чаще всего AUGA. (Это не был проверено на нуклеотидных последовательностях). Тогда можно предположить, что у данной бактерии наиболее часто встречающимся стоп кодоном может быть UGA, или же UGA будет коррелировать с наличием пересечений. В общем, это еще подлежит анализу на нуклеотидной последовательности генома.

Вышло так, что на обеих цепях количество пересечений одинаковы, но автор считает, что это случайность, или по крайней мере разумно интерпретировать данный факт не может, так как по длинам пересечения различны и это не ошибка программы.

remainder 3	0	1	2	sum
+ (main)	0	232	110	342
- (complementary)	1	228	113	342

Таблица 4. *Статистика пересечений по группам в зависимости от делимости длины пересечения на три*

Распределение

В геноме встречались три вида кодирующих последовательностей: кодирующие белки, кодирующие РНК, и последовательности, утратившие способность кодировать белок и не экспрессирующиеся. Как оказалось, по основной и комплементарным цепям все три класса генов распределены случайным образом, то есть каждый ген попал на + или – цепь с вероятностью 50%.

Таким образом можно предположить, что функционально в клетке бактерии эти цепи не отличаются между собой.

	protein	pseudogene	RNA
+ (main) strand	928	10	22
-(complementary) strand	942	8	27
difference	-14	2	-5
outarea percentage	0.732	0.497	0.375
distribution	random	random	random

Таблица 5. *Статистика распределения различных классов генов по двум цепям хромосомы*

Квазиопероны

Было выявлено 287 и 277 квазиоперонов на цепях. Больше всего приходилось на квазиопероны длиной 3 и 5 генов. Причем единичных генов было выявлено 96 на одной и 86 на другой (см Сопровождающий файл 1, лист “qvasi”), что по сравнению с длинами цепей крайне мало (928 и 942). Эти данные согласуются с наличием большого количества пересечений. В целом можно сделать вывод о том, что в геноме очень мало межгенных нуклеотидов и интронов.

Сопроводительные материалы

- Сопроводительный материал 1. Геном изучаемой бактерии и анализ (*https://kodomo.fbb.msu.ru/~sofyagdk26/term1/Hydrogenobacter_thermophilus_TK-6.xlsx*)
- Сопроводительный материал 2. Геном изучаемой бактерии - изначальные данные (*https://kodomo.fbb.msu.ru/~sofyagdk26/term1/GCF_000010785.1_ASM1078v1_feature_table.txt*)
- Сопроводительный материал 3. Программа: проверяет случайно ли распределение (*https://kodomo.fbb.msu.ru/~sofyagdk26/term1/randomornot.py*)
- Сопроводительный материал 4. Программа: считает количество квазиоперонов (*https://kodomo.fbb.msu.ru/~sofyagdk26/term1/qvasi.py*)
- Сопроводительный материал 5. Программа: анализирует пересечения генов (*https://kodomo.fbb.msu.ru/~sofyagdk26/term1/peresecheniya.py*)

Благодарности

Я хотела бы поблагодарить Даниила Бобровского за идеи по пересечениям, помощь с теорией по статистике и мозговой штурм при разработке алгоритмов.

Список литературы

1. http://www.microbiologyresearch.org/docserver/fulltext/ijsem/34/1/ijs-34-1-5.pdf?expires=1545194701&id=id&accname=guest&checksum=C1D925EBE2A0F3421ADAE920400467BB
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3111988
3. https://en.wikipedia.org/wiki/Hydrogenobacter_thermophilus
4. https://en.wikipedia.org/wiki/RNA_polymerase
5. https://en.wikipedia.org/wiki/Ribosomal_protein
6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215218/
7. https://itol.embl.de/itol.cgi
8. https://www.nature.com/articles/s41598-017-12619-6