ОБЗОР ПРОТЕОМА БАКТЕРИИ CLOSTRIDIUM THERMOCELLUM ATCC

ПОПОВ АЛЕКСЕЙ АЛЕКСЕЕВИЧ

ФББ МГУ им. М.В. Ломоносова, University Name, Moscow, Russia.
concitorem@fbb.msu.ru

Краткий обзор протеома Clostridium thermocellum, являющийся частью работы, посвящённой использованию Microsoft Excel. В данном обзоре большое внимание уделено распределению генов по хромосоме бактерии. В ходе работы было получено множество дополнительных данных, которые также частично освещены в осуждении.

1. Введение

Clostridium thermocellum - палочковидная грамположительная бактерия, живущая в анаэробных, термофильных условиях. Принадлежит C. thermocellum к классу Clostridia (подробная классификация в таб.1 [3]), однако, в отличие от большинства других организмов, входящих в данный таксон, она не выделяет токсичных веществ и не считается патогенным организмом. Одной из особенностей данного организма, привлекающей биотехнологов, является способность разлагать целлюлозу. Также способность расщеплять целлюлозу означает, что бактерия может жить практически в любой среде, содержащей органические соединения, например, в почве, отложениях на дне водоёмов, кишечном тракте животных [1]. Генетическая информация организма закодирована в двухцепочечной кольцевой молекуле ДНК, имеющей длину 3,843,301 пар оснований. Геном бактерии содержит около 3,376 генов. Из них 3,224 кодируют белки, а остальные 152 – РНК [2].

Основной целью нашей работы являлось изучение функционала Microsoft Excel.
В ходе работы было изучено распределение длин белков, синтезируемых данным организмом. Также мы подсчитали некоторые другие количественные и качественные параметры генома бактерии и исследовали закономерности распределения генов на цепях ДНК.

2. Материалы и методы

В своей работе я использовал GBK файл с геномом Clostridium thermocellum ATCC, содержащийся на сайте NCBI (https://www.ncbi.nlm.nih.gov/nuccore/125972525). Идентификатор хромосомы - NC_009012.1.
Данные были импортированы в файл Excel, где с ними проводилась дальнейшая работа. В ходе работы были использованы такие функции как СЧЁТЕСЛИМН(),

производящая подсчёт элементов, удовлетворяющих условию; СЛУЧМЕЖДУ(), позволившая имитировать случайное распределение; ПОИСК(), возвращающая начало искомого фрагмента; ПРАВСТР() и ЛЕВСТР(), возвращающие срез содержимого ячейки (справа и слева соответственно).

3. Результаты

3.1. *Число генов белков и генов РНК по категориям. Примерная оценка числа генов на 1 млн пар нуклеотидов (п.н.).*

3.1.1. *Число генов белков и генов РНК по категориям.*

Как уже было отмечено выше, геном бактерии содержит около 3,376 генов. Из них 3,224 кодируют белки, а остальные 152 – РНК. При этом 56 кодируют тРНК, 12 – рРНК, один – ncРНК, два – misc РНК. Имеется и 81 псевдоген. То есть ген, утративший способность кодировать белки.

3.1.2. *Примерная оценка числа генов на 1 млн пар нуклеотидов (п.н.).*

Рисунок 1

Для примерной оценки плотности генов P пользовались простейшей формулой (1). Где G – количество генов. L – длина генома (в млн пар нуклеотидов). Полученное значение выражали в количестве генов на 10^6 пар оснований (gen/Mbp).

$$P = G/L$$

Итоговая плотность - 878,4 gen/Mbp. Однако, выполнение этого задания породило гипотезу о том, что гены на хромосомах распределены не диффузно и оценка не

будет полностью описывать происходящее. Для проверки гипотезы была построена гистограмма, отражающая распределение генов бактерии на цепях ДНК (Рис.1). Видно, что определённые закономерности всё же существуют.

3.2. *Гистограмма длин белков из протеома своей бактерии или археи.*

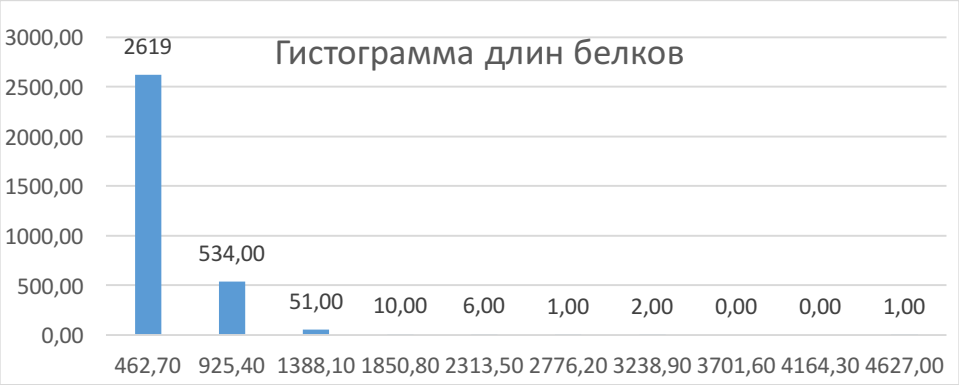

Рисунок 2

На основании имеющихся данных была построена гистограмма, отражающая количество белков, входящих в различные диапазоны длин (Рис. 2).

На ней видно, что подавляющее большинство белков организма имеют длину порядка 500 амк.

При этом самый большой белок имеет длину 4627 амк., а самый маленький – 38 амк.

3.3. *Таблица числа генов белков и генов РНК на прямой и комплементарной цепочках ДНК.*

С помощью данных из GBK файла и xlsx-файла с протеомом бактерии была построена таблица, в которой отражено расположение различных генов на прямой и комплементарной цепи (Таб. 1).

Таблица 1

Цепь	Ген белка	Ген РНК
Sense	1605	44
Antisense	1619	27

3.4. *Проверка гипотезы о том, что гены распределены по цепочкам случайно с вероятностями 0,5.*

Для ответа на этот вопрос я создал в Excel таблицу из 3376 строк и 100 столбцов, заполнив каждую из ячеек 0, или 1. При этом ячейки моделировали случайность расположения генов (например, 0 – sense, 1 – antisense) на цепях. Далее подсчитывалось отклонение от идеального случая (с вероятностью 0,5). И таблица выдавала долю экспериментов, в которых отклонение распределения генов в геноме моей бактерии меньше, чем полученные. Обычно таблица выдаёт 90% (+/- 10%), что говорит о случайности распределения генов.

3.5. *Количество "квазиоперонов" в геноме бактерии.*

Для ответа на вопрос было составлено две таблицы: ввода-вывода и обсчётная. На первой таблице есть графа для ввода расстояния между генами, которые находятся в одном квазиопероне. Следующая же графа выводит количество квазиоперонов.
Для расстояния между квазиоперонами 100 bp был составлен график, отражающий зависимость количества квазиоперонов от их длины (Рис. 3).

Рисунок 3

Что же такое квазиоперон. Является ли он чем то важным, или просто результатом случайного совпадения? Для ответа на этот вопрос обратимся к статистике. Вычтем из длины генома длину кодирующей части. После этого воспользуемся функцией СЛЗНАЧ(), расставив ей 3224 гена в некодирующем пространстве. Нетрудно теперь найти количество случаев, в которых расстояние между генами будет меньше 100

bp. Примерно 350 случаев. Тогда как в реальной бактерии таких случаев около 1200. Неслучайность доказана.

3.6. *Статистические данные о пересечениях генов.*

Таблица 2

Тип пересечения	Пересечения на одной цепи			Пересечения между двумя цепями		
Сдвиг рамки	0	-1	-2	0	-1	-2
Количество	0	91	76	4	94	68

Очевидно, что гены на одной цепи не будут иметь пересечения со сдвигом рамки на 0, так как при этом второй ген не будет предсказан (Таб. 2).

3.7. *Статистика белков по категориям достоверности их существования.*

Таблица 3

Category	Number
Evidence at protein level	57
Evidence at transcript level	6
Inferred from homology	759
Predicted	2003
No information	399

Было выделено 5 категорий достоверности существования белков (Таб. 3).

4. Обсуждение

Основным результатом своей работы я считаю большое количество доказательств (пункты 1 и 5) неслучайности расположения генов на хромосоме, которому противоречит пункт 4. В дополнительных результатах 1 пункта содержится информация, подтверждающая очевидный факт того, что расположение генов на комплементарных цепях зависит друг от друга. Они стремятся к наименьшему перекрыванию (при перекрывании на одной цепи возможно кодировать нормальную структуру, а при комплементарном перекрывании – нет. Так как имеется лишь две довольно чёткие границы, отделяющие зоны интенсивного расположения sense и antisense генов, то можно выдвинуть гипотезу, что изначально все гены располагались на одной цепи, но в какой-то момент произошли парные двуцепочечные разрывы на хромосоме, которые после репарации дали эти самые пороги.

6 и 7 пункты являются подтверждением того, что большая часть транскриптов данного организма являются лишь биоинформатически предсказанными и не имеют никакой связи с реальным состоянием дел.
2 пункт показывает, что идеальная для белка длина порядка 500 амк. остатков, что связанно с тем, что такой белок гораздо более просто сворачивается, имеет меньшую вероятность нарушения структуры, а в следствие этого находится в более выиграшной с точки зрения эволюции позиции по отношению к другому белку, выполняющему ту же функцию, но имеющему больший размер.

5. Заключение

В ходе работы были выявлены некоторые закономерности распределения генов на кольцевой хромосоме организма. Также стало понятно, что геном протеом данной бактерии есть лишь плод предсказания биоинформатиков.

6. Сопроводительные материалы

1) Файл с обработкой данных;

2) Таблица с протеомом.

7. Благодарности

Хочу выразить благодарность кафедре биоинформатики ФББ МГУ за предоставленное задание, которое было интересно делать, а так же за советы по мере его выполнения.

8. Список литературы

1) Todar, Kenneth. "The Pathogenic Clostridia". Online Textbook of Bacteriology. University of Wisconsin-Madison, Department of Bacteriology. 2006;

2) Clostridium thermocellum ATCC 27405, complete genome;

3) Taxonomy - Clostridium thermocellum (Ruminiclostridium thermocellum).