ОБЗОР ПРОТЕОМА БАКТЕРИИ RUMINICLOSTRIDIUM THERMOCELLUM AD2

СУТЫРИН ЕГОР АЛЕКСЕЕВИЧ

ФББ МГУ им. М.В.Ломоносова, Москва, Россия
egor_su@fbb.msu.ru

Краткий обзор протеома R.Thermocellum и простейшие выводы. Работа выполнена в рамках курса по изучению Exel. Содержит анализ количества белков и из распределения по длине, функции и кодирующим цепям.

Ключевые слова: Протеом; Ruminiclostridium thermocellum; Exel.

1. Введение

Ruminiclostridium thermocellum(или Clostridium thermocellum[1]) – грамположительная палочковидная бактерия из класса Clostridia[2].Она термофильна и анаэробна, но при этом не считается патогенной, в отличии от большинства Клостридий[3]. Для биологов данная бактерия интересна во многом своей способностью расщеплять целлюлозу[4], в частности штамм AD2 отличается от дикого типа дефектным механизмом адгезии целлюлозы[1].
Основными целями нашей работы являются изучение способов обработки данных и их прменение

2. Материалы и методы

Данные для анализа были взяты с сайта NCBI, из архива feature table, а после импортированы в Microsoft Excel. В результате была получена плоская таблица, содержащая информацию о генах, белках, транспортной и рибосомальной РНК бактерии.
Полученная таблица была обработана при помощи встроенных в Excel функций, таких как СЧЁТЕСЛИМН, позволяющей посчитать количество строк, удовлетворяющих заданному набору условий, МИН и МАКС, показывающих, соответственно, минимальное и максимальное значение ячейки на заданном промежутке, СРЗНАЧ, СТАНДОТКЛОН и МЕДИАНА, возвращающих соответствующую статистическую характеристику для заданного диапазона.
В том числе было посчитано количество строк с определёнными значениями первой и второй ячеек(feature и class), результаты представлены в таблице 1. Количество прсевдогенов, тРНК и иРНК было посчитано по количеству строчек с соответствующим значением ячейки class. Для белков использован столбец «with_protein», т.к. для каждого белка есть и строка с feature «gene», и строка с

feature «CDS», но feature «CDS» есть только у белков, в отличие от feature «gene». Белки, выполняющие определённую функцию были посчитаны путем поиска по ключевым словам в названии белков(). Данные при этом были отсортированы так, что отображались только строки с feature «CDS».

Таблица 1. Feature - class

	class	protein_coding	with_protein	tRNA		pseudogene	rRNA
# feature	1	0	0	0	0	0	0
gene	0	3035	0	56	0	12	12
CDS	0	0	3035	0	0	0	0
tRNA	0	0	0	0	56	0	0
rRNA	0	0	0	0	12	0	0

3. Результаты

3.1. *Белки*

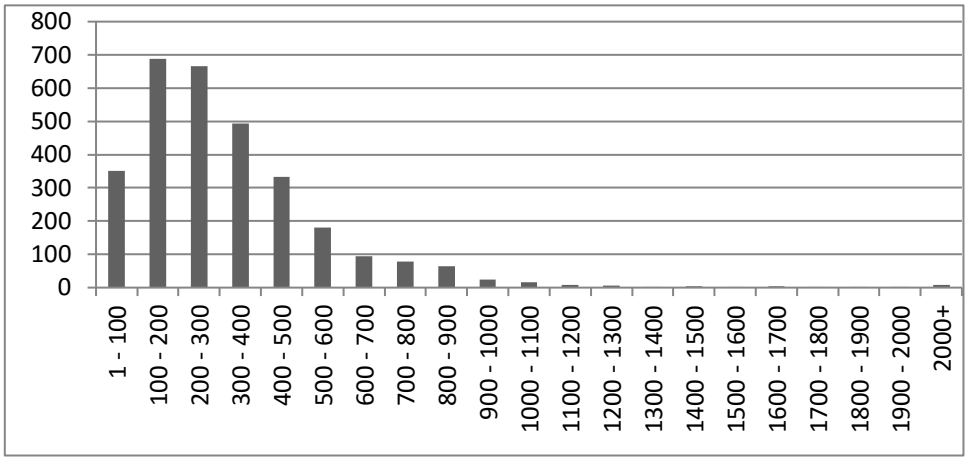
Согласно табл.1 в протеоме бактерии закодировано 3035 белков. Статистические данные о них представлены в таблице 2. Распределение белков по длине показано на гистограмме 1, по горизонтали - интервалы длины, по вертикали – количество белков.

Таблица2. Статистические данные о белках R.thermocellum AD2.

Характеристика	Значение, а.о.
Минимальная длина	30
Максимальная длина	8366
Средняя длина	329,192
Стандартное отклонениие	302,204
Медиана	269

Как видно из гистограммы, у бактерии больше всего белков с длиной от 100 до 300 аминокислотных остатков, количество более длинных белков стремительно уменьшается при увеличении длины.

Гистограмма 1. Распределение белков по длине.



Белки были разделены на 4 группы, в зависимости от их функции. Результаты показаны в таблице 3. Из неё видно, что значительная часть белков(22.4%) – гипотетические, рибосомальных и транспортные белков немного(2,4%-4,3%), несмотря на важность их функции.

Таблица 3. Группы белков R.thermocellum AD2.

Тип белка	**Кол-во**
Транспортные	131
Рибосомальные	79
Гипотетические	678
Все остальные	2147

3.2. *РНК*

Из таблицы 1 видно, что у R.thermocellum AD2 в протеоме есть только транспортные и рибосомальные РНК. 56 тРНК и 12 рРНК.

3.3. *Распределение генов в геноме*

Геном содержит в себе 3115 генов, представлен кольцевой молекулой ДНК и имеет длину 3554854 пар нуклеотидов. Средняя плотность расположения генов

876,5 генов на 1 миллион пар нуклеотидов, распределение генов по прямой(+) и обратной(-) цепям ДНК показано в таблице 4.

Таблица 4. Количество генов на прямой и обратной цепи ДНК.

Цепь	Белки	Псевдогены	РНК
+	1476	5	37
-	1559	7	31

4. Обсуждение

Как видно из пункта 3.1 длины белков различаются более чем на 2 порядка, но при этом длина почти всех белков находится в интервале 30 – 700 а.о. , что позволяет предположить, что бактерия стремится сделать свои белки короче. В сумме с маленьким процентом рибосомальных и транспортных белков, это наталкивает нас на вывод, что бактерия предпочитает решать даже самые важные задачи минимальными затратами. Также заметно большое количество гипотетических белков в протеоме, что говорит о том, что он исследован не до конца.
Странным кажется присутствие в геноме только транспортных и рибосомальных РНК. Этому есть 2 возможных объяснения: либо в бактерии отсутствуют любые РНК кроме транспортных, рибосомальных, и матричных, что выглядит неправдоподобно, либо оставшиеся РНК производятся путем модификации других РНК и тогда они присутствуют в протеоме, но не выделяются. Также можно заметить, что тРНК у бактерии 56, а аминокислот гораздо меньше. Поиск по исходным данным показал, что у Ruminiclostridium termocellum бывает до 5 тРНК на 1 аминокислоту. Это отчасти противоречит нашему первому выводу, объяснения этому у нас пока нет.
Количество генов белков, псевдогенов, и РНК на + и – цепях примерно равно.

5. Заключение

Протеом Ruminiclostridium termocellum все ещё нуждается в исследовании, после нашего анализа все ещё остаются вопросы, требующие дальнейшего исследования. При всем этом бактерия является хорошим модельным организмом для подобных работ.

6. Сопроводительные материалы

Файл с обработкой данных и таблицей с протеомом

Благодарности

Хочу выразить свою благодарность кафедре Биоинформатики ФББ МГУ за предоставленное задание, подсказки к нему, и навыки его выполнения.

Список литературы

1. NCBI data bases
2. MicrobeWiki - free wiki resource on microbes and microbiology.
3. Wikipeida - The Free Encyclopedia
4. JGI Genome Portal