Обзор протеома бактерии Elusimicrobium minutum Pei191

Р. А. Кудрин[1]
[1]Факультет биоинженерии и биоинформатики МГУ им. Ломоносова

РЕЗЮМЕ

В этом мини-обзоре анализируется протеом(совокупность белков) бактерии Elusimicrobium minutum Pei191. В частности, сделана попытка выявления закономерностей в распределении генов по их длинам, а также по положению в молекуле ДНК и др.* Работа сделана в рамках учебного курса практической информатики на факультете биоинженерии и биоинформатики МГУ и представляет скорее учебную, чем научную, ценность. Все вычисления производились с помощью программы Microsoft Office Excel 2007.

ВВЕДЕНИЕ

Предпосылки появления этого учебного мини-обзора появились в 2008 году, когда А.В.Алексеевский и Е.А.Аксянов предложили своим студентам на факультете биоинженерии и биоинформатики создать описания бактерий и их протеомов[1][2][3]. Это задание было немного модифицировано И.Русиновым для тренировки студентов в обработке данных с помощью MS Excel. Написание мини-обзора ознакомит нас со структурой написания научных статей, что, скорее всего, понадобится нам в будущем. Также мы получили импульс к самостоятельному выявлению закономерностей в параметрах протеома бактерии, что прямо показывает, как много мы можем сделать уже сейчас.

МАТЕРИАЛЫ И МЕТОДЫ

Данные о протеоме бактерии Elusimicrobium minutum Pei191 были взяты из открытой базы данных NCBI ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria. Для обработки средствами Excel файлы с расширениями .ptt и .rnt были экспортированы в таблицу формата xlsx, содержащие информацию о генах, кодирующих белки, и генах, кодирующих РНК бактерии соответственно. Затем из экспортированных файлов была создана общая таблица с добавлением графы «Тип гена». Для построения диаграммы сначала был добавлен столбец «Карманы», затем использовалась надстройка Excel «Пакет анализа данных»[4], далее вставка диаграммы по полученному распределению на листе «Гистограмма» (Дополнительные материалы, Elsimicrobium_minutum_Pei191_proteom.xlsx). Для подсчёта генов на прямой и комплементарной цепи использовалась функция СЧЁТЕСЛИ. Для проверки гипотезы о случайном распределении генов с вероятностью 0.5 использовалась функция БИНОМ.РАСП для вычисления p-value(для этого необходимо в поле для параметра «интегральная» прописать ИСТИНА. Эти результаты помещены на лист «Количество генов на цепях». Для подсчёта квазиоперонов и перекрытий между генами заново экспортировались файлы формата .rnt и .ptt с использованием двух разделителей: табулятора и точки, при этом считалось, что два идущих подряд символа-разделителя — тоже разделитель. Далее использовалась функции ЕСЛИ для создания двух бинарных столбцов «Добавляет ли ген квазиоперон» и «Перекрывается ли ген с предыдущим». Далее с помощью функции СУММ были посчитаны суммы значений в этих столбцах. Эти результаты вынесены на лист «Квазиопероны и перекрытия генов». Для выявления закономерностей в параметрах генов, длина которых не делится на 3, был создан лист «Длина в п.н. не делится на 3». Там была использована функция ОТБР и СЧЁТЕСЛИ.

* Более подробно все исследованные параметры описаны в разделе «Результаты».

РЕЗУЛЬТАТЫ

Распределение генов белков по длинам.

Было выявлено, что больше всего генов, кодирующих белки имеют длину от 100 до 400 триплетов. Полное распределение генов по длинам можно увидеть на рис.1

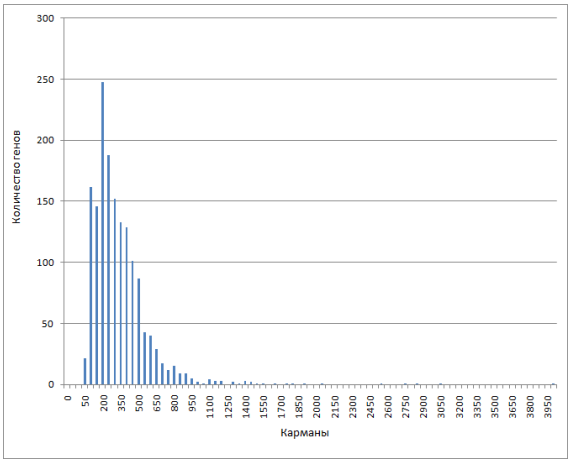

Рис. 1 Распределение генов по карманам различной длины.

Распределение генов по цепям

Таблица 1. Распределение генов по цепям

Тип гена	На прямой цепи	На комплементарной цепи
CDS	760	769
RNA	22	28

P-value для гипотезы случайного распределения генов по цепочкам равно 0,362305357, что соответствует гипотезе о случайном распределении генов по цепочкам с вероятностью 0.5

Квазиопероны и перекрытия между генами

Всего было найдено 1115 квазиоперонов и 202 перекрытия между генами. Перекрытием между генами считалась непрерывная область молекулы ДНК, которую накрывали бы, по крайней мере, два гена.

Количество генов, длина которых в п.н. не делится на 3

Всего найден 31 ген. Все транслируются в РНК-последовательности.

ОБСУЖДЕНИЕ

Длина генов: по результатам исследования, в геноме бактерии преобладают белок-кодирующие гены длины от 100 до 400 триплетов. Факт отстутствия белков меньше определённой(примерно 25 триплетов) длины объясняется тем, что такие короткие белки не могли бы выполнять сложных метаболических функций. Белки гигантского же размера(свыше 700 аминокислотных остатков) встречаются, но их очень мало, и многие из них транслируются только предположительно (hypothetical proteins).

Распределение генов по длинам: данные из таблицы 1 соответствуют гипотезе о случайном расположении генов на цепочках с вероятностью 0.5 при P-value, равном
0,362305357.

Квазиопероны и перекрытия между генами:
Количество генов меньше, чем в 1,5 раза превосходит количество квазиоперонов, что не может не удивлять, так как многие процессы в клетке выполняются группой белков. Явление перекрытия генов можно объяснить тем, что они могут находиться на комплементарных цепочках или в разных рамках считывания.

Гены, длина которых не делится на 3:
Этот феномен легко объясняется тем, что все эти гены кодируют РНК-последовательности, а для РНК триплетность необязательна.

БЛАГОДАРНОСТИ

Выражаю благодарность моим родителям, а также моей жене и детям, которые поддерживали меня в трудные моменты и дарили мне вдохновение во время написания этой статьи.

СПИСОК ЛИТЕРАТУРЫ

1.Artemov.doc
2.Tyshkovskiy.doc
3.Poverennaya.doc
4.Презентация к практикуму 15 Excel-2

ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ

1. Elusimicrobium_minutum_Pei191_proteom.xlsx