Описание протеома археи Methanosaeta concilii GP6

Ефремов А.А.
Факультет биоинженерии и биоинформатики МГУ им. М.В.Ломоносова

РЕЗЮМЕ

Данная работа проведена с целью изучения распределения длин белков археи M.concilii GP6, а также описания расположения генов белков и РНК в геноме. Конкретно под расположением генов в геноме подразумевается их положение на прямой или комплементарной цепях ДНК и пересечение генов друг с другом. Для выполнения этой работы использовалась программа Microsoft Office Excel 2003.

1 ВВЕДЕНИЕ

Известно, что совокупность всех белков организма являет собой протеом. Изучение и описание полного протеома открывает путь к более глубокому пониманию путей метаболизма, клеточного сигналинга и системной биологии организма в целом. В будущем это может привести к открытию новых сторон клеточной физиологии, изобретению перспективных методов лечения, а также к появлению удобных способов получения требуемых биоинженерных конструкций. В последние годы, в связи с развитием компьютерных технологий, растет количество статей с описаниями аспектов системной биологии, геномики, протеомики, метаболомики* и других направлений биологии, требующих анализа больших массивов данных.

2 МАТЕРИАЛЫ И МЕТОДЫ

В этой работе использовались данные полного секвенирования генома археи Methanosaeta concilii GP6, которые были получены из базы данных National Centre of Biotechnology Information (идентификаторы NC_015416 – хромосома и NC_015430 - плазмида). Работа проводилась в программе Microsoft® Office Excel 2003, версия 11.8404.8405 SP3.

Из записей NC_015416.ptt, NC_015416.rnt и NC_015430.ptt данные были импортированы в файл Excel (1) в листы chromosome, plasmid и rna соответственно, в качестве разделителя использован знак табуляции. Для установления формы распределения длин белков была построена столбчатая гистограмма. Все белки из листов chromosome и plasmid были собраны в лист prots и упорядочены по возрастанию длины в аминокислотных остатках (АО), далее были созданы карманы от 0-50 АО до 3050-3100 АО с шагом в 50 АО. В отдельном листе histo построена гистограмма, показывающая частотную плотность распределения длин белков в протеоме. Также были посчитаны среднее, мода, медиана, минимальное и максимальное значения, их величины представлены в листе prots.

Помимо распределения длин белков были установлены некоторые детали расположения генов белков и РНК в геноме. Сначала было установлено количество генов на прямой и комплементарной цепях ДНК. На листе strands дан перечень генов с указанием их типа (CDS или RNA) и положением на “+” - цепи или “-” - цепи. Далее эти данные сведены в таблицу на листе table, где также при помощи функции БИНОМРАСП было посчитано Р-значение, отражающее вероятность статистической случайности данного исхода при условии, что распределение – биномиальное.

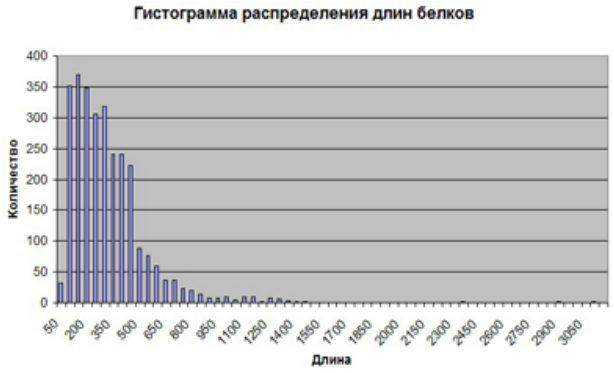
Подсчет количества пересечений генов потребовал импорта данных с использований нескольких разделителей – знака табуляции и точки. Полученная таблица была приведена к удобному виду удалением ненужных столбцов и упорядочиванием по возрастанию номера нуклеотида, которым заканчивается ген, после этого она была сохранена на лист cross. Далее составлен столбец, значение ячейки которого равно единице, если номер последнего нуклеотида соответствующего гена больше номера первого нуклеотида следующего гена. Далее была посчитана сумма единиц внутри этого столбца, которая является общим числом пересечений генов.

* Метаболомика – изучение совокупности метаболических путей организма.

3 РЕЗУЛЬТАТЫ

Таблица 1. Расположение генов на “+” и “-” цепях ДНК

Цепь	Белок	РНК	Общий итог
-	1402	22	1424
+	1448	28	1476
Общий итог	2850	50	2900



На гистограмме отражено распределение с правосторонней асимметрией. Средняя длина белка равна 295,2365, минимальная 37, максимальная 3064, мода 374, медиана 253.

Гены белков и РНК расположены в геноме следующим образом (см. таблицу 1):

генов РНК на “+” - цепи – 28;
генов РНК на “-” - цепи – 22;
генов белков на “+” - цепи – 1448;

генов белков на “-“ - цепи – 1402.
Проверка случайности показала, что такое расположение генов на прямой и комплементарной цепях ДНК с большой вероятностью случайно (P = 0,837487571).
Подсчет пересечений генов дал результат в 336 пересечений.

4 ОБСУЖДЕНИЕ

Полученная гистограмма полностью соответствует ожиданиям: понятно, что длина функционального белка ограничена снизу, поэтому в наименьший карман попало мало полипептидов (32 из 2850). Белки длиной, резко превышающей среднюю, будут встречаться редко, так как огромная полипептидная цепь редко бывает необходимой для выполнения функций. Распределение генов на ДНК тоже не вызывает удивления, так как обе цепи имеют 5’-3’ направленность, ничего не мешает проведению транскрипции как с прямой, так и с обратной цепи, а, значит, расположение генов может быть случайно. Большое количество пересечений генов может быть объяснено тем, что их преобладающее количество приходится на случаи, когда гены расположены на разных цепях. В остальных случаях, видимо, один ген содержит промотор другого гена (но, что наиболее вероятно, в другой рамке считывания).

5 ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ

(1) - Файл Excel proteome_M.concilii_GP6.xls