Bioinformatics

Обзор протеома бактерии Rhodobacter sphaeroides

Samborskaya Margarita

School of bioengineering and bioinformatics, MSU

ABSTRACT

В отчете описывается протеом бактерии *Rhodobacter sphaeriodes* , штамм ATCC 17029.Обзор был выполнен в рамках курса биоинформатики с помощью программы Microsoft Excel 2010.В ходе работы были установленны следующие факты:

1)В геноме наиболее велико количество белков с длиной 400-500 аминокислот.

2)Количество генов белков намного больше, чем количество генов РНК.

3)Многие гены пространственно сгруппированы в квазиопероны.

4)Большое количество генов, кодирующих белки, пересекается.

5)Есть основания предполагать, что гены в геноме распределены не случайным образом.

1 INTRODUCTION

Данный обзор представляет собой одно из заданий курса биоинформатики. Задание необходимо для освоения программы Microsoft Excel и формирования навыков сортировки и анализа данных. Важной задачей данного обзора также является написание структурированного отчета, содержащего все необходимые разделы.

2 METHODS

Исходные данные были получены на сайте ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Rhodobacter_sphaeroides_ATCC_17029_uid58449/

Взятые файлы:
NC_009040.ptt, NC_009040.rpt, AC = NC_009040.1
NC_009049.ptt, NC_009049.rnt, AC = NC_009049.1
NC_009050.ptt, NC_009050.rnt. AC = NC_009050.1
Содержимое данных файлов было взято полностью. Работа проведена с помощью Microsoft Excel 2010.Исходный геном содержал хромосому 1, хромосому 2 и плазмиду. Также все гены разделялись на белок-кодирующие и кодирующие РНК, для них были известны их длина, начало и конец транскрипции, полярность цепи.
Были использованы такие функции как “СЧЁТЕСЛИМ” (для подсчета количества генов на каждой цепочке ДНК) и “БИНОМРАСПР”(для подсчета вероятности случайного распределения генов по цепочкам ДНК), “МАКС” и “МИН”, а также пакетом “Анализ Данных” для построения гистограммы.
Все результаты доступны на странице http://kodomo.fbb.msu.ru/~margarita/term1/Excel/index.html
в файле
http://kodomo.fbb.msu.ru/~margarita/term1/Excel/index.html/ Prac15Samborskaya.xlsx

3 RESULTS

Гистограмма:
Была построена гистограмма длин всех белков, с помощью которой было установлено, что максимально количество белков, имеющих длину 300 - 400 аминокислот(24.35 %, 1006 белков). Гистограмма представлена на рисунке 1.

Рисунок 1 Частотное распределение длин белков у бактерии Rhodobacter sphaeroides (в аминокислотных остатках)

Также стоит отметить, что наблюдается очень быстрый рост в количестве белков в интервале 200-400 , после чего наблюдается стремительное уменьшение числа белков на интервале 400 – 700 п.н..

Подсчет количества генов:

Также было подсчитано количество генов белков и генов РНК на прямой и комплементарной цепочках ДНК. Результаты подсчета представлены на таблице 1.

Стоит отметить, что количество белков намного превосходит количество РНК(количество РНК составляет 1.7 % от общего количества кодирующих последовательностей).

На прямой и комплементарной цепях содержится примерно равное количество белков.

В то же время, число РНК на прямой цепи намного больше, чем на комплементарной.

Таблица 1. Число генов белков и генов РНК у бактерии Rhodobacter sphaeroides

	РНК	Белки
Цепь +	49	2075
Цепь -	23	2056

Распределение генов по цепочкам:

Была проверена гипотеза о том, что гены распределены по цепочкам случайно с вероятностью 0,5.

Было установлено, что вероятность такого события очень мала для РНК. Однако вероятность случайного распределения ДНК достаточно высока. Данные приведены в таблице 2.

Таблица 2. Вероятность наблюдаемого распределения генов бактерии Rhodobacter sphaeroides по цепям ДНК для случайного распределения

Вероятность для РНК	Вероятность для ДНК
0.00147	0.38972

Подсчет квазиоперонов:

В геноме моей бактерии 912 квазиоперонов (считая, что расстояние между генами в квазеопероне < 100 б.п.) Это значит, что многие гены пространственно сближены.

Пересечения генов:

Были составлены статистические данные о пересечениях генов. Большое количество пересечений может быть свидетельством того, что некоторые гены не экспрессируются. Данные представлены в таблице 3.

Таблица 3. Количество пересечений в геномных последовательностях у бактерии Rhodobacter sphaeroides

Тип последовательности	Кол-во пересечений
Хромосома 1 РНК	0
Хромосома 2 РНК	0
Хромосома 1 CDS	696
Хромосома 2 CDS	272
Плазмида CDS	21

Белок-кодирующие последовательности:

Было проверено, делятся ли все длины белок-кодирующих генов на 3.Было установлено, что они все кратны трем. Если бы такие случаи обнаружились, то это скорее всего была бы ошибка секвенирования либо другая потеря данных. Возможно, что длина, не кратная трем, обусловлена явлением рибосомального фреймшифта.

4 DISCUSSION

Большинство белков имеют длину около 400 аминокислот, что является средней длиной белка в бактерии. Практически нет белков длиной меньше 100 аминокислот.

Распределение CDS может быть случайным, поскольку вероятность такого распределения для случайного расположения равна 38.9 %.

Однако, распределение РНК по цепям не случайно (поскольку вероятность такого распределения для равновероятной модели 0.147%.

ACKNOWLEDGEMENTS

Выражаю благодарность Андрею Алексеевскому и Ивану Русинову, которые ведут данный блок практику курса биоинформатики
Funding: School of Bioengineering and Bioinformatics, MSU

REFERENCES

http://www.ncbi.nlm.nih.gov/ - National Center for Biotechnology Information
http://kodomo.fbb.msu.ru/wiki/ - Сервер компьютерного класса ФББ