Практическая биоинформатика

Краткий обзор генома бактерии *Burkholderia sp.* CCGE1001

Поддъяков Иван Дмитриевич[1]

[1]Факультет биоинженерии и биоинформатики МГУ им. М.В. Ломоносова

Резюме

Часто бывает необходимо найти различные общие сведения по геному или протеому организма. В данной работе я представил такие сведения для *Burkholderia sp.* CCGE1001, обработав их с помощью средств Microsoft Excel.

1 ВВЕДЕНИЕ

В данной работе я привел некоторые сведения о геноме и протеоме *Burkholderia sp.* CCGE1001. Подробнее об обработанных и представленных сведениях будет сказано ниже.

Род *Burkholderia* включает в себя более сорока различных видов, которые занимают широкий диапазон экологических ниш. Они встречаются в почве и воде, в растениях, животных, мицелии грибов[1]. Некоторые представители рода Burkholderia являюся паразитами. Так, Burkholderia pseudomallei является возбудителем мелиодоза, способна заражать людей[2]. *Burkholderia sp.* может быть использована для борьбы с насекомыми-вредителями[3].

Рис. 1. Burkholderia cepacia – возбудитель муковисцидоза [4][5].

Burkholderia sp. CCGE1001 - грамотрицательная свободноживущая β-протеобактерия. Относится к мезофилам[6]. Геном Burkholderia sp. CCGE1001 состоит из двух хромосом длинной 4063449 и 2770302 b.p., отсеквенирован 7 февраля 2011 года. Бактерия имеет 6537 ген, из них 6453 кодируют белки[7][8].

2 МЕТОДЫ

Результаты работы были получены с помощью программы Microsoft Office Excel 2010. Данные взяты из базы NCBI для упомянутой бактерии, использовано содержание архива «GCA_000176935.3_ASM17693v3_cds_from_genomic.fna.gz». Данный файл был импортирован в Excel документ и превращен в плоскую таблицу (лист flat_table)[1]. Каждый анализ данных приведен на новом листе с использованием ссылок на первый.

В рамках работы я пользовался такими методами для получения следующих результатов:

- Распределение генов по категориям (CDS, tRNA, rRNA и т.д.) и по хромосомным структурам, а также расчет кол-ва определенных генов на 1 млн пар нуклеотидов: функция СЧЁТЕСЛИМН, арифметические и логические операции – лист general.
- Построение гистограмм длин белков как всех, так и для каждой структуры: функция СЧЁТЕСЛИМН, построение гистограмм и работа с гистограммами – лист histCDS.
- Определение распределения генов по структурам и их цепям: СЧЁТЕСЛИМН, логические операции – листы table_part и table_common.
- Проверка случайного распределения генов по прямой и обратной цепям: функция ХИ2.ТЕСТ, принимающая в качестве аргументов фактическое распределение белков и ожидаемое при равном распределении – лист reliability.
- Подсчет числа квазиоперонов в геноме в целом и, в частности, для каждой структуры и цепи: функции СЧЁЕТЕСЛИМН, ЕСЛИ, МАКС, МИН, СУММ, математические и логические операции, построение гистограмм – лист koperon.
- Подсчет максимальной и минимальной длины белка, средней и медианы для каждой структуры: функции МАКС, МИН, СРЗНАЧ, МЕДИАНА – лист basic.

Для корректного отображния результатов я пользовался форматированием ячеек. Для удобства работы с данными использовал фильтр и закрепление областей.

3 РЕЗУЛЬТАТЫ

В данном разделе пердставлены основные результаты работы, для наглядности приведены таблицы и гистограммы.

3.1. Общие

Таблица 1. Основные статистические данные по протеому *Burkholderia sp.* CCGE1001.

Максимальная длина белка, п.н.	4347,00

[1] Здесь и далее будут указываться листы файла .xlsx, ссылка на который дана в разделе 5.

Минимальная длина белка, п.н.	24,00
Средняя дл. белка на хромосоме 1, п.н.	328,23
Средняя дл. белка на хромосоме 2, п.н.	330,72
Средняя дл. белка на плазмиде, п.н.	232,35
Медиана белка на хромосоме 1, п.н.	293,00
Медиана белка на хромосоме 2, п.н.	299,00
Медиана белка на плазмиде, п.н.	191,00

Таблица 2. Количество генов по категориям.

Gene	Total	Per_1_mil_bp
CDS	6453	890,10
tRNA	62	8,55
misc_RNA	1	0,14
ncRNA	2	0,28
rRNA	18	2,48
tmRNA	1	0,14

В Таблице 1 и Таблице 2 представлены некоторые основные характеристики протеома бактерии. В частности, максимальная и минимальная длина гена равна 4347 п.н. и 24 п.н. соответственно. Медианы длин генов меньше соответствующих средних. Всего генов, кодирующих белки, - 6453, а кодирующих РНК различных типов – 84. На 1 млн п.н. приходится около 900 белковых генов и не более 12 генов РНК

3.2. Распределение белков

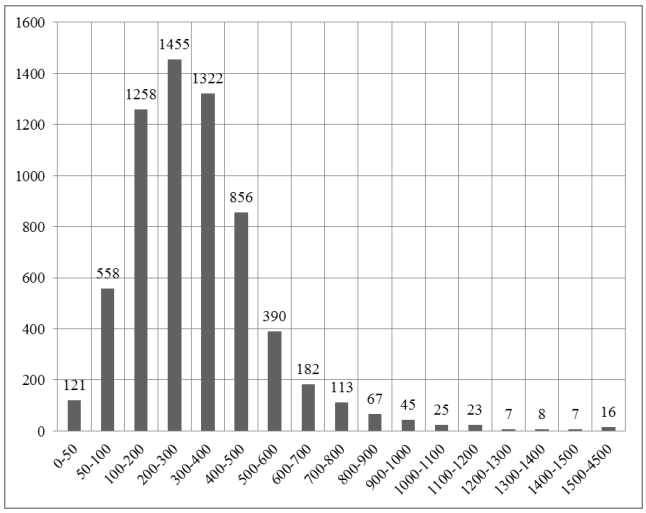

Рис. 2. Распределение всех белков.

На гистограмме Рис. 2 представлено общее распределение белков по длине. Наибольшее количество белков укладывается в интервал длины 200-300 п.н. Распределение белков каждой хромосомной структуры представлено в виде гистограмм на листе histCDS. Для обеих хромомсом также верно, что наибольшее количество белков лежит в интервале 200-300. Однако белки плазмиды распределены иначе, без явно выраженного пика, максимум белков лежит в интервале 50-100.

3.3. Проверка случайного распределения генов по цепям

Таблица 3. Распределение генов по цепям.

Вектор, цепь	CDS	RNA
chr1, +	1670	26
chr1, -	1875	41
chr2, +	1169	16
chr2, -	1251	1
pl, +	260	0
pl, -	228	0

Проверка была выполнена мной с использованием встроенной функции MsOffice Excel ХИ2.ТЕСТ. Функция принимет два диапазона – фактическое распределение белковых генов по цепям каждой структуры и ожидаемое распределение с вероятностью 0,5. За нулевую гипотезу было принято случайное распределение генов по цепям. В результате было получено значение P = 0,000232691, что меньше принятого уровня значимости 0,01. Это означает, верна альтернативная гипотеза: распределение генов по цепям не случайно. Распределение генов по цепям представлено в Таблице 3.

3.4. Подсчет числа квазиоперонов

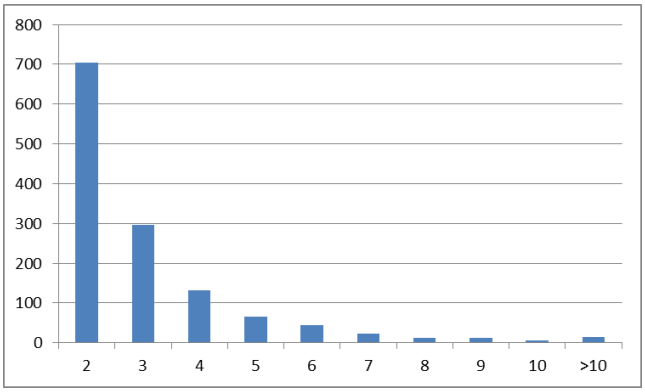

Рис. 3. Распределение квазиоперонов по их длинам.

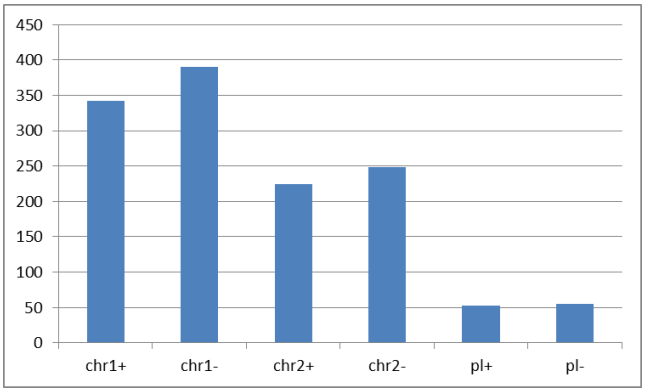

Рис. 4. Распределение квазиоперонов по цепям.

Всего на двух хромосомах и плазмиде обнаружено 1312 квазиоперонов при пороге 100 п.н. На Рис. 3 представлено распределение квазиоперонов по длине. Количество оперонов длиной 2 гена превышает количество всех остальных оперонов. По распределению оперонов по молекулам ДНК и цепям (представлено на Рис. 4) можно увидеть, что количество оперонов на разных цепях хромосом различается.

4 ОБСУЖДЕНИЕ

Геном бактерии состоит из двух хромосом и плазмиды. Отсутствие РНК на плазмиде может быть следствием того, что она была получена горизонтальным переносом генов, так как все необходимые РНК, в частности, – тРНК и рРНК – располагаются на двух хромосомах.

Статистически доказано, что распределение генов по цепям неслучайно. В пользу этого вывода говорит также то, что число квазиоперонов на разных цепях различно. Таким образом, гены, объединяясь в гипотетические опероны, не могут равномерно распределиться по цепям.

Распределение генов бактерии показывает, что разброс в числе генов разных длин велик, а большинство (95%) генов имеет длину до 800 п.н. Также для каждой молекулы ДНК медиана меньше среднего. Из этого следует, что больше половины генов имеют длину меньше, чем среднее значение. Подобное распределение может быть обусловлено тем, что только до определенной длины белки могут эффективно выполнять свои функции в физиологических условиях.

5 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Файл формата .xlsx с результатами.

6 СПИСОК ЛИТЕРАТУРЫ

1. Tom Coenye, Peter Vandamme; Burkholderia: Molecular Microbiology and Genomics; Wymondham Horizon Bioscience 2007
2. Microbiologybytes
3. Марроун Памела,Хуан Хуачжан, Койвунен Марья, Кордова-Крейлос Ана Люсия, Асолкар Ратнакар, Пестицидная композиция, включающая изолированный штамм burkholderia sp., соединения, выделенные из burkholderia sp., их способы получения и применения.
4. SciencePhoto
5. J. R. Govan, V. Deretic, V Deretic, Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia, Microbiol. Mol. Biol. Rev. September 1996 vol. 60 no. 3 539-574
6. Integrated microbal genomes
7. NCBI
8. NCBI