

Краткий обзор генома и протеома бактерии *BREVUNDIMONAS NAEJANGSANENSIS*

Бурнашов Иван

Факультет биоинженерии и биоинформатики, МГУ, Ленинские горы МГУ 1, стр. 74
Москва, Российская федерация
maskazo@fbb.msu.ru

В данном исследовании я проанализировал геном бактерии **Brevundimonas naejangsanensis** Kang et al. 2009. Исследование проводилось с целью структурализации информации из полной геномной последовательности для выявления полезных данных для будущих исследований и практического применения.

Ключевые слова: геном, бактерия, протеом, Brevundimonas naejangsanensis, обзор.

1. Введение:

Brevundimonas относятся к классу протеобактерий. Они грамотрицательные, неферментирующие, аэробные бациллы. Brevundimonas - одна из немногих бактерий, показывающий высокую выживаемость в условиях имитации марсианской поверхности. *Результаты одного из этих экспериментов по экспериментальному облучению в сочетании с предыдущим радиационным моделированием указывают на то, что Brevundimonas sp. MV.7 углублённые только на 30 см. в глубину марсианской пыли могли выдержать космическое излучение в течение до 100 000 лет, прежде чем пострадать от сокращения численности населения, превышающей 10^6 раз

Геном **Brevundimonas naejangsanensis** состоит из 2971535 пар нуклеотидов и включает в себя 2828 генов, что составляет 951,69 генов на 10^6 пар нуклеотидов

* <http://online.liebertpub.com/doi/abs/10.1089/ast.2009.0439>

2. Методы:

Brevundimonas naejangsanensis strain B1 chromosome, полный геном хранится на *серверах NCBI

Excel использовался в качестве базового программного обеспечения. Конкретные методы, которые для анализа генома и протеома будут наблюдаться в следующих подпунктах.

*ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/635/915/GCA_000635915.2_ASM63591v2/

3.1 Геном:

Для расчета количества кодирующих белок и РНК генов и псевдогенов на обоих цепях я сравнил информацию из столбцов '#feature' и 'class'. Были рассмотрены только те строки, которые содержали 'gene' в столбце '#feature'

Я определял ген как:

- (1) кодирующий белок, если он имел "protein_coding" в столбце "class"
- (2) псевдоген, если он содержал "pseudogene";
- (3) РНК кодирования если там были "ncRNA", "tRNA", "tmRNA", "SRP_RNA", "rRNA", "RNase_P_RNA" в столбце "class"

Тот же метод был использован для идентификации числа последовательностей ДНК, кодирующих различные типы продукции:

Если строка содержит "cds" в столбце "#feature", в CDS были определены разные типы кодирования:

- (1) транспортный белок, если он содержит "transporter" или "transport protein" и не содержит "binding" в столбце "name"
- (2) рибосомный белок, если в нем содержатся "ribosomal" и не содержат "transferase" в столбце "name"
- (3) гипотетический белок, если он содержал "hypothetical" в колонке "name" и "with_protein" в столбце "class"
- (4) другой белок, если он не соответствовал каким-либо перечисленным требованиям (количество = все cds – (1) – (2) – (3))

Для генов кодирующих РНК та же схема индексов в столбцах, что и для таблицы 1, использовано: "gene" в столбце "#feature" и тип РНК соответственно в столбце "class"

3.2 Протеом:

Для построения гистограммы длины белка я использовал значения из столбца product_length в изначальном файле refseq. Гистограмма была построена в Excel. Статистические метки длины белка (минимальные, максимальная длина и др.), которые представлены в таблице 3, были определены с помощью Excel

4. Итоги:**4.1 Геном**

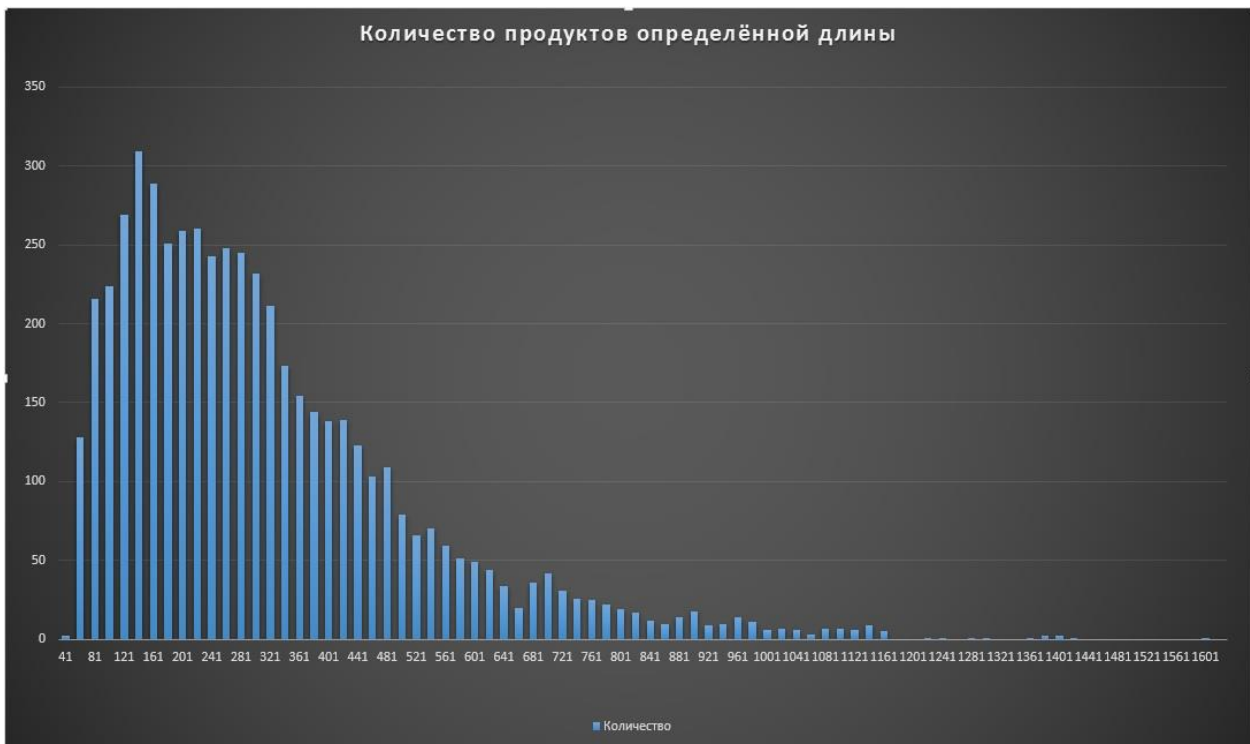
Одной из целей исследования было проанализировать распределение генов по цепям. Таблица ниже (Таблица 1) показывает количество генов на цепи ДНК.

Chain	protein coding	preudogenes	RNA coding
+ chain	1393 (50,58%)	6 (40%)	25 (42,37%)
- chain	1361 (49, 42%)	9 (60%)	34 (57,63%)

Процент от общего количества генов в классе генов указывается в скобках. Видно, что в генах, кодирующих белки, отсутствует большое отклонение от случайного (50%) распределения кодирования генов группы. В других группах отклонение значительное, но они содержат меньше, чем сто образцов, поэтому оценка распределения случайности не оправдана.

Из этой таблицы также видно, что группа генов белкового кодирования намного больше РНК кодирующих. Это естественно, так как белки более разнообразны в клетке

На диаграмме ниже (рис.1) иллюстрирует распределение длины белка. Каждый столбец соответствует количеству белков соответствующей длины (каждый зазор равен 20 п.н.). Понятно из гистограммы, есть один пик из наиболее распространенных длина белка (141-161 пар нуклеотидов), и частота уменьшается постепенно. XLSX-файл с гистограммой доступен в дополнительных материалах.



5. Заключение

Как и ожидалось, большого разброса по цепям в генах, кодирующих белки, не наблюдается (отклонение не превышает 0,6%). Но для более мелких групп, таких как псевдогены, наблюдается существенное отклонение от предполагаемого результата (50%). В отношении ген – псевдоген не было замечено ничего необычного.

6. Благодарности:

Отдельная благодарность преподавательскому составу факультета биоинженерии и биоинформатики МГУ им.М.В. Ломоносова.

7. Дополнительные материалы:

Excel файл: http://kodomo.fbb.msu.ru/~maskazo/term1/GCA_000635915.2.xlsx

8. Используемые материалы:

- (1) ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/635/915/GCA_000635915.2_ASM63591v2/