Мини-обзор генома и протеома *Thermus parvatiensis*

Адамович Арина

Факультет Биоинженерии и биоинформатики МГУ, 1 курс

РЕЗЮМЕ

Данный мини-обзор является отчётом о работе с геномом и протеомом *Thermus parvatiensis* с использованием программирования на языке Python и функционала электронных таблиц.

Ключевые слова

Thermus parvatiensis, геном, протеом

1 ВВЕДЕНИЕ

Thermus parvatiensis -- экстремальный термофил. Данный род, в том числе и конкретная бактерия, обитают при аномально высоких температурах, чем вызывают внимание биотехнологов. Ранее у других представителей этого рода были выделены термостабильные ДНК-полимеразы (Tripathi et al.,2017).

Впервые *T. parvatiensis* была выделена в 2015 году из пробы, взятой в горячих источниках близ Маникаран, Индия. Диапазон жизнедеятельности описывается температурой в 60-80 градусов по Цельсию, pH среды 7-9 (нейтральный/слабощелочной). Другими особенностями данного вида является невозможность сбраживать глюкозу. Далее стоит отметить, что *T. parvatiensis* неустойчива ни к одному из антибиотиков (Dwivedi et al., 2015).

Сама грамотрицательная бактерия имеет геном 1,8Mb (Tripathi et al.,2017). Содержание G+C пар в геноме: 68,7% (Dwivedi et al., 2015).

Домен: Бактерии (Bacteria)
Филлиум: Deinococcus-Thermus
Класс: Deinococci
Порядок: Thermales
Семейсво:Thermacare
Род: *Thermus*
(thermus parvatiensis in Taxonomy (uniprot.org))

В данной работе представлены результаты анализа генома и протеома *Thermus parvatiensis.*

2 МАТЕРИАЛЫ И МЕТОДЫ

Результаты данной работы основаны на даных о геноме и протеоме из базы NCBI Genome(4).

Для обработки данных была использована программа Google Sheets. Для составления сводных таблиц по конкретным темам на основе общей таблице по геному использовались ряд методов электронных таблиц, а именно функции: COUNTIFS, AVERAGE, STDEV, MEDIAN, MIN, MAX; а также другие методы: специальная вставка, распространение формул, вставка и форматирование гистограммы, сортировка и форматирование ячеек.

Для подсчёта нуклеотидов и их частот, частот кодонов и аминокислот были написаны соответствующие скрипты на Python. Также был использован сервис Webskew(5) для построения графика GC-Skew cumulative.

3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Размер и состав генома

В данном разделе представлена общая информация о геноме *Thermus parvatiensis.*

Геном данной бактерии представлен одной хромосомой и одной плазмидой. В хромосоме содержится 1872821 пар оснований, а в плазмиде 143277. В таблице 1 приведены численные значения встречаемости нуклеотида в хромосоме и плазмиде. Полученные данные в очередной раз подтверждают правило Чаргаффа: количество A и T сопостовимо и количество C и G .

Содержание G+C пар в хромосоме 68,54%, а плазмиде 68,41%,, что примерно соответствует литературным данным (Dwivedi et al., 2015). Данное содержание GC пар довольно высокое, что часто наблюдается у термофильных бактерий, видимо это приспособление к столь экстремальным условиям жизни, ведь G+C пары более устойчивы за счет образования тройной связи.

Таблица 1. Нуклеотидный состав

нуклеотид	A	T	C	G
количество количество нуклеотидов в хромосоме	293627	295528	643168	640498
частота встречаемости	0,157	0,158	0,343	0,342
количество нуклеотидов в плазмиде	22432	22832	48440	51110
частота встречаемости	0,155	0,157	0,334	0,353

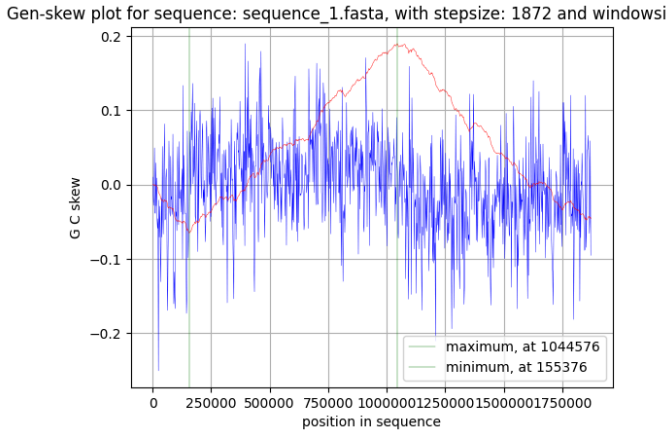
© ФББ 2021

3.2 Нахождение координаты начала репликации и координаты терминации репликации.

Расчёт GC-Skew совершается по формуле: GCSkew = (G − C)/(G + C), где G и C соответственные количества гуанинов и цитозинов в окне заданной ширины.

Максимальное значение соответствует участку терминации репликации, а минимальное - ориджину. Гипотетически, ориджин находится примерно на 155000 нуклеотиде, а терминатор - на 1050000.

Рисунок 1. GC-Skew cumulative



3.3 Длина белков

Исходя из рисунка 2, наибольшее количество белков протеома исследуемой бактерии (1552) имеет длину от 90 до 180 аминокислот, далее при увеличении количества аминокислот количество белков, лежащих в соответствующем диапазоне длин постепенно снижается . Кроме того, большое количество белков имеют длину, не превышающую 120 аминокислот. Минимальная длина белка составляет 27 аминокислот, максимальная - 1524 аминокислот, эти и некоторые другие данные о белках представлены в таблице 3.

Рисунок 2. Гистограмма длин белков

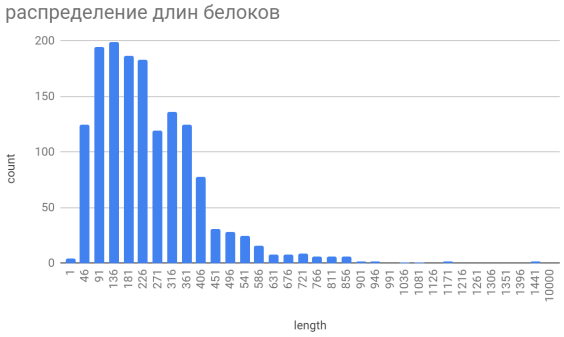


Таблица 3. Длины белков протеома *T. parvatiensis*

средняя длина	269
стандартное отклонение	167,8842834
медиана	234
минимальное значение	27
максимальное значение	1 524

3.4 Распределение генов

В Таблице 4 представлено количество генов белков, псевдогенов и генов РНК на “+”- и “-”-цепях ДНК. Также была рассчитана вероятность случайно получить такое же различие или больше в распределении генов/псевдогенов по цепям.

Исходя из полученных результатов, можно сказать что различия в нахождение генов белков, а также псевдогенов, на “+” цепи будет являться статистически достоверным. Про гены РНК нельзя сделать такой вывод, но с другой стороны таких генов сильно меньше, чем обсуждаемых ранее.

Таблица 4. Распределение генов белков, РНК и псевдогенов по и “+”- и “-”-цепям ДНК у *T. parvatiensis*

	"+" цепь	"-" цепь	вероятность
гены белков	850	701	0,000169472
псевдогены	322	257	0,007765840
гены рнк	29	26	0,787706189

3.5 Частоты использования кодонов

В таблице 5 (а также вспомогательная таблица, лист кодоны) представлены количества вхождений того или иного кодона в геном. Самыми частыми оказались GGG (34740) и CCC (34428), которые кодируют глицин и пролин соответственно. Следующий после них по частоте это аланин (GCC, 30185).

Данные результаты неплохо согласуются с действительностью, так как глицин самая популярная аминокислота, пролин же обладает исключительной конформационной жесткостью, которая обеспечивает повышение устойчивости белка(6) и, наконец, аланин имеет несложное строение и высокую температуру плавления(около 300°C). Для экстремального термофила важна высокая стойкость белков, которая отчасти обеспечивается и стойкостью аминокислот.

Самыми редкими кодонами у *T. parvatiensis* являются аспарагин, изолейцин и тирозин. Редкие встречи тирозина, на мой взгляд, поддаются объяснению. У тирозина в радикале содержится ароматический участок, который вероятно сложно и энергозатратно синтезировать.

Таблица 5. Количество вхождений различных кодонов (фрагмент)

кодон	количество вхождений	аминокислота
GGG	34740	глицин
CCC	34428	пролин
GCC	30185	аланин
TAT	904	тирозин
ATT	779	изолейцин
AAT	770	аспарагин

Также были проанализированы стоп-кодоны (Таблица 6). Ниже приведены количества вхождений для каждого стоп-кодона, а также его доля среди всех стоп-годонов.

Таблица 6. Встречаемость стоп-кодонов в кодирующих белок последовательностях бактерии.

кодон	встречаемость	%
TGA	5487	49,80484705
TAA	1657	15,04039212
TAG	3873	35,15476082

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Ссылка на гугл-папку со всеми сопроводительными материалами: https://drive.google.com/drive/folders/1mvsltaDytAMtAskd9VzbJZB76qFyKgLs?usp=sharing

Ссылка на все скрипты Python, используемые при получение данных: https://colab.research.google.com/drive/1PUyHLPcqEzvMOSSST85b4eHEYNoudpo-?usp=sharing

Ссылка на вспомогательную таблицу, используемую для получения данных методом электронных таблиц: https://docs.google.com/spreadsheets/d/18p8yq_yvlr5-mgwkIFLS3OfJYg4jExdFU76ZGHGK0g0/edit?usp=sharing

БЛАГОДАРНОСТИ

Мы благодарим преподавателей биоинформатики ФББ МГУ, которыми был составлен курс информатики, в рамках которого был написан данный обзор.

СПИСОК ЛИТЕРАТУРЫ

1. Dwivedi, Vatsala; Kumari, Kirti; Gupta, Sanjay Kumar; Kumari, Rekha; Tripathi, Charu; Lata, Pushp; Niharika, Neha; Singh, Amit Kumar; Kumar, Roshan; Nigam, Aeshna; Garg, Nidhi; Lal, Rup (2015). *Thermus parvatiensisRLTsp. nov., Isolated from a Hot Water Spring, Located Atop the Himalayan Ranges at Manikaran, India. Indian Journal of Microbiology, 55(4), 357–365.* doi:10.1007/s12088-015-0538-4

2. Tripathi, Charu; Mishra, Harshita; Khurana, Himani; Dwivedi, Vatsala; Kamra, Komal; Negi, Ram K.; Lal, Rup (2017). *Complete Genome Analysis of Thermus parvatiensis and Comparative Genomics of Thermus spp. Provide Insights into Genetic Variability and Evolution of Natural Competence as Strategic Survival Attributes. Frontiers in Microbiology, 8(), 1410–.* doi:10.3389/fmicb.2017.01410

3. thermus parvatiensis in Taxonomy (uniprot.org)

4. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/535/545/GCF_001535545.1_ASM153554v1/
5. https://genskew.csb.univie.ac.at/webskew
6. https://en.wikipedia.org/wiki/Proline#Properties_in_protein_structure