

Мини-обзор генома и протеома *Thermogutta terrifontis*

Коростина Мария

ФББ, 1 курс

КЛЮЧЕВЫЕ СЛОВА

Thermogutta terrifontis, бактерия, геном, протеом

1 ВВЕДЕНИЕ

Целью данного обзора является изучение генома и протеома *Thermogutta terrifontis*. Это граммотрицательная термофильная бактерия, представитель филума Planctomycetes.

Впервые была выделена из микробного мата в наземном горячем источнике на одном из Курильских островов Кунашир (штамм R1T). Растет в среднем при температуре 55-60 градусов по Цельсию при pH 6,0-6,5.¹

Способна использовать анаэробное дыхание, в частности нитратредукцию. Конечными продуктами ферментации глюкозы являются водород, ацетат, лактат.¹

Было показано, что бактерия может расти на ксантановой камеди - полисахариде, который синтезируется *Xanthomonas campestris*.¹ Ксантановая медь широко используется в промышленности (пищевой, косметической, фармацевтической).

Резистентна к пенициллину и ванкомицину, но чувствительна к стрептомицину.

Классификация:

- › Bacteria
- › PVC group
- › Planctomycetes
- › Planctomycetia
- › Pirellulales
- › Thermoguttaceae
- › *Thermogutta*²

2 МЕТОДЫ

Информация о геноме и протеоме бактерии получена с использованием базы данных NCBI.³

Для поиска участков начала и конца репликации были использованы возможности сервиса GenSkew.⁴

Программы [gc](#), [percentage acgt](#), [gcskew](#) были написаны на языке Python 3.

Программа Linux: [b-dna](#) (доступны по ссылке 2 в Сопроводительных материалах).

Для построения гистограммы, нахождения длин конкретных белков были использованы функции Google Таблицы (см. ссылку 1 в Сопроводительных материалах).

3 РЕЗУЛЬТАТЫ

3.1 Общие данные о геноме

Бактерия *Thermogutta terrifontis* содержит всего одну кольцевую цепь ДНК. Длина кольцевой хромосомы: 4810751 bp. Содержание G+C пар в геноме: 57,34% (согласно данным статьи [1], также проверено с помощью [gc](#)).

Частота GC пар выше, чем высота AT пар. Скорее всего, это является приспособлением бактерии к выживанию в экстремальных условиях. Между гуанином и цитозином образуются три водородные связи, тогда как между аденином и тимином две. Водородные связи укрепляют связь между двумя цепями ДНК, что при повышении температуры помогает бактерии сохранять в целостности генетический материал.

3.2 Нуклеотидный состав

При помощи [percentage acgt](#) был выявлен нуклеотидный состав и частота встречающихся нуклеотидов. В геноме нет никаких других нуклеотидов, кроме A, T, C, G. Вывод программы отображен в таблице 1.

Частота комплементарных оснований приблизительно одинакова, соответственно правила Чаргаффа выполняются. Первое: количество аденина приблизительно равно количеству тимина, а количество гуанина - количеству цитозина. Соответственно выполняется и второе: количество пуринов (A, G) равно количеству пиримидинов (T, C).

Таблица 1. Частота нуклеотидов

нуклеотид	количество	частота
A	1024594	0.2130
C	1360881	0.2829
G	1397591	0.2905
T	1027685	0.2136

3.3 GC-Skew

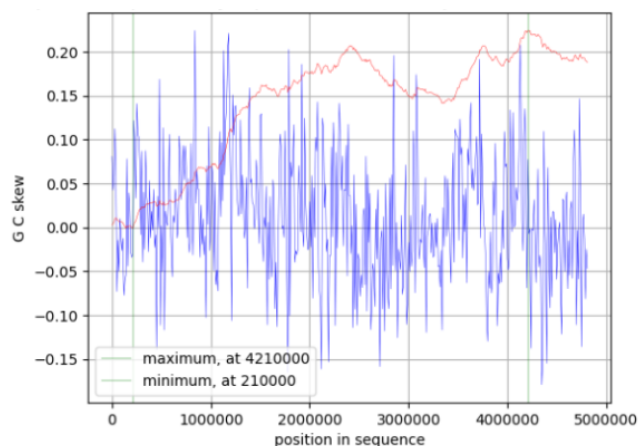
Используя `gcskew` и функции ЭТ, был построен график, который показывает закономерность в значении cumulative GC-skew. Он присутствует по ссылке 1 (см. Сопроводительные материалы, лист “gc skew”). Здесь же показан график, построенный с помощью сервиса GenSkew (рисунок 1). Формула, по которой рассчитывается GC-skew:

$$GCSkew = (G - C)/(G + C)$$

GC-Skew cumulative - сумма всех предыдущих и настоящего GC-Skew. Этот метод позволяет найти точки ориджина и терминации на кольцевой ДНК, исходя из факта, что у некоторых бактерий число G на лидирующей цепи больше числа C, а на запаздывающей - наоборот.

Максимальное значение соответствует участку терминации репликации, а минимальное - ориджину. Гипотетически, ориджин находится примерно на 210000 нуклеотиде, а терминатор - на 4210000. Однако можно заметить, что график очень неровный из-за локальных скачков значений GC-skew, поэтому можно сделать вывод, что для бактерии *Thermogutta terrifontis* нельзя точно определить местонахождение начала и конца последовательности (соотношение числа гуанинов к цитозинам не закономерно на лидирующей и запаздывающих цепях).

Рисунок 1. GC-Skew cumulative



3.4 Гистограмма длин белков

Представленное на гистограмме (рисунок 2) соотношение отражает частоту встречаемости белков различной длины в протеоме *Thermogutta terrifontis* (таблицу и гистограмму также можно найти в Приложении 1, лист “гистограмма”, данные брались из таблицы в листе “genome table”).

Результаты показывают, что длины большей части белков лежат в пределах 300-350 аминокислотных остатков.

Пик гистограммы соответствует длине белка 350. В этом промежутке находятся различные киназы, гидролазы, трансферазы, оксигеназы, в том числе белки, участвующие в ЭТЦ.

Белки с минимальными длинами: белки рибосомы субъединицы 50S (37, 56, 64 а.о.), субъединицы 30S (61, 66 а.о.), ингибитор ДНК-гиразы, а также множество гипотетических белков (таблица 2).

Самый большой белок - сериновая пептидаза, катализирующая протеолиз белков (6758 а.о.). Активный центр содержит серин (таблица 3).

Статистические данные о протеоме находятся в таблице 4 (расчеты по ссылке 1 Сопроводительных материалов, лист “статистика”).

Таблица 2. Белки с минимальной длиной

название	длина
50S ribosomal protein L36	37
DUF1156 domain-containing protein	46
50S ribosomal protein L33	56
4Fe-4S binding protein	57
type Z 30S ribosomal protein S14	61
Flp family type IVb pilin	62
50S ribosomal protein L35	64
DNA gyrase inhibitor YacG	64
30S ribosomal protein S21	66
DUF6485 family protein	67
twin-arginine translocase TatA/TatE family subunit	67

Рисунок 2. Гистограмма длин белков

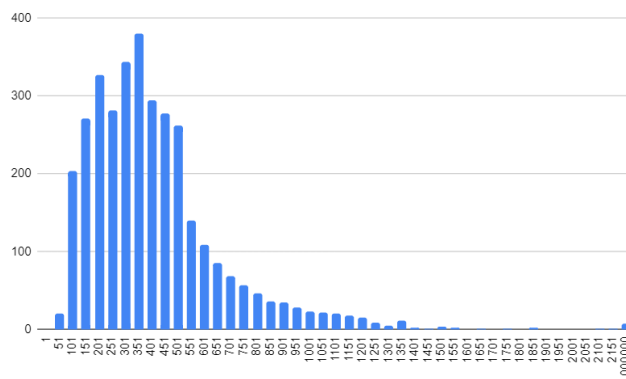


Таблица 3. Белки с максимальной длиной

название	длина
S8 family serine peptidase	37
tetratricopeptide repeat protein	46
excinuclease ABC subunit UvrA	56
DUF2380 domain-containing protein	57

Таблица 4. Статистика данных о протеоме

показатель	значение
средняя длина	391,1
медиана	336
стандартное отклонение	294,4
min	37
max	6758

3.5 Распределение различных последовательностей по + и - цепям ДНК

С помощью **b-dna** было найдено распределение последовательностей по + и - цепям ДНК. Результаты показаны в таблице 5.

На + цепи среди CDS оказались 2 белок-некодирующие последовательности, среди генов - 2 псевдогена, 3 рРНК, 1 тмРНК (участвует в высвобождении рибосом, если они “застряли” на мРНК в ходе трансляции и устраняет дефектный белок), 25 тРНК.

На - цепи среди CDS есть 5 белок-некодирующих последовательностей, а среди генов - 5 псевдогенов, 1 SRP_RNA (signal recognition particle, участвует в котрансляционном транспорте белков), 22 тРНК, RNase_P_RNA (РНК-аза Р, участвует в расщеплении РНК).

Таблица 5. Распределение генов белков по цепям ДНК

название	+	-
CDS	1768	1594
gene	1796	1617
tRNA	25	22
rRNA	3	0
tmRNA	1	0
ncRNA	0	2

Вероятность случайного распределения равна 0,0024 (ссылка 1 в Сопроводительных материалах, лист “вероятность случайного распределения”), соответственно, различия статистически значимы.

4 БЛАГОДАРНОСТЬ

Благодарю преподавательский состав по практической информатике за помощь в усвоении методов, которые были применены в процессе написания обзора.

5 СПИСОК ИСТОЧНИКОВ

1 - Slobodkina, G. B., Kovaleva, O. L., Miroshnichenko, M. L., Slobodkin, A. I., Kolganova, T. V., et al. (2014) *Thermogutta terrifontis* gen. nov., sp. nov. and *Thermogutta hypogea* sp. nov., thermophilic anaerobic representatives of the phylum Planctomycetes, *Int. J. System. Evol. Microbiol.*, 65, 760-765, <https://doi.org/10.1099/ij.0.000009>.

2 - <https://www.uniprot.org/taxonomy/1331910>

3 - NCBI: <https://www.ncbi.nlm.nih.gov/nucleotide/CP018477.1>

4 - GenSkew: <https://genskew.csb.univie.ac.at/>

6 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

1. Ссылка на файл в Google Таблицы: <https://docs.google.com/spreadsheets/d/1duVwZRw8xph0nYs0QjOtw2GnoL92xPI6rFEEA8nOYV0/edit?usp=sharing>

2. Ссылка на программы: https://colab.research.google.com/drive/1K51yOmD2ga7qMYJPx-vnLWjL2exoOzN0#scrollTo=8ehZm6JB_6ol