

# Обзор протеома археи *Pyrobaculum oguniense* TE7

Бойко Александра

Факультет биоинженерии и биоинформатики МГУ им. М.В.Ломоносова, 1 курс

## РЕЗЮМЕ

Данная статья посвящена исследованию протеома археи *Pyrobaculum oguniense*, штамм TE7, а конкретнее – изучению расположения генов на разных цепях ДНК, а также распределения длин белков этого организма. Была предложена и проверена гипотеза о случайности распределения генов по цепочкам, а также была выявлена самая частотная длина белков у этого организма. В ходе исследования использовалась программа Microsoft Excel 2007.

## 1 ВВЕДЕНИЕ

Архея *Pyrobaculum oguniense* TE7 относится к классу Thermoprotei. Как и большинство представителей рода *Pyrobaculum*, это термофильный гетеротрофный организм, для которого оптимальными условиями является температура 90-94°C и pH 7.0-7.5. Эти археи могут обитать как в аэробной, так и в анаэробной среде. Обнаружены представители данного вида были в горячих источниках Японии<sup>[2]</sup>

Геном данного штамма полностью секвенирован, он состоит из одной кольцевой хромосомы и одной плазмиды. Всего в геноме закодировано 2972 белков, из них 2835 генов кодируют белки, а остальные 137 – РНК.

В ходе работы было изучено распределение длин белков *Pyrobaculum oguniense* TE7, а также было подсчитано количество белков на прямой и обратной цепях обеих молекул ДНК. Была высказана гипотеза о том, что гены распределены по цепочкам случайно с вероятностью 0,5.

Основная цель работы заключалась в исследовании протеома археи и приобретении навыков работы в программе Microsoft Excel.

## 2 МАТЕРИАЛЫ И МЕТОДЫ

В этом исследовании использовались материалы секвенирования генома археи *Pyrobaculum oguniense* TE7, найденные в базе данных NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/>) по идентификаторам NC\_016885 и NC\_016886 для хромосомы и плазмиды соответственно.

Данные из полученных файлов были импортированы в Excel. Лишняя информация была удалена, все белки организма собраны на отдельном листе (bar chart) и отсортированы по

длине (в аминокислотных остатках). Затем была построена гистограмма распределения длины белков в протеоме с карманами в 50 а.о., для этого была подключена надстройка «Пакегет анализа».

Для подсчета количества генов белков и РНК на «-» и «+» цепях использовалась функция СЧЁТЕСЛИ(COUNTIF), кроме того, была задействована функция БИНОМРАСП для проверки гипотезы о случайности распределения генов по цепям (таблица на листе strands). Эта функция позволяет получить P – число, отражающее вероятность статистической случайности успешного исхода при допущении, что распределение биномиальное.

## 3 РЕЗУЛЬТАТЫ

### 3.1 Распределение длин белков

Распределение длин белков археи *Pyrobaculum oguniense* показано на гистограмме, приведенной ниже (рис. 1). При построении графика были объединены данные по белкам хромосомы и плазмиды организма. По рисунку можно увидеть, что большинство белков состоит из 50 – 500 аминокислот. Средняя длина белка равна ~ 253,4 а.о.; самый маленький белок состоит из 29 а.о., самый большой – из 4372 а.о. Больше всего белков по длине попадают в диапазон 100 – 150 а.о., их 464 штуки.



**Рисунок 1.** Распределение длин белков в протеоме археи *Pyrobaculum oguniense* TE7. На горизонтальной оси указаны длины белков (в аминокислотных остатках), на вертикальной – количество белков с такой длиной.

### 3.2 Распределение генов по цепям

Результаты по количеству генов на разных цепях ДНК представлены в таблице 1. Информация приведена для всех генов (и тех, которые находятся в хромосоме, и тех, которые находятся в плазмиде).

**Таблица 1.** Распределение генов, кодирующих белки и РНК, по цепям ДНК.

Цепь	Белки	РНК	Общий итог
Прямая цепь	1443	72	1515
Обратная цепь	1392	65	1457
<b>Всего</b>	<b>2835</b>	<b>137</b>	<b>2972</b>

Значения для количества генов на прямой и обратной цепочках, приведенные в таблице, близки между собой, то есть гены распределены примерно поровну. Подсчитанная вероятность статистической случайности равна  $\sim 0,86$ .

## 4 ОБСУЖДЕНИЕ

В диапазон длины 50 – 500 а.о. попало 89% всех белков организма, и это неудивительно, так как эта длина является наиболее оптимальной для существования и функционирования белка. Слишком короткие белки не были бы так высоко специализированы, к примеру, как им нужно быть для того, чтобы нормально функционировать, а длинные полипептидные цепи имели бы громоздкие четвертичные структуры, из-за чего также не имели бы необходимых свойств.

Биномиальное распределение – это распределение вероятностей числа «попаданий» генов на прямую или комплементарную цепь, если в каждом случае вероятность «попадания» гена на ту или иную цепь 0,5 (а это подтверждено результатами, приведенными в таблице). В ситуации с *Pyrobaculum oguniense TE7* его значение велико, значит, можно сделать вывод, что распределение генов по цепочкам ДНК случайно с большой вероятностью (0,86). Гипотеза подтвердилась.

## СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

(1)[http://kodom.fbb.msu.ru/~boyko.s/term1/Pyrobaculum\\_proteom.xls](http://kodom.fbb.msu.ru/~boyko.s/term1/Pyrobaculum_proteom.xls) - Файл Excel с исходными данными, гистограммой и таблицей распределения генов по цепям.

## БЛАГОДАРНОСТИ

Хочу выразить свою благодарность Андрею Владимировичу Алексеевскому и Ивану Русинову за обучение основам использования программы Microsoft Excel.

## СПИСОК ЛИТЕРАТУРЫ И ССЫЛКИ

[1] NCBI PubMed Bernick DL, Karplus K, Lui LM, Coker JK, Murphy JN, Chan PP, Cozen AE, Lowe TM. Stand Genomic Sci. 2012 Jul 30;6(3):336-45. <http://www.ncbi.nlm.nih.gov/pubmed/23407329>