

Изучение особенностей археи *Halanaeroarchaeum sulfurireducens* на основе данных генома

Соколов Сергей Ильич

Факультет биоинженерии и биоинформатики, Московский государственный университет им. М.В. Ломоносова, Ленинские горы, д. 1, стр. 73, 119234, Москва, Россия

Резюме

Предпосылки: исследования гиперсолёных озёр в Кулундской степи (Алтай, Россия) 2009-2013 годов привели к открытию нового вида галоархей с ранее несвойственным царству архей метаболизмом [1]. В данном мини-обзоре я буду проводить анализ некоторых данных о геноме этого организма.

1 ВВЕДЕНИЕ

Вид является единственным описанным в своём роде, представляет из себя анаэробный организм, неподвижный, плоский коккоид или дощевидную палочку (рис. 1). Метаболизм чрезвычайно узкий: в качестве донора электронов/источника углерода используются ацетат и пируват, а в качестве акцептора-элементарная сера; в качестве источника азота используется аммоний. Архея была обнаружена в рассолах и аэробных сульфидных поверхностных отложениях (2-10 см), полученных из гиперсолёных хлоридно-сульфатных озёр. Оптимальный рост происходил в присутствии 4М (Моль/л) NaCl и в диапазоне от 3 до 5 М при температуре 37-40 °C. Удивительно то, что в одной из культур (штамм HSR2) наблюдались такие восстановленные соединения серы, как CS_2 , CH_3SH , которые никогда не встречались у известных серовосстанавливающих прокариот [1].

Таксономическое положение археи:

Домен: *Archaea*
Царство: *Euryarchaeota*
Класс: *Halobacteria*
Отдел: *Halobacteriales*
Семейство: *Halobacteriaceae*
Род: *Halanaeroarchaeum*
Вид: *Halanaeroarchaeum sulfurireducens* [1]

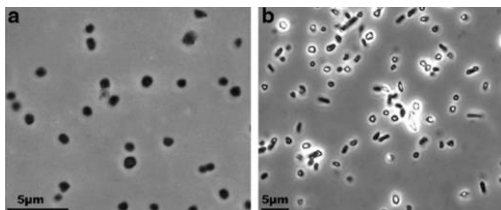


Рис. 1. Изображение *Halanaeroarchaeum sulfurireducens*, выращенных в ацетате (a) и пирувате (b) в качестве доноров электронов и элементарной сере в качестве акцептора электронов [1]

2 МЕТОДЫ

При изучении археи использовались нижеперечисленные платформы и их средства:

1. Банк генов:
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/305/655/GCF_001305655.1_ASM130565v1/, с которого были получены таблицы генетических особенностей, последовательностей и участков CDS, а именно файлы:

GCF_001305655.1_ASM130565v1_feature_table.txt,
GCF_001305655.1_ASM130565v1_genomic.fna,
GCF_001305655.1_ASM130565v1_cds_from_genomic.fna.

2. Интерпретатор команд Bash, в котором при помощи команды `infseq *_cds_from_genomic.fna -nocolumns -out cds.csv`, я создал файл с кодирующими последовательностями, а затем через подключение к серверу по протоколу `sftp` скачал его и таблицу генетических особенностей, полученную из банка генов: GCF_001305655.1_ASM130565v1_feature_table.txt. Все вышеперечисленное было импортировано в Google таблицы.

3. Google таблицы (Excel) (см. сопроводительные материалы, п.1), где и проводился анализ генома с использованием формул и функций (последние я привел ниже для каждого задания):

3.1: MAX, MIN, COUNTIFS

3.2: COUNTA

3.3: COUNTIF

3.4: MAX, MIN, COUNTIF, COUNTIFS

3.5: COUNTIFS

3.6: -

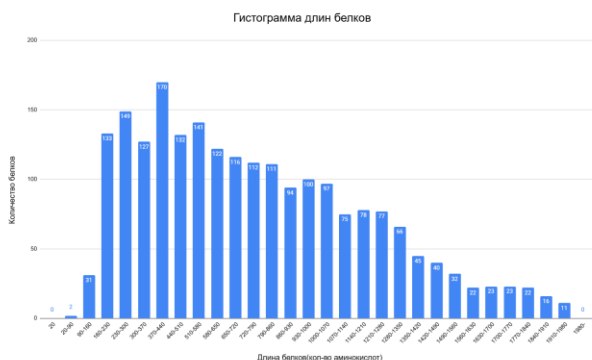
3.7: MAX, MIN, COUNTIFS

3 РЕЗУЛЬТАТЫ

3.1 Гистограмма длин белков, закодированных в геноме археи *Halanaerobacterium sulfurireducens*

Из гистограммы можно сделать некоторые выводы о белках археи (рис. 2):

1. Степень полимеризации молекулы. Чем большим количеством аминокислот закодирован белок, тем, вероятнее, сложнее его пространственная структура. Пространственная же структура определяет функции белка, количественно или качественно [2].
2. На гистограмме отсутствуют пробелы, однако наблюдаются пики, т.е. определённые белки с длинами от 160 до 860 являются наиболее функционально значимыми для археи. Белки с большей длиной встречаются реже, и это естественно: их сложнее синтезировать.
3. Наиболее распространены белки длиной 370-410 аминокислот – 170 единиц. Белки с длиной менее 20 и более 1980 аминокислот не встречаются. В среднем белки эукариотических организмов длиннее бактериальных, а последние, в свою очередь, длиннее архейских. Этот факт можно интерпретировать как эволюционную тенденцию к увеличению размера белка, которая свидетельствует о слиянии однофункциональных белков в расширенные многофункциональные [2].



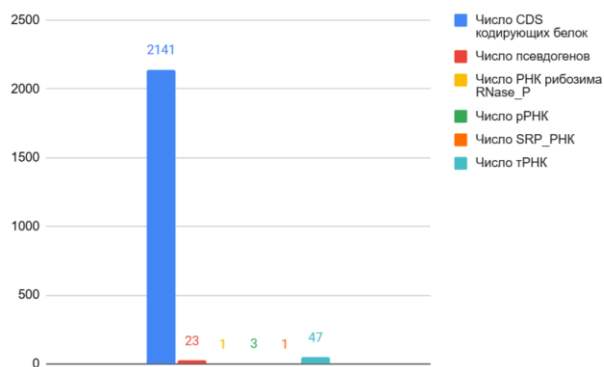


Рис 3. Гистограмма количества генов и их функций на наибольшем репликоне

Наибольший репликон (рис. 3) содержит гены со всеми функциональными единицами. Т.к. роль белков в организмах невероятно обширна, генов, их кодирующих, преобладающее большинство. Гены, несущие псевдогены, более не функциональны и не представляют особого интереса, их доля составляет 0,01, т.е. 1%. Отдельно стоит обсудить гены, несущие различные виды РНК, а именно сами РНК:

1. RNase_P_RNA: данная РНК является каталитической частью рибонуклеопротеина RNase P (рибонуклеаза P). RNase P – это рибозим, т.е. фермент на основе РНК (один из немногих), который участвует в процессе доработки РНК, а именно в созревании тРНК. Рибозим чрезвычайно консервативен во всех организмах его содержащих, т.е. выполняет строго определённую функцию. Сама РНК является одной из древнейших, т.к. без неё невозможно синтезировать тРНК, которая необходима для трансляции белков [6].
2. rRNA: такие РНК входят в состав рибосом, составляя их каркас, участвуя в формировании пептидных связей (Большой субюнит) и расположении мРНК и тРНК для успешной трансляции (Малый субюнит). Как и предыдущая РНК, чрезвычайно консервативна.
3. SRP_RNA: РНК представляет собой компонент сигнального распознающего комплекса (Signal Recognition Particle), который участвует в процессе транспортировки белков через клеточную мембрану. Сигнальные пептид обычно располагается на конце белка, и указывает, что белок должен быть транспортирован через какую-либо мембрану. Данная РНК способствует связыванию комплекса с рибосомой, синтезирующий белок с таким пептидом [7].

4. tRNA: РНК транспортирует аминокислоты к рибосоме, где происходит синтез полипептидной цепи. Каждая тРНК имеет уникальную последовательность для связывания с определённой аминокислотой. На одной из сторон молекулы находится антикодон, комплементарный кодону на мРНК, для точности связывания.

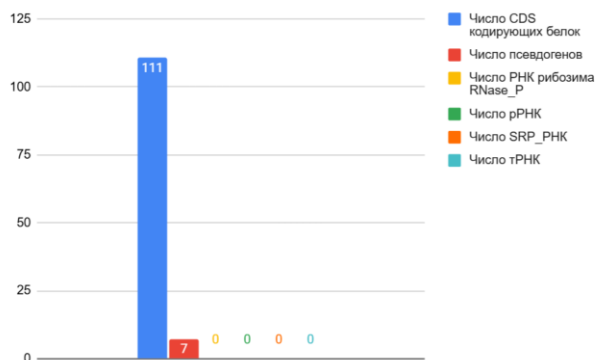


Рис 4. Гистограмма количества генов и их функций на рМ27-SA2-01

Следующий и последний репликон (рис. 4) уже имеет название. Из гистограммы видно, что гены, содержащиеся в нём, несут информацию о белках, а также имеются псевдогены, их доля составляет 0,06, т.е. 6%. Такой репликон мог появиться в связи с требованием повышения мобильности и адаптивности археи к различным изменениям окружающей среды, т.е. более эффективно выполнять необходимые задачи без помощи большого репликона, в котором содержится очень большое количество информации [8].

Если сравнить репликоны по составу, очевидно, что репликон рМ27-SA2-01 зависит от большого. Во-первых, вероятнее всего, большой репликон кодирует жизненно-необходимые белки, в то время как меньший отвечает лишь за какие-то несколько функций. При этом, все РНК закодированы в большом, следовательно, меньший зависит как минимум по процессам трансляции.

3.4 Гистограмма расстояний между участками CDS

Из гистограммы можно сделать несколько выводов (рис. 5):

1. В основном расстояния не превышают 500 нуклеотидов, но есть также участки, расстояние между которыми отрицательное, т.е. считывание начинается с конца другого участка.
2. Расстояний более 8000 нуклеотидов почти не наблюдается, но отдельные большие участки могут представлять определённый интерес, т.к. возможно в процессе эволюции произошло расширение генетического материала, не затронувшее CDS, но изменившее структуру генома [8]



Рис 5. Гистограмма межгенных промежутков

Анализ предоставляет полезную информацию о структуре генома, функциональной организации и структуре организма. Обычно у прокариотических организмов эти расстояния гораздо меньше, нежели у эукариотических, содержащих интроны, регуляторные элементы и некодирующие области в большем количестве.

Близкое расположение участков или их перекрывание может свидетельствовать о наличии **оперонов** – участков, включающих в себя несколько CDS, но выполняющих очень близкие или зависимые друг от друга функции.

3.5 Гистограмма расстояний между пересекающимися CDS

Из гистограммы (рис. 6) видно, что значительная доля перекрываний приходится на значение 4 (то есть перекрываются по 4 нуклеотидам). Это значение может быть обусловлено перекрыванием стоп-кодонов предшествующего участка с старт-кодоном следующего участка. Например, стоп-кодон **TGA** перекрывается с старт-кодоном **ATG**. Самое большое перекрывание достигает промежутка от 85 до 90 нуклеотидов.



Рис 6. Гистограмма пересекающихся CDS

Как обсуждалось ранее, перекрывание нуклеотидов между разными участками CDS может свидетельствовать о наличии оперонов. Гены в составе оперона транскрибируются в одну молекулу мРНК. Наличие таких участков помогают клетке скоординированно активировать или дезактивировать определённые биохимические процессы, реализующиеся несколькими белками. Так как прокариотам зачастую необходимо быстро адаптироваться к изменениям окружающей среды, данные участки жизненно необходимы для взаимодействия с окружающей средой с минимальными затратами. Чем больше длина перекрывания между участками – тем более они зависимы друг от друга [9].

3.6 Опероны в CDS

Говоря об оперонах, то найти хотя бы несколько таких участков самостоятельно довольно несложно. Необходимо учитывать несколько факторов: расстояние между участками (это может быть и перекрыванием, и просто расположением участков на расстоянии нескольких нуклеотидов друг от друга) и природа того, что этот участок кодирует. Это могут быть белки или РНК, выполняющие функции, относящиеся к одному и тому же процессу в организме.

В последующих таблицах (таб. 1-3) представлены предположительные опероны:

Таблица 1. Первый предположительный оперон

Номер CDS	Расстояние между CDS	Что кодируют
101	-14	ATP synthase archaeal subunit H
102	6	V-type ATP synthase subunit I
103	15	F0F1 ATP synthase subunit C
104	-4	V-type ATP synthase subunit E
105	-4	ATP synthase A1 subunit C
106	-4	V-type ATP synthase subunit F
107	-4	ATP synthase subunit A
108	12	ATP synthase subunit B
109	30	V-type ATP synthase subunit D

1. Гены находятся близко к друг другу, т.е. либо перекрываются, либо располагаются поблизости

2. Все перечисленные гены кодируют субъединицы АТФ-синтазы, которая является мультисубъединичным комплексом (множество субъединиц, объединяющиеся для выполнения определённой биологической функции). Скорее всего, они все экспрессируются одновременно для построения данного комплекса.

Таблица 2. Второй предположительный оперон

Номер CDS	Расстояние между CDS	Что кодируют
151	-4	ECF transporter S component
152	-4	energy-coupling factor ABC transporter ATP-binding protein
153	5	ABC transporter ATP-binding protein

1. Гены снова находятся близко друг к другу

2. Все указанные гены связаны с транспортными белками (ABC – транспортеры и ECF – компоненты), которые выполняют одну задачу – транспорт веществ через мембрану. Вероятнее всего, экспрессируются они все вместе

Таблица 3. Третий предположительный оперон

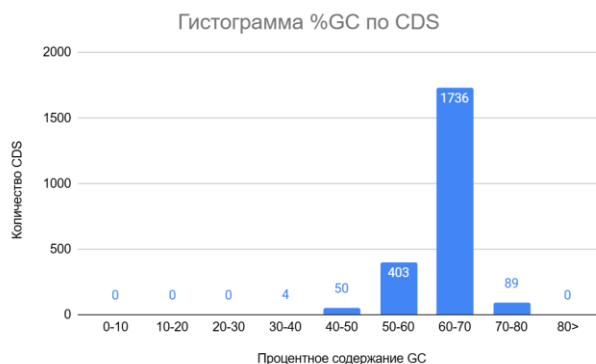
Номер CDS	Расстояние между CDS	Что кодируют
239	-1	DNA-directed RNA polymerase subunit H
240	1	DNA-directed RNA polymerase subunit B''
241	1	DNA-directed RNA polymerase subunit B
242	-8	DNA-directed RNA polymerase subunit A'
243	2	DNA-directed RNA polymerase subunit A''

1. Гены вновь расположены близко друг к другу

2. Данные последовательности кодируют субъединицы ДНК-зависимой РНК-полимеразы. Этот фермент участвует в транскрипции – синтезе РНК на матрице ДНК. С большой долей вероятностью, эти участки экспрессируются вместе.

3.7 Гистограмма %GC по CDS

Из гистограммы (рис. 7) видно, что содержание GC в CDS велико: в среднем это 60-70%, отдельные участки содержат от 40 до 60%, или от 70 до 80%.

**Рис 7. Количество кодирующих последовательностей, распределённых по процентному содержанию GC**

Вычисление отношения гуанина и цитозина к общему количеству азотистых оснований необходимо для понимания нескольких важных аспектов об геноме в целом:

1. Высокое процентное содержание GC указывает на большую стабильность ДНК, т.к. связи между гуанином и цитозином более прочные (три водородные связи), чем между аденином и тимином (две водородные связи). Такая молекула более устойчива к разрушению и деформации, разрывам [10].

2. Если процентное содержание GC высокое, это также может говорить об ареале обитания археи, например, о высокой температуре, а как мы знаем, для этой археи условия благоприятны при 37-40 градусах [10].
3. CDS с высоким процентным содержанием GC, вероятнее всего, синтезируют белки, т.к. эти азотистые основания влияют на экспрессию генов, т.е. на их активность при передаче генетического материала в виде аминокислотных последовательностей, а также уменьшают количество ошибок при репликации [11].
4. Кодированные CDS содержат гораздо больший процент GC по сравнению с некодирующими. Последние зачастую служат местом нейтральных мутаций, поэтому количество GC может сильно уменьшаться [12].

Благодарности

Благодарю своих бабушку и маму за то, что они готовили кушать, пока я изучал архею выше).

Сопроводительные материалы

1. <https://kodomo.fbb.msu.ru/wiki/Users/sergox> - Моя вики-страница на сайте Kodomo со всеми таблицами Google

Ссылки

1. Dmitry Y. Sorokin, Ilya V. Kublanov, Mikhail M. Yakimov, W. Irene C. Rijpstra, Jaap S. Sinninghe Damste (2016), ***Halanaeroarchaeum sulfurireducens* gen. nov., sp. nov., the first obligately anaerobic sulfur-respiring haloarchaeon, isolated from a hypersaline lake.** Microbiology society
2. Axel Tiessen, Paulino Pérez-Rodríguez, Luis José Dela-ye-Arredondo (2012), **Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes.** BMC
3. Nobuhiko Tokuriki, Dan S Tawfik (2009), **Stability effects of mutations and protein evolvability.** ScienceDirect
4. Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korb, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder (2007), **What is a gene, post-ENCODE? History and updated definition.** Cold Spring Harbor Laboratory Press Bookstore (CSH)

5. ZhaoLei Zhang, Mark Gerstein (2004), **Large-scale analysis of pseudogenes in the human genome.** ScienceDirect

6. Cecilia Guerrier-Takada, Kathleen Gardiner, Terry Marsh, Norman Pace, Sidney Altman (1983), **The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.** 50 Cell

7. Christian Zwieb, Jerry Eichler (2001), **Getting on target: The archaeal signal recognition particle.** PMC

8. Christopher M. Thomas, Kaare M. Nielsen (2005), **Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria.** Nature

9. S Sáenz-Lahoya, N Bitarte, B García, S Burgui, M Vergara-Irigaray, J Valle, C Solano, A Toledo-Arana, I Lasa (2019), **Noncontiguous operon is a genetic organization for coordinating bacterial gene expression.** PMC

10. Alexander E Vinogradov (2003), **DNA helix: the importance of being GC-rich.** Oxford Academic

11. T Ikemura (1985), **Codon usage and tRNA content in unicellular and multicellular organisms.** Oxford Academic

12. N Sueoka (1993), **Directional mutation pressure, mutator mutations, and dynamics of molecular evolution.** Springer Nature