Biostatistics (2018), **0**, 0, *pp.* 1–12
doi:10.1093/biostatistics/samplebibtex

Обзор протеома бактерии *Salinibacter ruber* DSM 13855

Руслан Салаватович Салимгареев

Факультет биоинженерии и биоинформатики,
Московский государственный университет имени М. В. Ломоносова,

Ленинские горы д. 1 стр. 73, 119234, Москва, Россия

russal2010@kodomo.fbb.msu.ru

Аннотация

В настоящее время в открытом доступе появляется всё больше отсеквенированных и аннотированных геномов. Примечательно, что получить содержательные результаты из этих данных можно сравнительно простыми способами. В частности, при обработке таблиц удобна программа Microsoft Excel. В этой работе с использованием MS Excel 2013 на основе открытых данных NCBI проанализированы особенности кодирования белков в геноме галофильной бактерии *Salinibacter ruber*.

Ключевые слова: Галофилы; Прокариоты; Протеом.

1. Введение

Солёные озёра — одно из самых суровых местообитаний на планете. Лишь немногие виды микроорганизмов способны там существовать, среди которых *Salinibacter ruber* – бактерия, открытая в 2000 году путём секвенирования гена 16S-субъединицы рибосомной РНК (Antòn *et al.*, 2000): образцы были получены из воды прудов Braç del Port в Испании. До этого в

условиях столь высокой солёности (до 37%) были известны только археи.

Примечательно, что *S. ruber* имеет множество общих черт с археями. Часть из этих черт, по всей видимости, результат конвергенции, например, крайне высокая концентрация ионов калия в цитоплазме. В то же время предполагается возможность горизонтального переноса генов с археями в прошлом, в частности, генов некоторых родопсинов (Oren *et al.*, 2002; Mongodin *et al.*, 2005).

В данной работе проанализирована таблица, описывающая транскриптом *S. ruber* штамма DSM 13855, при этом отмечены некоторые особенности распределения генов белков и РНК.

2. Материалы и методы

Исходные данные получены с сайта National Center for Biotechnology Information (2018), информация о точной длине цепочек ДНК бактерии взята из работы Mongodin *et al.* (2005): хромосома длиной 3551823 пары оснований, плазмида pSR35 длиной 35505 пар оснований. Данные о классах существования белков приводятся в соответствии с сайтом UniProt (2018). Для обработки использовались возможности программы MS Excel 2013. Оценка числа генов на 1 млн нуклеотидов была выполнена по формуле:

$$n = \frac{N_g}{L_{bp} \cdot 1 \cdot 10^{-6}}$$

где n – искомая величина, N_g – число генов, L_{bp} – длина цепи ДНК в парах оснований.

Работа в MS Excel

Для выбора данных широко использовались сводные таблицы. При этом применялся фильтр по значению CDS в поле "# feature", если нужно было отсечь белки (у РНК там написан вид РНК). Колонка "name" была использована, чтобы определять классы белков, причём

численность класса «остальные» вычислена путём вычитания уже приведённых в таблице белков из общего числа (см. лист "proteins" таблицы `features.xlsx`).

Сопоставление граф "class" и "strand" в сводной таблице дало материал для табл. 3 (лист "pivot-strands" таблицы `features.xlsx`).

Длины белков для статистической обработки скопированы из колонки "product_length" на лист "pr13-2-histogram".

Координаты начал и концов генов получены с помощью фильтров: значение "gene" в столбце "# feature" и значения в столбцах "seq_type" и "strand", задающие конкретную цепь хромосомы или плазмиды. (Результат операции вынесен на листы "chr−" – "pl+" таблицы). На каждом из 4 листов с помощью формул были вычислены размеры попарных пересечений соседних генов и межгенные промежутки. Для того чтобы убрать нули из столбца чисел, был использован сценарий `compress.py` на языке программирования Python версии 3.6.

При составлении диаграммы распределения белков по категориям существования полученные с сайта UniProt данные были помещены на лист "uniprot" таблицы `features.xlsx`.

3. Результаты

3.1 *Распределение генов по категориям*

Распределение генов белков и генов РНК по категориям их продуктов приведено в табл. 1 и табл. 2 соответственно. Видно, что доля гипотетических белков в протеоме составляет около одной трети. Из закодированных РНК больше всего транспортных (44), есть также 3 рибосомальных и 1 РНК другого вида.

Число генов на 1 млн пар оснований (с точностью до десятков): в хромосоме – 810, в плазмиде – 930.

3.2 *Распределение белков по длинам*

На рис. 1 приведена гистограмма длин всех белков, а на рис. 2 – гистограмма для белков длины не более 1200 а.к. в более крупном масштабе. Из рисунков видно, что если в вероятностном пространстве рассматривать все закодированные белки как равновероятные исходы, то наиболее вероятной будет длина ≈ 250 а.к.

Таблица 3 показывает статистические данные по длинам белков. Из неё следует, что и медиана, и среднее на диаграммах находились бы *правее* самого высокого столбца.

3.3 *Распределение категорий генов по цепям ДНК*

Распределение генов по двум комплементарным цепочкам хромосомы и двум цепочкам плазмиды приведено в табл. 4. Интересно, что на плазмиде гены РНК не встречаются. В целом самая обширная категория – гены белков.

3.4 *Пересечения генов и межгенные промежутки*

Статистические данные по пересечениям генов в хромосоме представлены в табл. 5. В плазмиде присутствует только одно пересечение длины 8 на обратной цепи. Гены не пересекаются в пределах одной рамки считывания: все пересечения на одной цепочке имеют длину, не кратную 3.

Среди межгенных промежутков подавляющее большинство небольших, поэтому при построении гистограмм была использована логарифмическая шкала. Результаты представлены на рис. 3–6.

3.5 *Классы существования белков*

Соотношение количества белков различных классов приведено на рис. 7. Видно, что 70 % из них предсказаны, то есть их существование не подтверждено ни наблюдением транскрипции, ни наблюдением самого белка, а также не следует из гомологии с известными белками родственных организмов. Лишь менее 1 % протеома – белки, существование которых подтверждено непосредственно; оставшиеся предполагаются из гомологии.

Замечание. По неясным причинам наблюдается расхождение в числах: в UniProt присутствует 2780 белков, кодируемых генами на хромосоме, и 32 – на плазмиде, по данным анализа таблицы с NCBI же на хромосоме на 21 больше кодирующих белки последовательностей, то есть 2801.

4. Обсуждение

Как выяснилось, число транспортных РНК намного меньше количества кодонов, задающих аминокислоты: 44 против 61. Это может означать, что некоторые тРНК взаимозаменяемы.

Отсутствие генов РНК на плазмиде может определяться их важностью для бактерии: она не может себе позволить потерять РНК вместе с плазмидой, если выведет её во внешнюю среду.

Гены могут пересекаться, хотя происходит это не очень часто (≈ 170 случаев на ≈ 2800 генов). По одной рамке считывания они не пересекаются вовсе, хотя можно было бы представить, например, два альтернативных старт-кодона, соответствующих одному стоп-кодону.

Промежутки между генами, которые всё-таки не пересеклись, могут быть всевозможными: от нуля до тысяч пар оснований. Распределение промежутков очень неравномерное, и короткие встречаются намного чаще длинных.

Большинство белков в протеоме *S. ruber* никогда не были сами секвенированы, и их существование предполагается на основании косвенных данных.

5. Сопроводительные материалы

Сопроводительные материалы доступны по ссылке `http://kodomo.fbb.msu.ru/~russal2010/term1/block4/features.zip`.

Благодарности

Хочу поблагодарить Софью Александровну Гайдукову, Ольгу Алексеевну Салимгарееву, Илью Александровича Семерикова и Бориса Александровича Фенюка за мотивацию к поступлению на ФББ МГУ. Без них я вряд ли написал бы эту статью.

Список литературы

Antòn, J., Rossellò-Mora, R., Rodrìguez-Valera, F. and Amann, R. (2000). Extremely halophilic *Bacteria* in crystallizer ponds from solar salterns. *Applied and environmental microbiology* **66**, 3052—-3057.

Mongodin, E., Nelson, K., Daugherty, S., DeBoy, R., Wister, J., Khouri, H., Weidman, J., Walsh, D., Papke, R., Sanchez Perez, G., Sharma, A. *et al.* (2005). The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18147–18152.

National Center for Biotechnology Information. (2018). OnlineDoc: `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/013/045/GCA_000013045.1_ASM1304v1/GCA_000013045.1_ASM1304v1_feature_table.txt.gz`.

Oren, A., Heldal, M., Norland, S. and Galinski, E. (2002). Intracellular ion and organic solute concentrations of the extremely halophilic bacterium *Salinibacter ruber*. *Extremophiles* **6**, 491–498.

UniProt. (2018). Веб-страница: *organism:"Salinibacter ruber <...>" in UniProtKB*. https://www.uniprot.org/uniprot/?query=proteome:UP000008674.

[]

Рис. 1. Гистограмма длин белков. На горизонтальной оси промежутки указаны в виде полуинтервалов, где начало не включается.

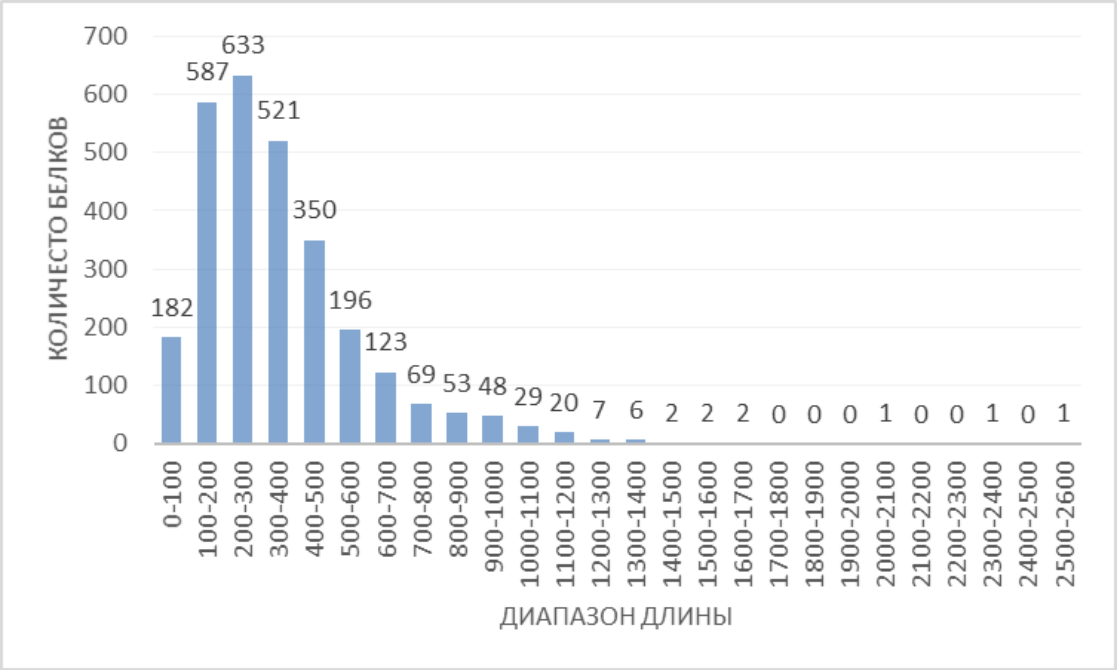

Рис. 2. Гистограмма длин белков, диапазон до 1200 а.к. На горизонтальной оси промежутки указаны в виде полуинтервалов, где начало не включается.

Рис. 3. Распределение межгенных промежутков на прямой цепи хромосомы по длинам. Правый конец промежутка на оси не включается.

Рис. 4. Распределение межгенных промежутков на обратной цепи хромосомы по длинам. Правый конец промежутка на оси не включается.

Рис. 5. Распределение межгенных промежутков на прямой цепи плазмиды по длинам. Правый конец промежутка на оси не включается.

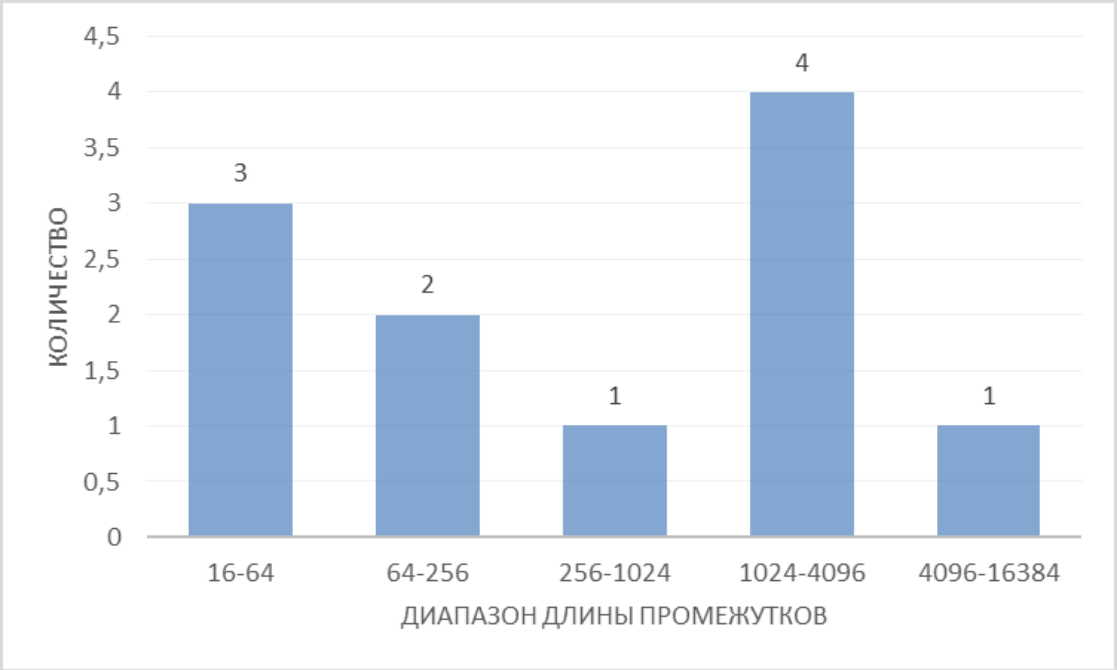

Рис. 6. Распределение межгенных промежутков на обратной цепи плазмиды по длинам. Правый конец промежутка на оси не включается.

Рис. 7. Круговая диаграмма, показывающая классы существования белков по UniProt. Указано абсолютное число белков, имеющих данный статус, и их доля во всём протеоме.

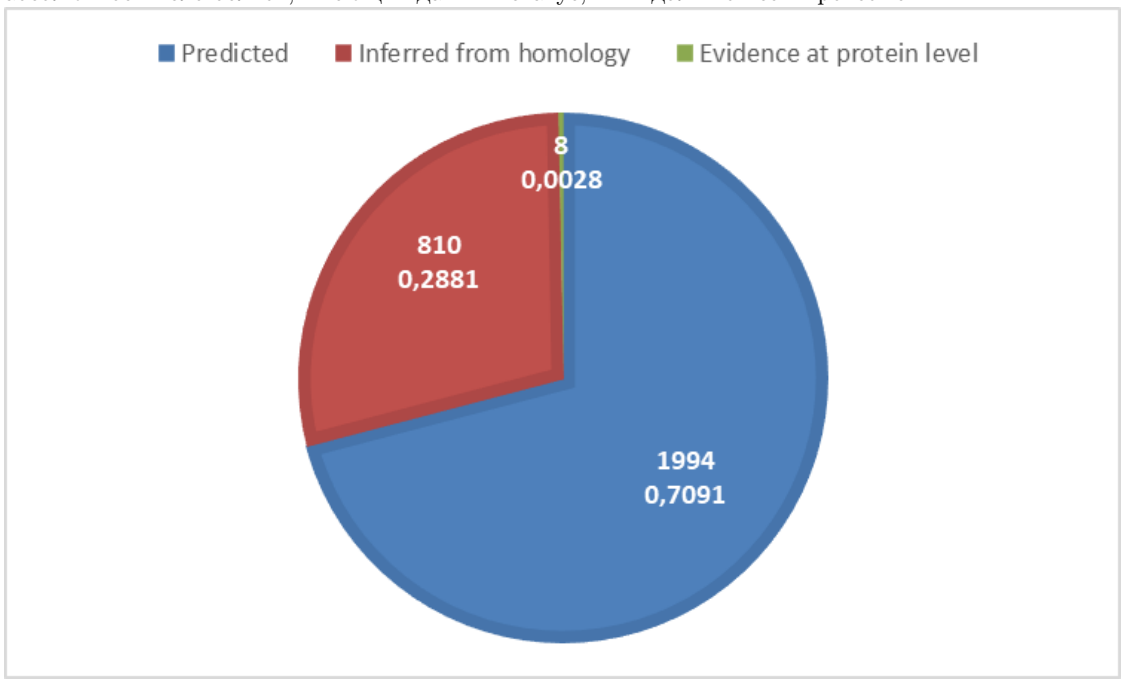

Таблица 1. Число генов белков по категориям

Категория	Число
Рибосомальные	56
Транспортные	34
Гипотетические	946
Остальные	1797

Таблица 2. Число генов РНК по категориям

Категория	Число
Транспортные	44
Рибосомальные	3
Остальные	1

Таблица 3. Статистические данные для длин белков

Минимум	Максимум	Среднее	Медиана	Ст. отклонение
30	2597	357,20	303	242,15

Таблица 4. Распределение генов по цепочкам

	Хромосома			Плазмида		
Цепочка	гены белков	псевдогены	гены РНК	гены белков	псевдогены	гены РНК
Прямая	1444	7	20	12	–	–
Комплементарная	1357	4	28	20	1	–

Таблица 5. Пересечения генов в кольцевой хромосоме

Цепочка	Число пересеч.	Ср. длина	Мед. длина	Станд. отклонение	Минимум	Максимум
Прямая	100	72,32	21,5	169,16	1	719
Компл.	73	27,68	16	38,07	1	274