# Обзор протеома бактерии Mycobacterium haemophilum

Покровский Сергей Юрьевич

1 курс, Факультет биоинженерии и биоинформатики, Московский Государственный Университет имени М. В. Ломоносова, Москва, Россия

**РЕЗЮМЕ**

В данной работе с использованием Microsoft Excel 2010 и исходных данных о протеоме Mycobacterium haemophilum DSM 44634 было проведено исследование, в результате которого были построены гистограммы распределения белков по длине, квазиоперонов по числу генов, определена доля гипотетических белков в протеоме, отмечены особенности перекрывания генов.

**КЛЮЧЕВЫЕ СЛОВА**

Протеом, распределение белков, квазиопероны, перекрывание генов, Mycobacterium haemophilum DSM 44634.

## 1 ВВЕДЕНИЕ

Mycobacterium haemophilum – вид патогенных для человека бактерии, вызывающих поражения кожи у лиц с ослабленным иммунитетом. Как и другие представители рода Mycobacterium, данные бактерии являются грамположительными неподвижными аэробными бациллами, проявляющими свойства кислото- и спиртоустойчивости. Характерными особенностями Mycobacterium haemophilum являются необходимые для оптимального роста достаточно низкая температура (30 – 32 °C) и содержание в субстрате соединений железа: гемина либо цитрата железа(III)-аммония (Lindeboom J.A., Bruijnesteijn van Coppenraet L.E.S., van Soolingen D., Prins J.M., Kuijper E.J., 2011). Как было показано JoAnn M. Tufariello и соавторами (2015), геном Mycobacterium haemophilum представлен 1 кольцевой хромосомой, которая содержит 3,964 гена и состоит из 4,235,765 пар нуклеотидов.

Данная работа имеет целью выявить некоторые протеомные и геномные особенности Mycobacterium haemophilum (штамм DSM 44634), а также проверить следующие гипотезы:

1) распределение генов по прямой и комплементарной цепи носит случайный характер;
2) доля гипотетических белков зависит от рассматриваемого диапазона длин;
3) перекрывание генов не зависит от того, расположены гены на одной цепи или на разных;
4) среди квазиоперонов преобладают короткие (из 2-3 генов)

## 2 МАТЕРИАЛЫ И МЕТОДЫ

Исходные данные:
https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/Mycobacterium%20haemophilum%20DSM%2044634
Инструменты: Microsoft Excel 2010.

Исходные данные в формате текстовой таблицы были импортированы в Microsoft Excel при помощи встроенного мастера импорта. Для удобства работы значения в столбце strand были заменены с + и – на +1 и -1 соответственно. Далее строки, содержащие основную и дополнительную информацию о генах, были отделены друг от друга при помощи фильтра (по столбцу # feature) и скопированы на отдельные листы (1 и 2 соответственно). Затем столбец locus_tag на листе 2 был помещён в начало таблицы и использован как ключ, чтобы с помощью функции VLOOKUP сопоставить каждой строке таблицы 1 соответствующую строку таблицы 2.

Для исследования распределения белков по длинам была построена таблица со столбцом чисел от 0 до 1000 с шагом 50, где каждая пара чисел задаёт диапазон длины белка, и столбцом формул COUNTIFS, при помощи которых было подсчитано количество белков в каждом диапазоне, а также с помощью функции COUNTIF – количество белков с длиной больше 1000; по полученным данным была построена гистограмма.

Аналогично были построены гистограммы распределения доли гипотетических белков в различных диапазонах длин, межгенных промежутков по длинам и квазиоперонов по числу генов. Гипотетические белки определялись по значению hypothetical protein в поле name. Межгенные промежутки определялись как расстояния между концом одного гена и началом следующего; перекрывания генов – как промежутки, не превышающие 0; для определения длины квазиоперонов были отмечены гены, длина промежутка после которых превышала 100 (отдельно рассмотрены гены на разных цепях), и подсчитано число генов между ними.

Гены белков определялись по значению protein_coding, гены РНК – по значениям rRNA, tRNA или Rnase_P_RNA, псевдогены – по значению pseudogene в поле class. Вероятность распределения генов по цепям была определена путём генерации 2000 случайных последовательностей длины, равной количеству генов, из 1 и 0, подсчёта количества 1 в каждой из них, вычисления отклонения этого количества от ожидаемого (половина длины) и определения доли последовательностей, в которых отклонение не меньше, чем в распределении генов.

Для выявления рибосомальных белков и РНК, самых длинных, самых коротких белков и белков, отвечающих за устойчивость к медицинским препаратам, были использованы фильтр по значению ribosomal в поле name, сортировка по убыванию/возрастанию значений поля product_length и фильтр по значениям resistance и multidrug в поле name.

## 3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

### 3.1 Распределение белков по длине

Распределение белков Mycobacterium haemophilum DSM 44634 по длине отражено на гистограмме (рис. 1). По горизонтальной оси расположены категории белков по длине в аминокислотных остатках (а. о.), по вертикальной оси – количество белков, принадлежащих каждой категории; шаг гистограммы был выбран равным 50 а. о.

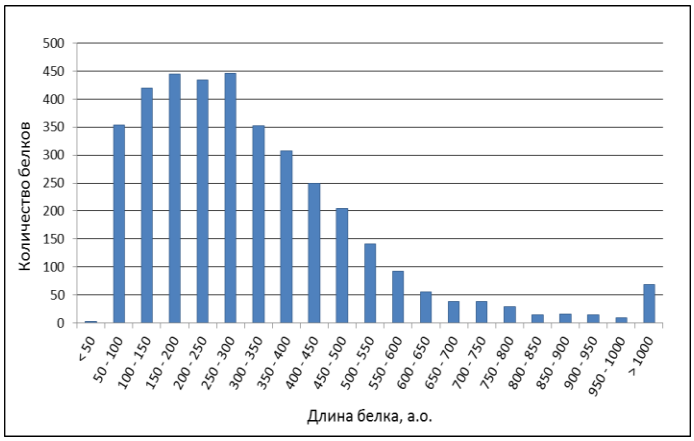


**Рисунок 1.** Гистограмма распределения белков протеома по длине

Как видно по рис. 1, большинство белков протеома имеют длину от 50 до 400 а. о.; медианное значение длины белка при этом составляет 274 а. о. По-видимому, белки длиной 50 – 400 а. о. являются типичными для данной бактерии.

Определённый интерес представляют белки с наибольшим отклонением длины от среднего значения. Оказалось, что, например, 11 самых длинных белков протеома являются поликетидсинтазами. Скорее всего, это явление объясняется тем, что настолько длинные белки нетипичны для данной бактерии, и вероятность того, что среди них будут присутствовать два или более абсолютно различных белков, крайне мала, в то время как взаимно схожие белки с большими значениями длин в отсутствие других оказываются сгруппированы.

### 3.2 Рибосомальные белки и рРНК

У Mycobacterium haemophilum обнаружены:

1. 21 белок малых субъединиц рибосом: 30S ribosomal protein S1 – 30S ribosomal protein S20, 30S ribosomal protein S30;
2. 31 белок больших субъединиц рибосом: 50S ribosomal protein L1 – 50S ribosomal protein L11, 50S ribosomal protein L13 – 50S ribosomal protein L25, 50S ribosomal protein L27 – 50S ribosomal protein L33, 50S ribosomal protein L35;
3. 3 рРНК: 5S ribosomal RNA, 16S ribosomal RNA, 23S ribosomal RNA.

### 3.3 Некоторые особенные белки, присутствующие в протеоме

#### 3.3.1 Поликетидсинтазы

Являются самыми длинными белками в протеоме Mycobacterium haemophilum. Поликетидсинтазы – это ферменты, катализирующие синтез поликетидов – поликарбонильных соединений со сложной структурой, многие из которых проявляют свойства токсинов или антибиотиков. В связи со строением поликетидов для их биосинтеза требуются сложные ферменты, что объясняет относительно большие значения длины полипептидной цепи поликетидсинтаз (до 4185 а. о.)

#### 3.3.2 Предшественник микофактоцина

Является самым коротким белком в протеоме данной бактерии (29 а. о.). В процессе посттрансляционной модификации из него образуется микофактоцин – общий для многих представителей рода Mycjbacterium белок, функционирующий как переносчик электронов. Был описан в 2011 году в статье Daniel H. Haft.

#### 3.3.3 Белки, обусловливающие устойчивость к медицинским препаратам

В протеоме Mycobacterium haemophilum были обнаружены белки длиной от 107 до 803 аминокислотных остатков, обеспечивающие устойчивость бактерии к некоторым медицинским препаратам:

1. Multidrug transporter, multidrug ABC transporter permease, multidrug ABC transporter ATP-binding protein, multidrug MFS transporter – обеспечивают устойчивость ко многим препаратам, выводя их из клетки;
2. Camphor resistance protein CrcB – обеспечивает устойчивость к камфоре;
3. Daunorubicin resistance protein DrrC – обеспечивает устойчивость к даунорубицину (антрациклиновому антибиотику);
4. Bleomycin resistance protein – обеспечивает устойчивость к блеомицину (гликопептидному антибиотику).

### 3.4 Распределение генов по прямой и комплементарной цепям ДНК

Информация о числе генов белков, генов РНК и псевдогенов на каждой из цепочек ДНК представлена в таблице 1.

| | Гены белков | Гены РНК | Псевдогены | Всего |
|---|---|---|---|---|
| **Прямая цепь** | 1961 | 26 | 80 | 2067 |
| **Комплементарная цепь** | 1767 | 23 | 73 | 1863 |

**Таблица 1.** Распределение генов белков, генов РНК и псевдогенов по прямой и комплементарной цепям ДНК

Можно заметить, что общее число генов (3930) отличается от полученного JoAnn M. Tufariello в 2015 году; это связано с последующим обновлением базы данных.

Из таблицы ясно, что во всех категориях генов их число на прямой цепи несколько превосходит число на комплементарной цепи. По результатам моделирования вероятность распределения 3930 генов: 2067 на прямой цепи и 1863 на комплементарной цепи оказалась равной 0,0005; это позволяет сделать вывод о том, что гипотеза (1) о случайном распределении ошибочна с высоким уровнем значимости.

## 3.5 Доля гипотетических белков среди белков различных длин

Ген гипотетического белка – это ген, схожий с геном белка, но доказательств его экспрессии не получено. В протеоме Mycobacterium haemophilum DSM 44634 гипотетическими являются 1345 из 3728 белков, что составляет 36,08 %. На рис. 2 приведена гистограмма, отображающая процент гипотетических белков в различных диапазонах длин. Видно, что среди белков длиной от 300 до 500 а. о. и более доля гипотетических белков составляет немногим более 20 % от общего числа, в то время как среди более коротких белков она тем выше, чем меньше их длина, и достигает 68 % в диапазоне длин менее 100 а. о. Гипотеза (2) оказалась верна.


**Рисунок 3.** Гистограмма зависимости доли гипотетических белков от рассматриваемого диапазона длин

Предположительные объяснения этого результата:

1) для более короткой кодирующей последовательности с большей вероятностью найдётся кодирующая последовательность гомологичного гипотетического белка у родственного вида, наличие которой подтвердит его существование;
2) возможно, в силу некоторых причин экспериментально подтвердить экспрессию гена короткого белка более затруднительно.

## 3.6 Распределение межгенных промежутков по длине

На рис. 3 приведена гистограмма распределения промежутков между генами по длине в парах нуклеотидов (п. н.).

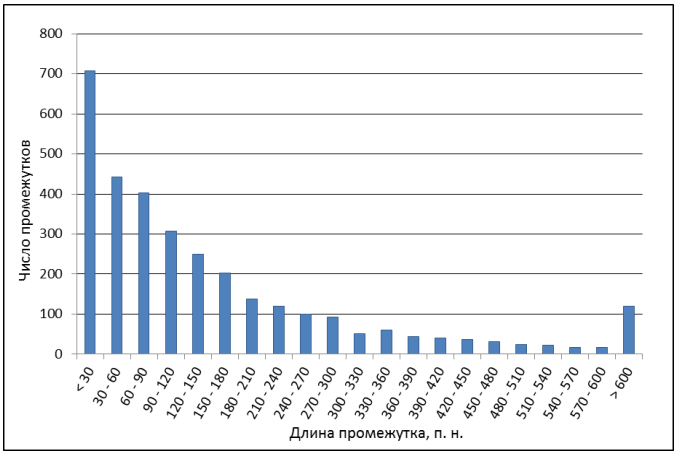


**Рисунок 2.** Гистограмма распределения межгенных промежутков по длине

Можно видеть, что для генома Mycobacterium haemophilum DSM 44634 наиболее характерны межгенные промежутки длиной менее 30 п. н., и количество промежутков определённой длины тем меньше, чем больше длина. Незначительное число длинных промежутков ожидаемо, поскольку межгенные промежутки не несут в себе информации, а компактный геном, для воспроизведения которого не требуется много ресурсов, определённо выгоден для одноклеточного быстро размножающегося организма.

## 3.7 Особенности перекрывания генов

При анализе длин перекрываний генов было выяснено, что большинство из них (576 из 688, или 83,7 %) имеют длину менее 10 пар нуклеотидов. Дальнейшее исследование показало, что 61,48 % от всех перекрываний составляют перекрывания по 4 парам нуклеотидов. Было также изучено, каким именно образом перекрываются гены, поскольку они могут находиться как на прямой, так и на комплементарной цепи. Все перекрывания были условно разделены на 4 типа в зависимости от того, на какой цепи расположены первый и второй перекрывающиеся гены. Например, перекрывание типа 1-ый + / 2-ой + по 4 парам нуклеотидов означает, что старт-кодон 2-го гена перекрывается по 1 нуклеотиду с предпоследним кодоном и по 2 нуклеотидам со стоп-кодоном 1-го гена; типа 1-ый + / 2-ой - – что триплет, комплементарный перевёрнутому (записанному от 3’ к 5’-концу) стоп-кодону 2-го гена, перекрывается по 1 нуклеотиду с предпоследним кодоном и по 2 нуклеотидам со стоп-кодоном 1-го гена.

Полученная информация представлена в таблице 2.

| Тип перекрывания | Число перекрываний по 4 п. н. | Всего перекрываний | Доля перекрываний по 4 п. н. |
|---|---|---|---|
| 1-ый +<br>2-ой + | 196 | 304 | 64,47% |
| 1-ый +<br>2-ой - | 51 | 109 | 46,79% |
| 1-ый -<br>2-ой + | 0 | 6 | - |
| 1-ый -<br>2-ой - | 176 | 269 | 65,43% |
| **Всего:** | 423 | 688 | 61,48% |

**Таблица 2.** Доля перекрываний по 4 парам нуклеотидов в зависимости от локализации перекрывающихся генов на прямой или комплементарной цепи

Из таблицы видно, что гены, находящиеся на одной цепи, часто перекрываются по 4 парам нуклеотидов; гены, перекрывающиеся по типу 1-ый + / 2-ой -, подвержены этому в меньшей степени; по типу 1-ый - / 2-ой + (когда участок, комплементарный перевёрнутому начальному участку 1-го гена, перекрывается с начальным участком 2-го гена) гены перекрываются крайне редко и вообще не перекрываются по 4 парам нуклеотидов. Таким образом, гипотеза (3) опровергнута.

Объяснение данному явлению, скорее всего, заключается в генетическом коде бактерии, а именно в невозможности перекрывания со старт-кодоном триплета, комплементарного перевёрнутому старт-кодону.

### 3.8 Распределение квазиоперонов по числу входящих в них генов

Было выяснено, что большая часть генов Mycobacterium haemophilum DSM 44634 группируется в квазиоперон как минимум с ещё одним геном. Не входят в квазиопероны 32,7 % генов на прямой цепи и 35,59 % генов на комплементарной цепи. На рис. 4 изображена гистограмма распределения квазиоперонов по числу генов, которые в них входят.

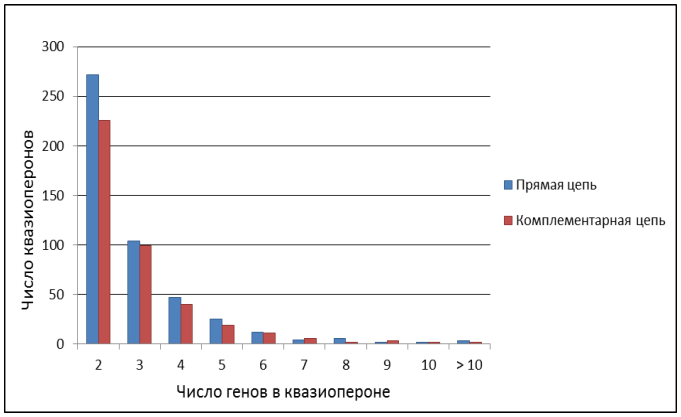


**Рисунок 4.** Гистограмма распределения квазиоперонов по числу входящих в них генов

Из рис. 4 можно сделать вывод, что последовательности из 2 идущих друг за другом генов встречаются намного чаще, чем более длинные квазиопероны; гипотеза (4) подтверждена. Также можно заметить небольшое различие в количестве квазиоперонов на разных цепях, но это объясняется общей неравномерностью распределения генов.

## 4 ЗАКЛЮЧЕНИЕ

В результате исследования было выяснено следующее:

1. Наиболее характерная для протеома Mycobacterium haemophilum длина белка – 50 - 400 аминокислотных остатков;
2. Mycobacterium haemophilum имеет 52 рибосомальных белка и 3 рРНК, а также ряд белков, придающих устойчивость к определённым лекарственным средствам;
3. Распределение генов по прямой и комплементарной цепочкам ДНК с высокой вероятностью носит неслучайный характер;
4. Доля гипотетических белков наиболее высока среди белков небольшой длины;
5. Для генома Mycobacterium haemophilum характерны небольшие по длине межгенные промежутки и квазиопероны из 2-3 генов;
6. Особенности перекрывания генов зависят от того, на каких цепочках расположены перекрывающиеся гены.

## 5 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Основной файл исследования:
https://kodomo.fbb.msu.ru/~pork7007/term1/Pokrovskiy_suppl.xlsx
Моделирование распределения генов:
https://kodomo.fbb.msu.ru/~pork7007/term1/Pokrovskiy_random_test.xlsx

## 6 БЛАГОДАРНОСТИ

Выражаю благодарность всем преподавателям ФББ МГУ, участвующим в подготовке и реализации курса биоинформатики.

## 7 СПИСОК ЛИТЕРАТУРЫ

1. Lindeboom JA, Bruijnesteijn van Coppenraet LES, van Soolingen D, Prins JM, Kuijper EJ. 2011. Clinical manifestations, diagnosis, and treatment of Mycobacterium haemophilum infections. Clin Microbiol Rev 24:701–717. doi:10.1128/CMR.00020-11
2. Tufariello JM, et al. 2015. The complete genome sequence of the emerging pathogen Mycobacterium haemophilum explains its unique culture requirements. MBio 6(6):e01313-01315
3. Haft, Daniel H. (2011). "Bioinformatic evidence for a widely distributed, ribosomally produced electron carrier precursor, its maturation proteins, and its nicotinoprotein redox partners". BMC Genomics. 12: 21. doi:10.1186/1471-2164-12-21. PMC 3023750. PMID 21223593