# Практикум по биоинформатике, организм: Mycoplasma Genitalium

Никита Карташов, Екатерина Ломерт, Екатерина Эсаулова

6 декабря 2014 г.

**Аннотация**

Your abstract.

## Описание организма

*Mycoplasma genitalium* относится к классу *Mollicutes* (микоплазмы). Микоплазмы имеют маленькие размеры (300 нм), из-за чего они не видны даже в световой микроскоп, у них нет собственной клеточной оболочки, и это сближает их с вирусами. Как и вирусы, микоплазмы не могут существовать иначе, чем паразитируя в клетках хозяина, из которых микоплазмы получают основные питательные вещества. Однако, в отличие от вирусов, микоплазмы способны расти в богатой бесклеточной среде.

*Mycoplasma genitalium* — паразитическая бактерия, живущая в половых и дыхательных системах приматов. Ещё недавно считалась организмом с наименьшим известным размером генома – 582 970 пар оснований (до обнаружения археи *Nanoarchaeum equitans* в 2002 году). Секвенирование генома началось с установления последовательности случайных генов в 1993 году Петересоном и было завершено в 1995 году.

В 2008 году в институте Крейга Вентера была полностью синтезирована молекула ДНК *Mycoplasma genitalium*. В искусственный геном были искусственно вставлены несколько маркеров, а также удалены гены, отвечающие за патогенность бактерии. В 2012 году ученые из Стэнфордского университета и Института Дж. Крейга Вентера создали первую программную симуляцию жизни *Mycoplasma genitalium*. Программа способна предсказывать биохимические процессы при воздействии различных факторов, в том числе фармакологических препаратов. Компьютерная модель работает на кластере из 128 компьютеров, имитируя полный цикл жизни микоплазмы. В ней учитывается взаимодействие всех ключевых молекул, один только перечень категорий которых составляет 28 пунктов.

### Используемые данные

**Штамм**: *Mycoplasma genitalium* G37
**NCBI Reference Sequence**: NC000908.2
**Размер (Мб)**: 0.580076
**GC%**: 31.7

**Количество белков**: 475
**Количество генов**: 524
**Количество псевдогенов**: 6
Ссылка на используемые данные:

`https://www.ncbi.nlm.nih.gov/genome/474?genome_assembly_id=166957`

# Поиск белков в геноме

Первым заданием было найти пять заданных протеиновых последовательностей в геноме. Мы выполнили поиск с помощью TBLASTN и Exonerate для штаммов G37, M2288, M2321, M6282, M6320. Так как у нас маленький организм, все белки нашлись с match $> 95$.

```
#!/bin/bash

genomes=$(ls ./genomes)
query=sample_proteins.fa

for gen in ${genomes}
do
db=db_"$gen"
makeblastdb -in ./genomes/${gen} -out ${db} -dbtype nucl
tblastn -query ${query} -db ${db} -outfmt 7 -out result_"$gen"
done
```

# Поиск генов. Сравнение инструментов: Glimmer и компания

## Работа с Glimmer'ом

Glimmer — набор программ для определения генов в бактериальных геномах. Принцип работы программы заключается в построении Марковской модели на тренировочном наборе данных, и затем использование получившейся модели для поиска генов в геноме.

В качестве тренировочных данных могут выступать:

1. Известные гены из генома (поиск по гомологии)
2. Достаточно длинные, неперекрывающиеся ORF (рамки считывания), которые могут быть определены при помощи скриптов Glimmer'а
3. Гены близкородственных организмов

Поскольку у *Mycoplasma genitalium* маленький геном, мы могли позволить себе запустить Glimmer много раз и сравнить полученные результаты. Поэтому мы попробовали все возможные варианты, а именно:

1. Взяли несколько штаммов *Mycoplasma genitalium*

2. Для каждого нашли гены на основе их ORF'ов

3. В качестве тренировочного набора взяли все гены близкородственной к *Mycoplasma genitalium* бактерии *Ureaplasma urealyticum serovar 10* и применили ко всем штаммам

### 0.1 Анализ результатов

Помимо Glimmer мы запустили Genemarks.

Был написан скрипт, проверяющий, что у гена есть рамка считывания, старт-кодон, стоп-кодон (у *Mycoplasma genitalium* стоп-кодоны только TAG и TAA) , что рамка считывания кратна трем и присутствует RBS.

Затем мы сравнили результаты с данными с NCBI. Glimmer нашел лишь 60% генов. Genemarks правильно определил 98% генов организма (515 ил 525), и нашел 40 лишних генов. Что характерно, все пропущенные гены были короче 400 нуклеотдов и имели GC–контент меньше 37%.

## Филогенетическое дерево

Были взяты белки трех организмов из филы *Firmicutes* (*Mycoplasma genitalium, Mesoplasma florum, Ureaplasma urealiticum*). Построены кластеры с помощью kClust. Полученная таблица загружалась в CAFE с настройками длины ветвей(в файле). По результатам CAFE построено дерево в R-studio (пакет ape, функция read.tree).