

# Домашнее задание №2

---

## Введение

Известно, что наша бактерия из Баренцева моря относится к виду *Thalassolituus oleivorans* (назовем ее, например, *T. oleivorans Barents*). При этом известен геном другого штамма этого вида бактерий с берегов Сицилии, Италия (*T. oleivorans MIL-1* – <http://www.ncbi.nlm.nih.gov/nuccore/HF680312>). Имеет смысл давать имена генам и белкам нашей бактерии схожие с теми, что были использованы для генов и белков двух других штаммов.

## Обязательная часть задания (8 баллов)

- На сайте [github.com](https://github.com) создаем публичный репозиторий и приводим ссылку на этот репозиторий в общей гугл-таблице (**вкладка HW2**)  
<https://docs.google.com/spreadsheets/d/1bdKBIDMFvhX8xStZUL2YHRTGLpmdiLNU5JcVmy8K-Yw/edit?usp=sharing>
- Рекомендуется выполнять работу в Google Collab ноутбуках.
  - Если вы будете выполнять работу на сервере или на своем компьютере, необходимо будет также загрузить написанный код на Github
- Подготавливаем файлы для аннотации
  - Пример Google Colab ноутбука:  
<https://colab.research.google.com/drive/1cwcaCz5kZhra8GwUcoABsHN0S9RA6UX8?usp=sharing>
- Используем полученные файлы для создания аннотированного генома бактерии в формате GenBank
  - Это можно сделать с помощью библиотеки Biopython
  - Пример работы с библиотекой:  
[https://colab.research.google.com/drive/1kv-bhhU1ykCOhl7OxzmKZN1l1qzOLL\\_r?usp=sharing](https://colab.research.google.com/drive/1kv-bhhU1ykCOhl7OxzmKZN1l1qzOLL_r?usp=sharing)

## Бонусная часть задания (2 балла)

- Предсказать рибосомальные РНК (5S, 16S, 23S) с помощью -- BLASTn
  - Взять нт посл-ти 5S, 16S, 23S рРНК бактерии *Thalassolituus oleivorans* и найти соотв участки в полученной сборке
  - Добавить эту инфу в полученный GenBank файл, а в отчете указать % сходства между соответствующими рРНК
  - Ноутбук с примером
  - <https://colab.research.google.com/drive/15CGH-caD6BODKifhM7sv8Sv10ZOf2qfO?usp=sharing>

## Список файлов для сдачи

- В репозитории в файле *README.md*
  - Ссылки на google colab ноутбуки
  - Статистика о том, сколько было предсказано генов всего, сколько из них удалось аннотировать с помощью сравнения с бактерией MIL-1, сколько с помощью БД SwissProt и какое кол-во белков остались без аннотации функции
  - Процент сходства между рибосомальными рнк
  - В репозитории в папке data:
    - GENOME.gbk -- аннотированный геном бактерии в формате GenBank
    - gms2.lst -- координаты всех предсказанных генов
    - proteins.fasta -- аминокислотные последовательности всех предсказанных генов
    - scaffolds.hits\_from\_MIL\_1.txt -- информация о схожести белков из нашей бактерии с белками из бактерии MIL-1
    - scaffolds.hits\_from\_SwissProt.txt -- информация о схожести белков из нашей бактерии с белками из БД SwissProt
- В репозитории в папке src – любой другой код, который был использован для выполнения задания

## Форма отчетности

Github репозиторий, содержащий все полученные результаты.

**Последний срок сдачи: среда, 19 октября до 23:59 (будет отслеживаться по последнему коммиту в репозиторий). Штраф -0.5 балла за каждый день просрочки.**

В случае возникновения вопросов обращаться по каналам связи:

tg: Iv\_sk

Почта: isgevorkov@edu.hse.ru