Выполненное ДЗ присылать на почту [dmitrii.lv.konovalov@gmail.com](mailto:dmitrii.lv.konovalov@gmail.com)

По вопросам писать в telegram +7-915-455-46-41

Дедлайн: 18:10 10.02.22

За каждый день после дедлайна штраф -1 балл.

ДЗ пересдать нельзя.

1. GenBank и EMBOSS
   1. Найти в Genebank и скачать полный геном бактерии**\*** в формате FASTA.
   2. Установить [EMBOSS](http://emboss.sourceforge.net/download/)**\*\***
   3. Разрезать геном на 10 частей. (команда **splitter**)
   4. Посчитать число слов AAAA, ATAT, ATTA, AATT (команда **compseq**)
   5. Перемешать каждую из 10 частей (команда **shuffleseq)**
   6. Посчитать число слов AAAA, ATAT, ATTA, AATT в перемешанных последовательностях
   7. Сравнить результаты 1.4 и 1.6 друг с другом и с ожидаемым в соответствии с моделью Бернулли.
2. Uniprot
   1. Сколько результатов найдено в UniprotKB по запросу в виде названия белка**\*\*\***? В Swiss-Prot? В TrEMBL?
   2. Сколько результатов найдено при использовании расширенного поиска в поле protein name? В Swiss-Prot? В TrEMBL?
   3. Сколько результатов остается при добавлении фильтра Homo sapiens в поле Taxonomy? В Swiss-Prot? В TrEMBL?

Откройте запись о вашем белке и ответьте на следующие вопросы:

* 1. Какова функция белка?
  2. К какому семейству он принадлежит?
  3. К скольки кластерам UniRef с идентичностью 1.0, 0.9, 0.5 принадлежит белок (включая изоформы)?

**\*** в таблице к ДЗ указаны accession numbers бактерий

**\*\*** Для установки на Windows может понадобится установка [Java 32bit](https://java.com/ru/download/manual.jsp)

**\*\*\*** белки указаны в таблице к ДЗ