Использование нейронных сетей для распознавания 16s РНК по вторичной структуре

Автор: Лунина Полина Сергеевна, 344 группа
Научный руководитель: доцент, к.ф-м.н. Григорьев С.В.

Санкт-Петербургский государственный университет
Кафедра системного программирования

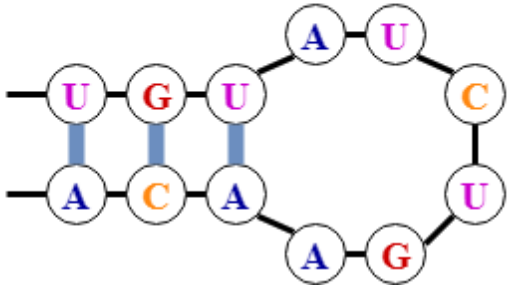
22 мая 2018г.

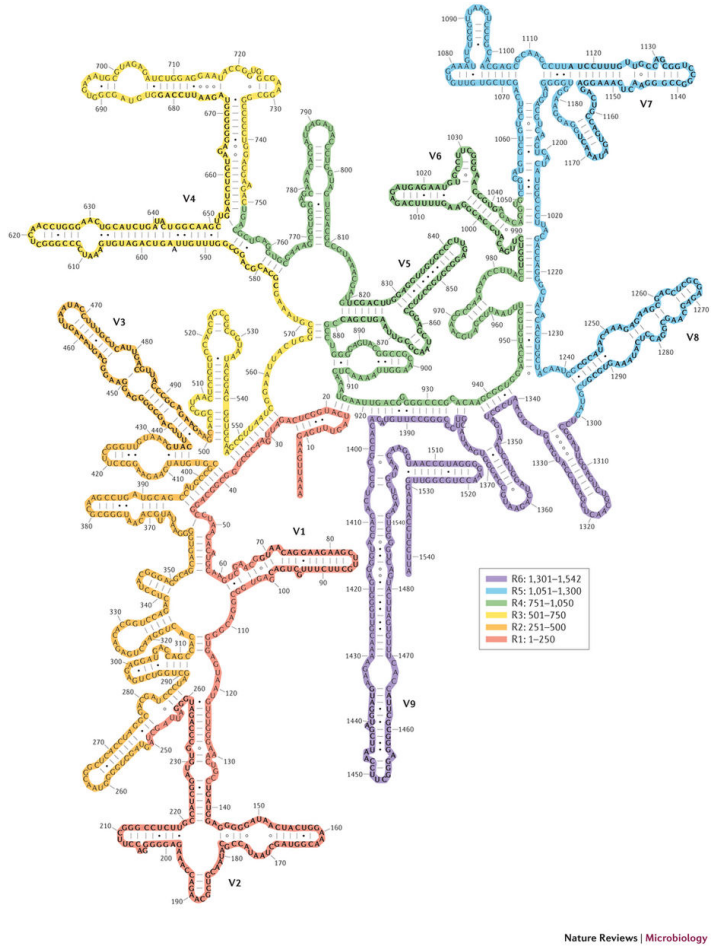
16s РНК

- Первичная структура — последовательность нуклеотидов
- Вторичная структура может быть задана грамматикой

```
[<Start>]
s1: stem<s0>
s0: A U C U G A
stem<s>:
      A s U
    | G s C
    | U s A
    | C s G
```

...UGUAUCUGAACA...





Цель и задачи

Цель:
исследование возможности распознавания бактерий на основе данных о вторичной структуре их 16s РНК, полученных методами синтаксического анализа, с помощью машинного обучения.

Задачи:

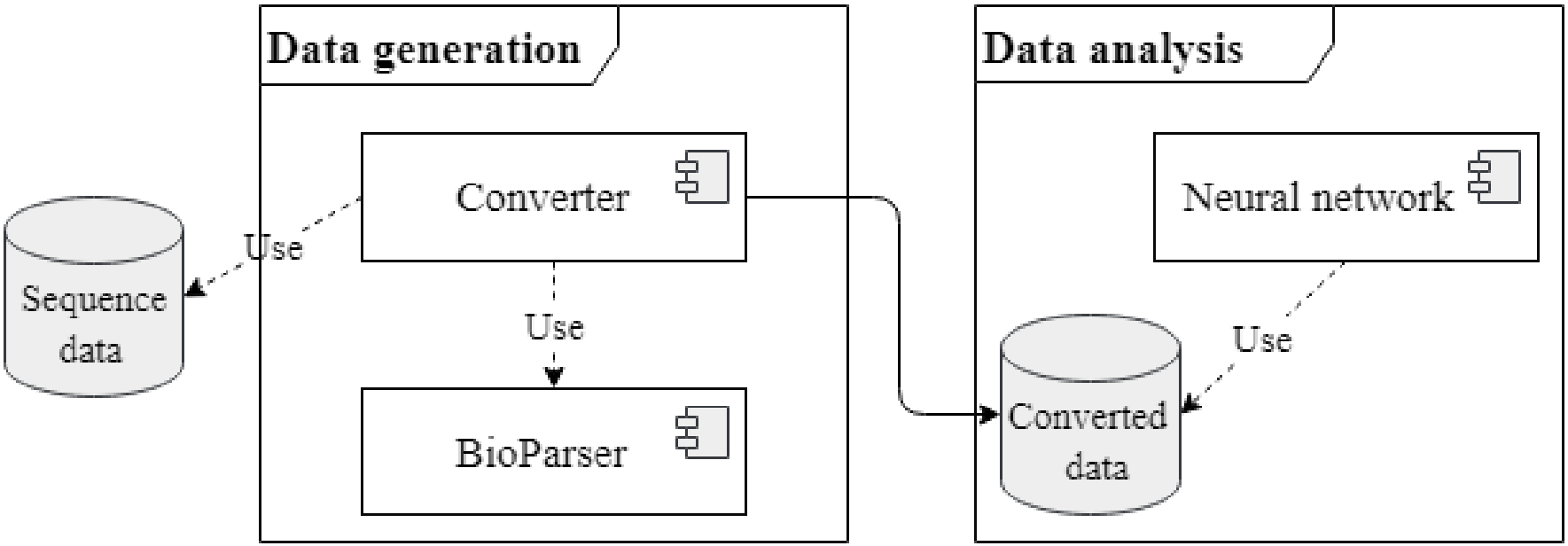
- Разработка архитектуры решения
- Выбор формата представления данных о вторичной структуре бактерий и реализация процесса их генерации
- Создание нейронной сети для распознавания 16s РНК среди прочих нуклеотидных последовательностей
- Экспериментальные исследования и анализ полученных результатов

Похожие решения

- BLAST — поиск гомологов белков и нуклеиновых кислот на основе их первичной структуры
- HMMER — моделирование вторичной структуры с использованием теории скрытых марковских моделей
- Infernal — моделирование вторичной структуры с помощью вероятностных моделей и стохастических контекстно-свободных грамматик
- Humidor — классификация данных в формате CIGAR strings с помощью сверточных нейронных сетей

Архитектура решения

- Платформа YaccConstructor
- Библиотека Keras



Генерация данных

- Имеется последовательность нуклеотидов и некоторая грамматика
- С помощью алгоритма синтаксического анализа получается лес разбора, по которому строится матрица
- Матрица линеаризуется и преобразовывается в числовой вектор

Нейронная сеть

- Keras, TensorFlow
- Dense и Dropout слои
- Данные из БД SILVA

Эксперименты

- Accuracy = 0.87
- Precision = 0.96
- Recall = 0.77
- Specificity = 0.97
- Всего 35706 образцов
- 7345 тестовых образцов

	classified as positive	classified as negative
positive	2789	856
negative	108	3592

Результаты

- Разработана архитектура решения
- Реализован процесс генерации данных в виде числовых векторов
- Создана нейронная сеть для распознавания 16s РНК бактерий по данным о вторичной структуре
- Проведены экспериментальные исследования на участках геномов бактерий из базы данных SILVA