Комбинирование нейронных сетей и синтаксического анализа для обработки вторичной структуры последовательностей

Автор: Лунина Полина Сергеевна, 444 группа
Научный руководитель: доцент, к.ф-м.н. Григорьев С.В.
Рецензент: специалист по анализу данных
ООО "Интеллоджик" Малыгина Т.С.

Санкт-Петербургский государственный университет
Кафедра системного программирования

24 мая 2019г.

Введение

- Анализ последовательностей, обладающих синтаксической структурой
- Существующие подходы:
  - N-граммы
  - Скрытые марковские модели
  - Ковариационные модели
  - Вероятностные грамматики

Постановка задачи

Цель — разработка подхода для анализа вторичной структуры последовательностей с использованием комбинации синтаксического анализа и нейронных сетей

Задачи:

- Разработать архитектуру решения, независимую от конкретной области применения и используемых технологий
- Провести экспериментальные исследования

Описание подхода

- Описание характерных особенностей вторичной структуры исследуемых последовательностей с помощью грамматики
- Извлечение этих особенностей с помощью алгоритма синтаксического анализа
- Обучение нейронных сетей на полученных данных для решения конкретной задачи

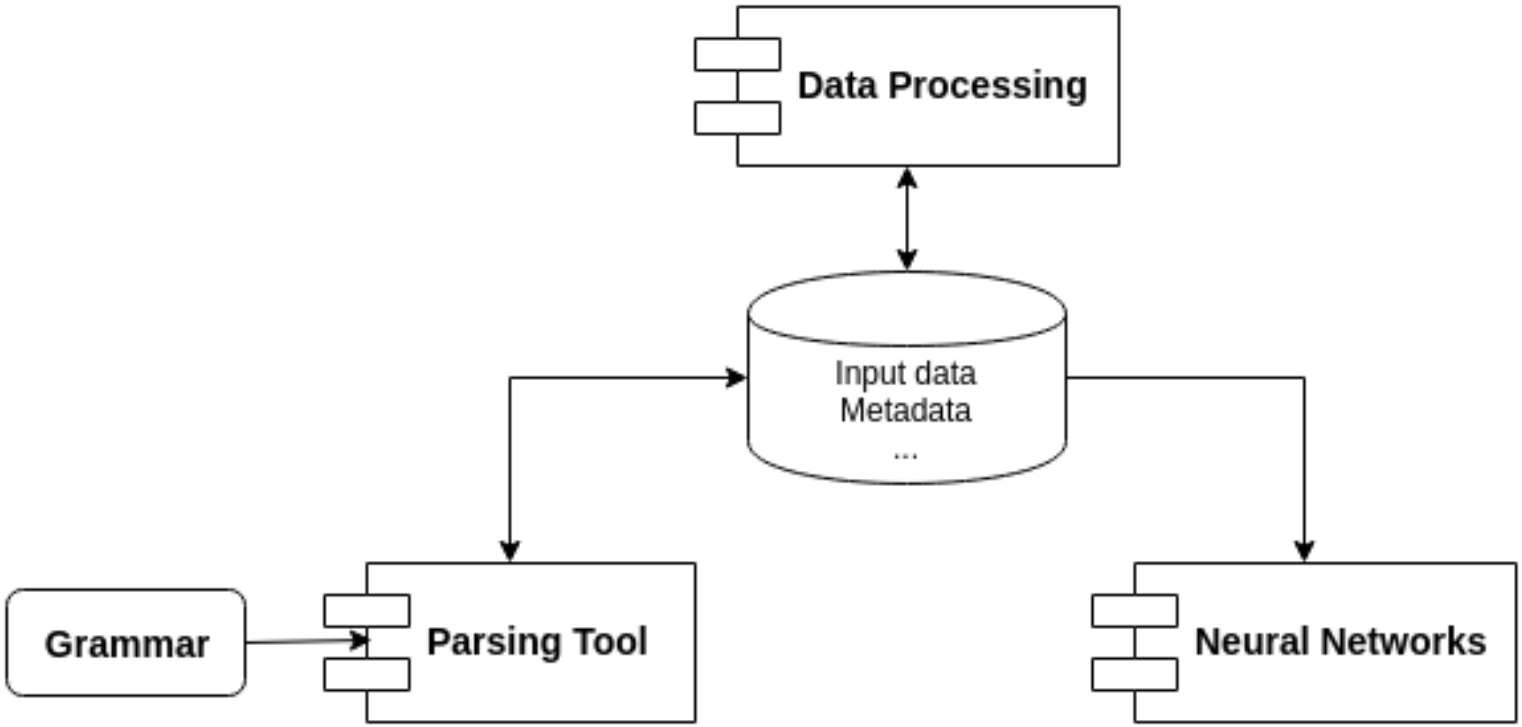
Генерация данных с помощью синтаксического анализатора

- Входные данные — грамматика и исследуемые последовательности
- Результат работы для входной строки w и нетерминала N — верхнетреугольная булева матрица разбора M_N, где $M_N[i, j] = 1$, если подстрока $w[i, j-1]$ выводима из N
- Форматы выходных данных:
  - Числовые вектора
  - Черно-белые изображения

Обучение нейронных сетей

- Обучение нейронной сети на векторах или изображениях
- Для использования обученной модели необходимо снова использовать синтаксический анализатор
- Проблема: времязатратность синтаксического анализа
- Решение: расширить обученную нейронную сеть верхними слоями, принимающими исходную последовательность

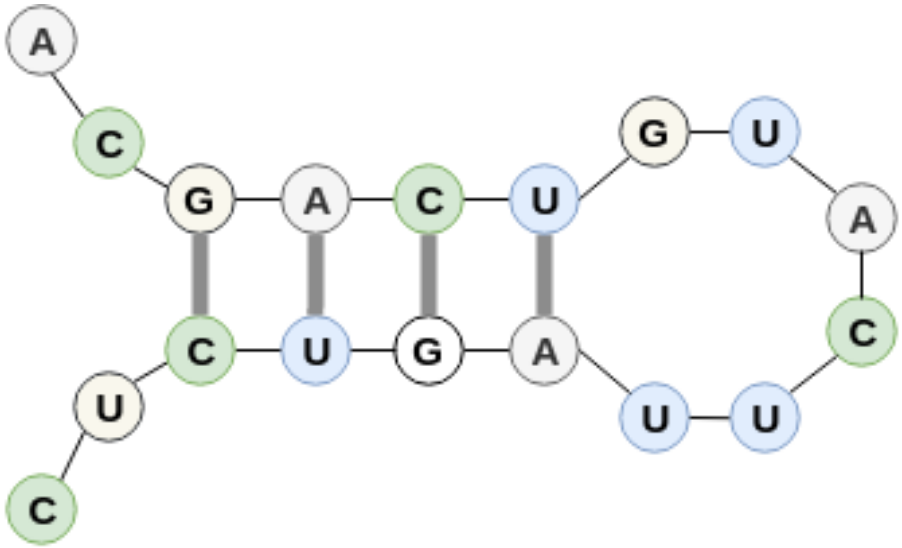
Архитектура решения



Биоинформатика — область применения

- Задачи анализа нуклеотидных и аминокислотных последовательностей
- Шпильки вторичной структуры РНК можно описать грамматикой

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



Эксперименты

Задачи:

- Классификация тРНК эукариотов и прокариотов
- Классификация тРНК архей, бактерий, грибов и растений

Технологии:

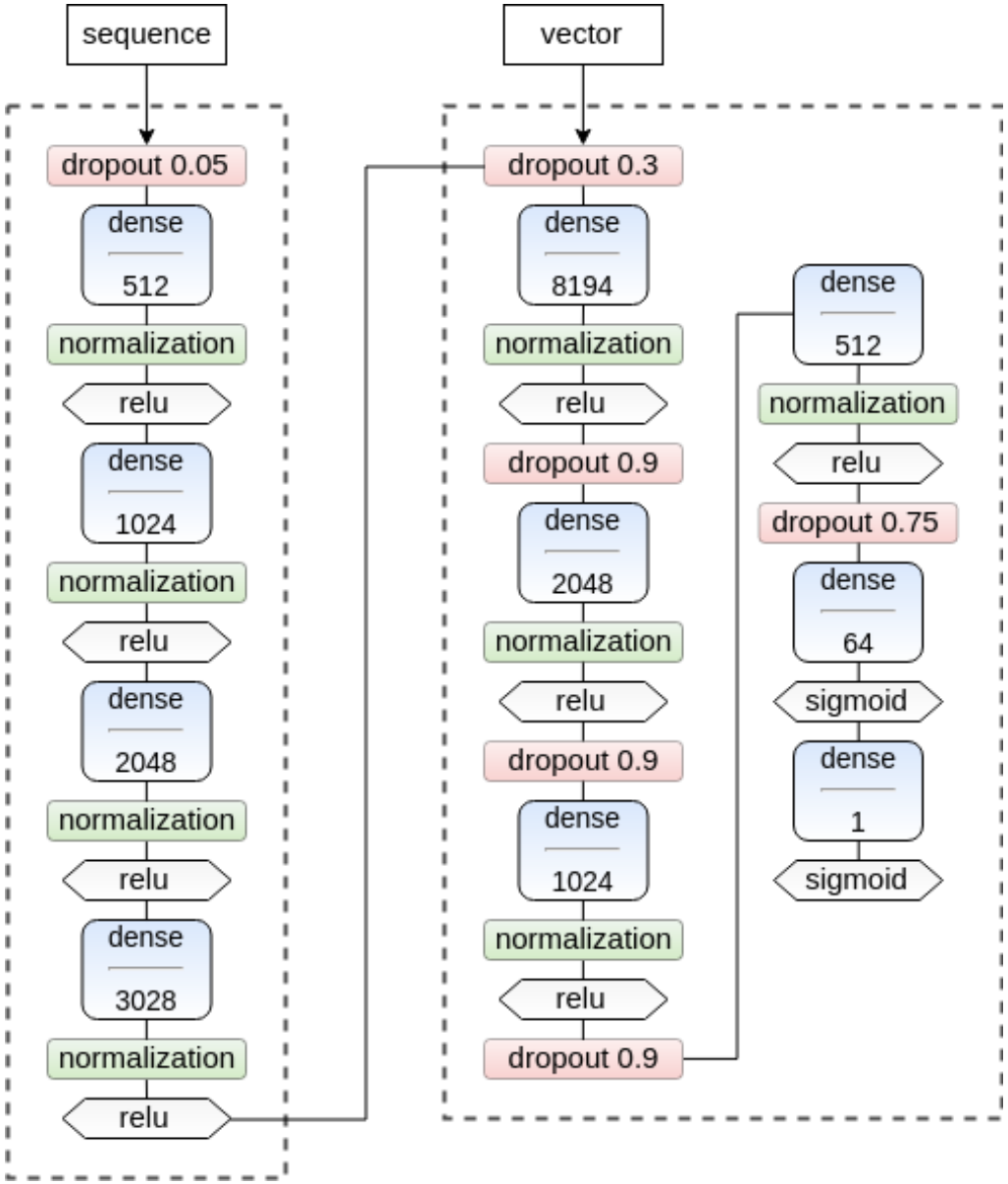
- Платформа YaccConstructor, разработанная на кафедре системного программирования СПбГУ
- Библиотека Keras и фреймворк Tensorflow

Базы данных:

- tRNADB-CE
- Genomic tRNA database

Эксперименты

- Обучение нейронных сетей на векторах и на изображениях
- Обучение на их основе нейронных сетей, принимающих исходные последовательности
- Тестирование и сравнение результатов



Классификация тРНК: эукариоты и прокариоты

train:valid:test = 20000:5000:10000

accuracy	vector-based approach	image-based approach
base model	94.1%	96.2%
extended model	97.5%	97.8%

class	vector-based approach		image-based approach	
	precision	recall	precision	recall
prokaryotic	95.8%	99.4%	96.2%	99.4%
eukaryotic	99.4%	95.6%	99.4%	99.5%

Классификация тРНК: археи, бактерии, растения и грибы

train:valid:test = 8000:1000:3000

accuracy	vector-based approach	image-based approach
base model	86.7%	93.3%
extended model	96.2%	95.7%

class	vector-based approach		image-based approach	
	precision	recall	precision	recall
archaeal	91.1%	99.2%	91.6%	98.5%
bacterial	96.6%	95.1%	95.2%	95.5%
fungi	98.5%	94.9%	97.5%	94.3%
plant	99.4%	95.7%	99.2%	94.7%

Результаты

- Разработана архитектура решения для использования предложенного подхода
- Проведены экспериментальные исследования предложенного подхода применительно к следующим задачам биоинформатики:
  - Классификация тРНК эукариотов и прокариотов
  - Классификация тРНК архей, бактерий, грибов и растений
- Результаты представлены на конференциях:
  - Постер "16s rRNA Detection by Using Neural Networks"на конференции Biata 2018
  - Статья "The Composition of Dense Neural Networks and Formal Grammars for Secondary Structure Analysis"на конференции BIOINFORMATICS 2019
  - Постер "Improved Architecture of Artificial Neural Network for Secondary Structure Analysis"на конференции Biata 2019