



Комбинирование нейронных сетей и синтаксического анализа для обработки вторичной структуры последовательностей

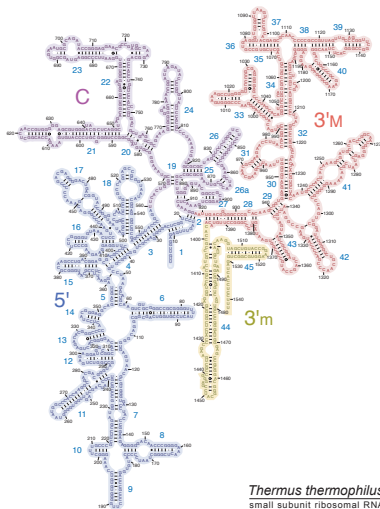
Полина Лунина

JetBrains Research, Programming Languages and Tools Lab
Санкт-Петербургский государственный университет

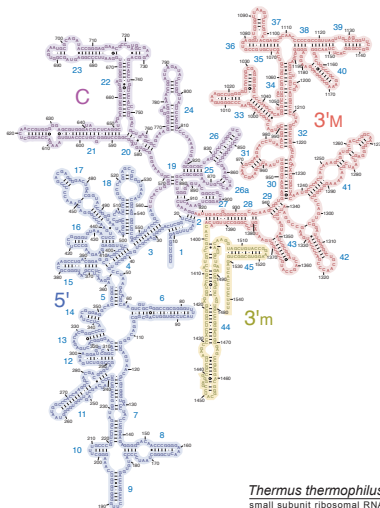
14 декабря 2019г.

• Задачи

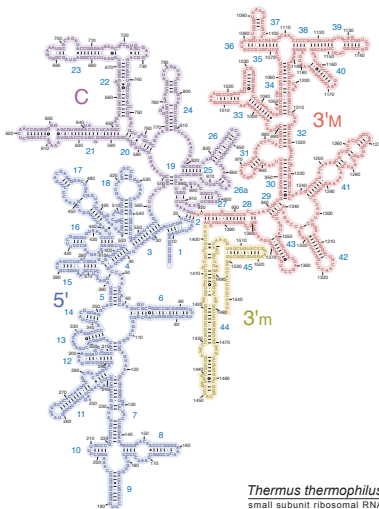
- ▶ Распознавание
- ▶ Классификация
- ▶ Предсказание
вторичных структур
- ▶ ...



- Задачи
 - ▶ Распознавание
 - ▶ Классификация
 - ▶ Предсказание вторичных структур
 - ▶ ...
- Формальное задание вторичной структуры

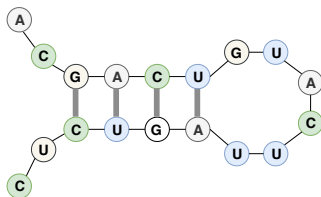


- Задачи
 - ▶ Распознавание
 - ▶ Классификация
 - ▶ Предсказание вторичных структур
 - ▶ ...
- Формальное задание вторичной структуры
- Вероятностная оценка



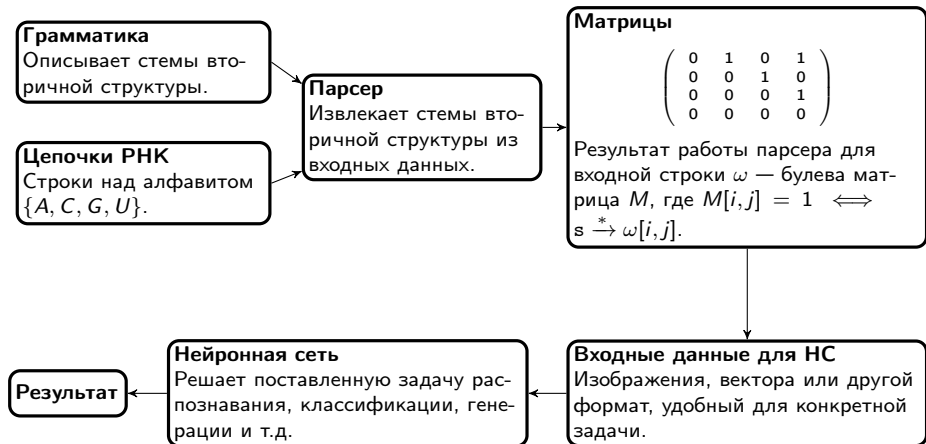
Наш подход

- Задать основные элементы вторичной структуры (стеми) с помощью грамматики
- Для вероятностной оценки использовать нейронные сети

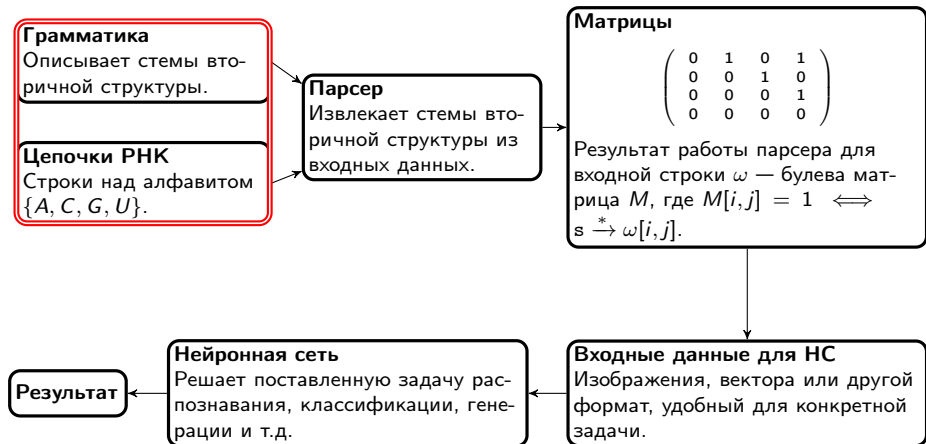


```
s1: stem<s0>
s0: G U A C U U
stem<s>:
    A s U
    I G s C
    I U s A
    I C s G
```

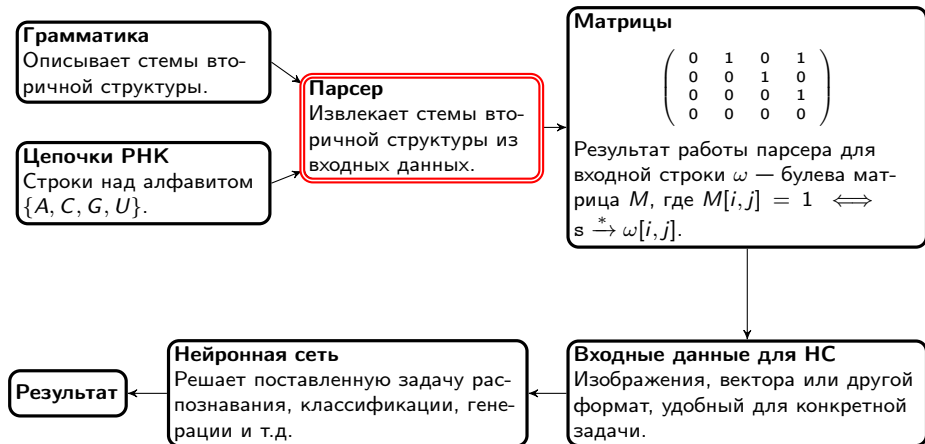
Структура решения



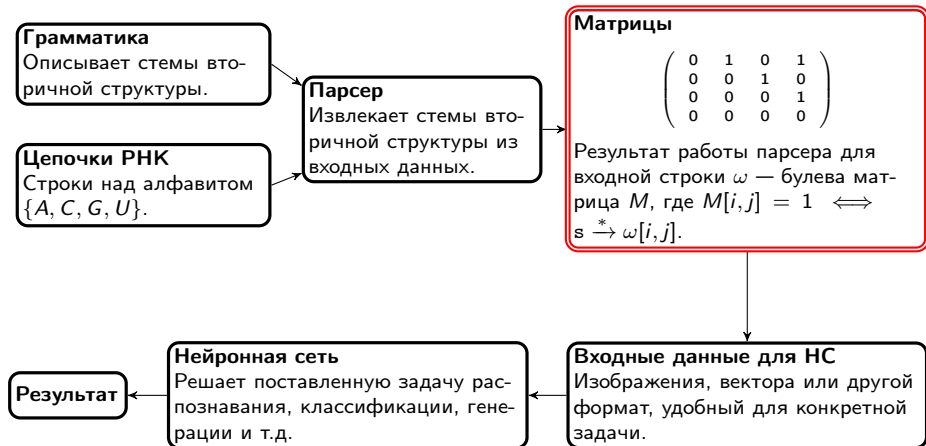
Структура решения



Структура решения



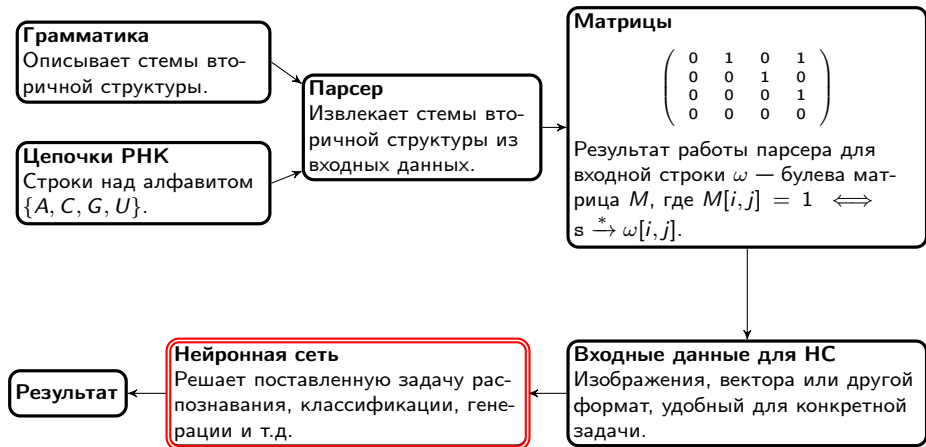
Структура решения



Структура решения



Структура решения



Структура решения



Проблема: времязатратность синтаксического анализа

Решение:

- Создать нейронную сеть, обрабатывающую непосредственно цепочку РНК
- Обучение в 2 этапа
 - ▶ Обучить нейронную сеть на результате работы парсера
 - ▶ Расширить ее верхними слоями, принимающими исходную последовательность

Задачи:

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

Технологии:

- Платформа YaccConstructor
- Библиотека Keras и фреймворк Tensorflow

Базы данных:

- tRNADB-CE
- Genomic tRNA database

- Два формата представления матриц
 - ▶ Вектора
 - ▶ Черно-белые изображения
- Обучение нейронных сетей на этих данных
- Обучение нейронных сетей, принимающих непосредственно последовательности РНК и использующих веса предыдущих моделей

Результаты

EP — эукариоты/прокариоты

ABFP — археи/бактерии/растения/грибы

Classifier	EP		ABFP	
Approach	Vector-based	Image-based	Vector-based	Image-based
Base model accuracy	94.1%	96.2%	86.7%	93.3%
Extended model accuracy	97.5%	97.8%	96.2%	95.7%
Samples for train:valid:test	20000:5000:10000 (57%:14%:29%)		8000:1000:3000 (67%:8%:25%)	

Результаты

EP — эукариоты/прокариоты

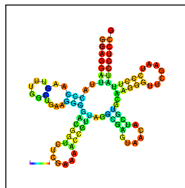
ABFP — археи/бактерии/растения/грибы

Classifier	Class	Vector-based approach		Image-based approach	
		precision	recall	precision	recall
EP	prokaryotic	95.8%	99.4%	96.2%	99.4%
	eukaryotic	99.4%	95.6%	99.4%	99.5%
ABFP	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%

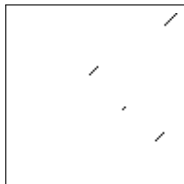
- Парсер находит в цепочке все возможные стемы, однако не все они действительно будут входить в состав вторичной структуры
- Хотим сконструировать нейронную сеть, которая отфильтрует лишние контакты между нуклеотидами и предскажет вторичную структуру цепочки

Текущее исследование

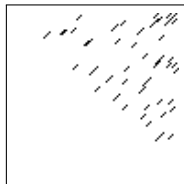
- Парсер находит в цепочке все возможные стемы, однако не все они действительно будут входить в состав вторичной структуры
- Хотим сконструировать нейронную сеть, которая отфильтрует лишние контакты между нуклеотидами и предскажет вторичную структуру цепочки



Вторичная структура



Contact map



Результат парсера

Задача: предсказание вторичных структур цепочек тРНК длины 90

Данные:

- RNACentral (последовательности тРНК)
- CentroidFold (эталонные структуры)

Результаты тестирования на 11000 образцов:

- Precision = 84% (сколько из предсказанных контактов действительно являются контактами в эталоне)
- Recall = 89% (сколько из требуемых контактов было найдено)

- Предсказание вторичных структур для цепочек различных РНК любой длины
- Улучшение точности результата путем увеличения количества данных и настройки параметров нейронной сети
- Выбор оптимального источника эталонных данных
 - ▶ Лучшие результаты на бенчмарках
 - ▶ Возможность предсказания псевдоузлов

- **BIOINFORMATICS-2019**

- ▶ Семён Григорьев, Полина Лунина. The Composition of Dense Neural Networks and Formal Grammars for Secondary Structure Analysis
- ▶ Публикация: Scopus

- **VIATA-2019** (постерный доклад)

- ▶ Семён Григорьев, Полина Лунина. Improved Architecture of Artificial Neural Network for Secondary Structure Analysis
- ▶ Публикация: BMC Bioinformatics, Scopus

- **CIBB-2019** (доклад)

- ▶ Полина Лунина, Семён Григорьев. On Secondary Structure Analysis by Using Formal Grammars and Artificial Neural Networks
- ▶ Публикация: ожидается