



Комбинирование нейронных сетей и синтаксического анализа для предсказания вторичных структур генетических цепочек

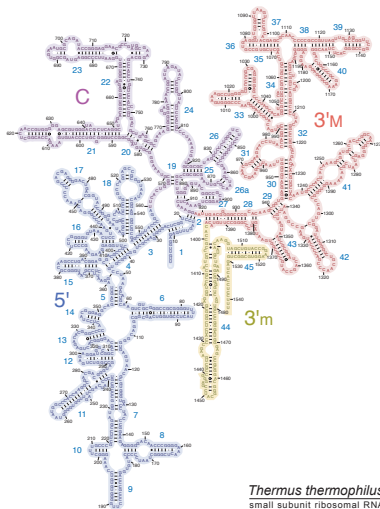
Лунина Полина Сергеевна, 571 группа
Научный руководитель: доцент, к.ф-м.н. Григорьев С.В.

Санкт-Петербургский государственный университет
Кафедра системного программирования

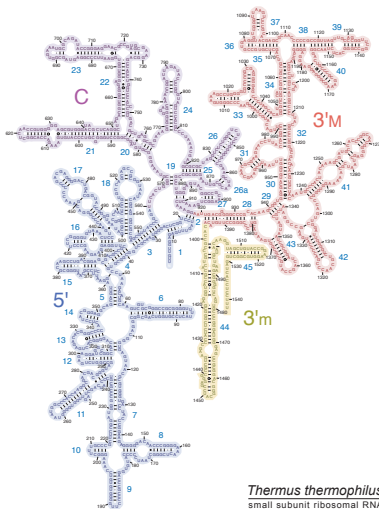
10 июня 2020г.

• Задачи

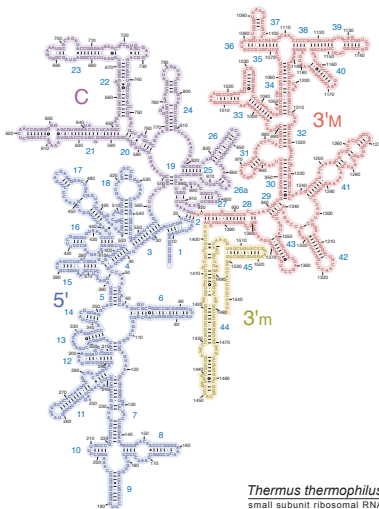
- ▶ Распознавание
- ▶ Классификация
- ▶ Предсказание вторичных структур
- ▶ ...



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

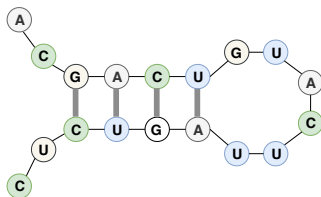


- Задачи
 - ▶ Распознавание
 - ▶ Классификация
 - ▶ Предсказание вторичных структур
 - ▶ ...
- Формальное задание вторичной структуры
- Вероятностная оценка



Наш подход

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

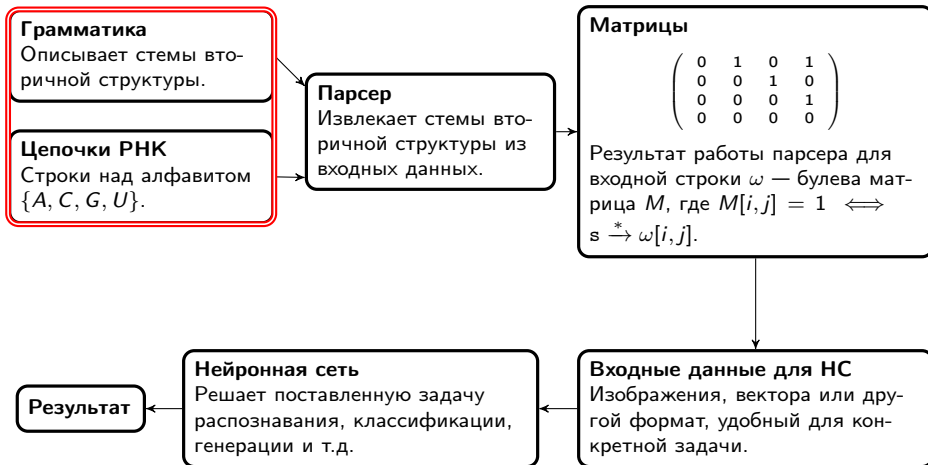


```
s1: stem<s0>  
s0: G U A C U U  
stem<s>:  
    A s U  
    I G s C  
    I U s A  
    I C s G
```

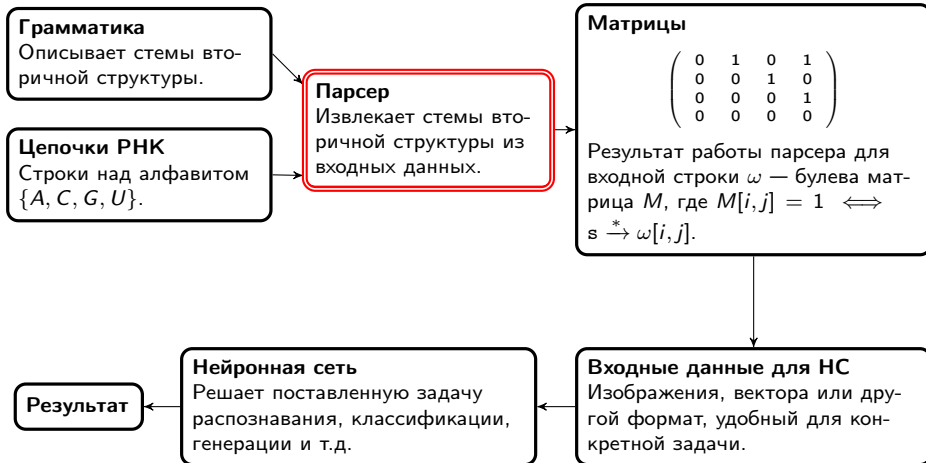
Наш подход



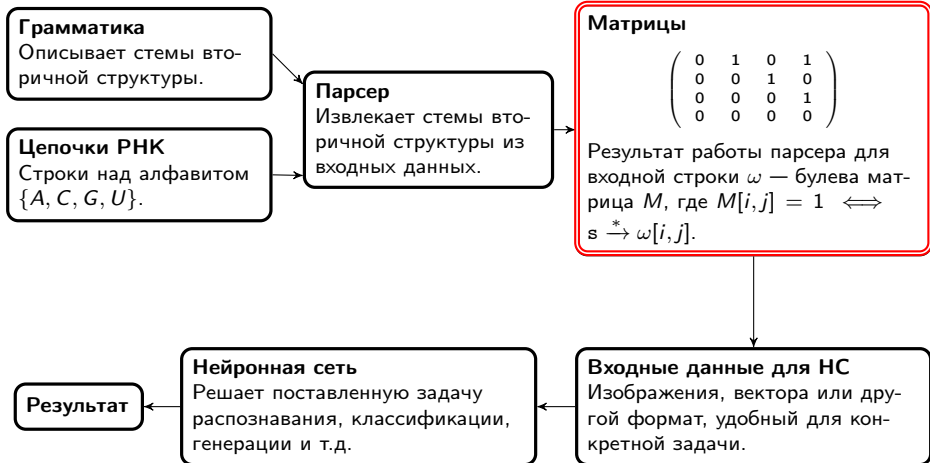
Наш подход



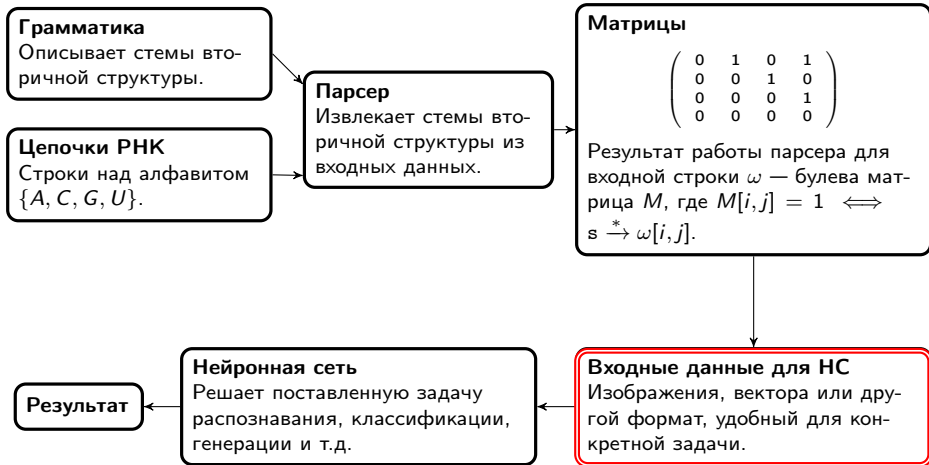
Наш подход



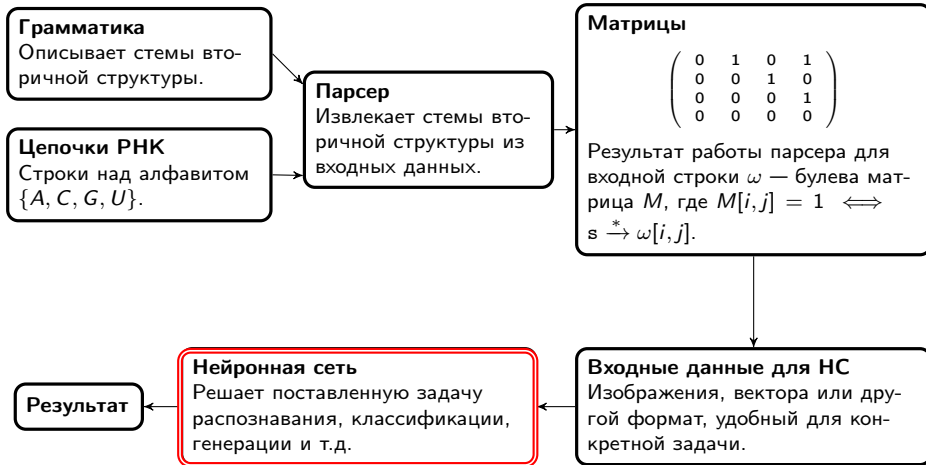
Наш подход



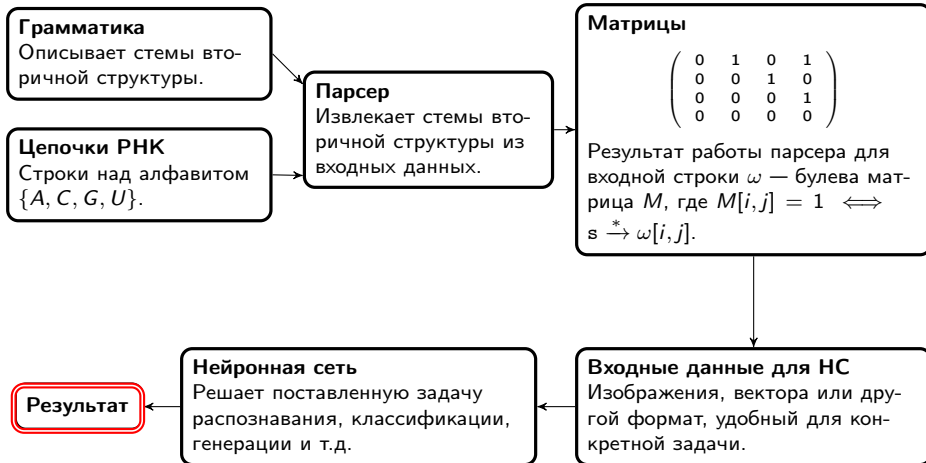
Наш подход



Наш подход



Наш подход

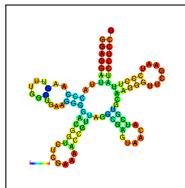


Идея текущего исследования

- 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

Идея текущего исследования

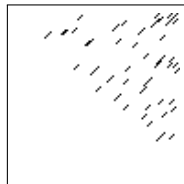
- Парсер находит в цепочке все возможные стемы, однако не все они действительно будут входить в состав вторичной структуры
- Хотим сконструировать нейронную сеть, которая отфильтрует лишние контакты между нуклеотидами и предскажет вторичную структуру цепочки



Вторичная структура



Матрица контактов



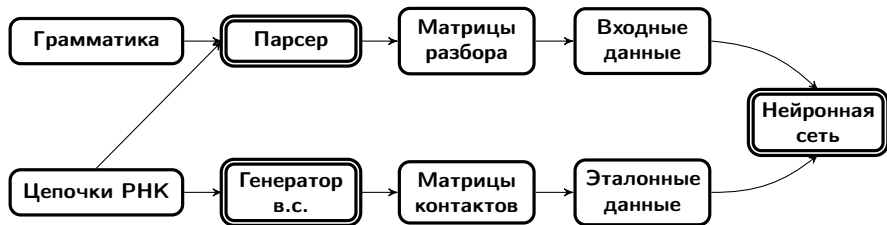
Матрица разбора

Цель — исследование возможности применения предложенного подхода к задаче предсказания вторичных структур геномных последовательностей

Задачи

- Разработка общей архитектуры решения
- Проведение экспериментальных исследований
 - ▶ Предсказание вторичных структур транспортных РНК с различной длиной цепочки
 - ▶ Исследование возможности предсказания псевдоузлов, невыразимых средствами используемой грамматики

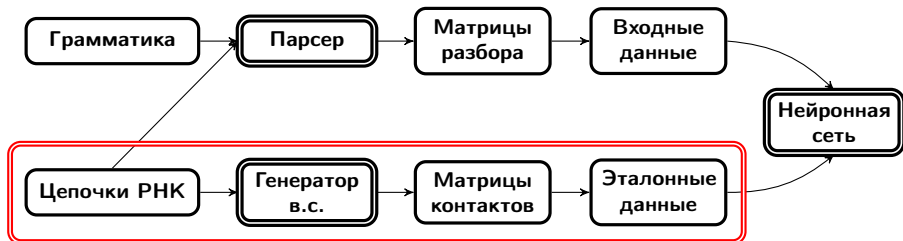
Архитектура решения



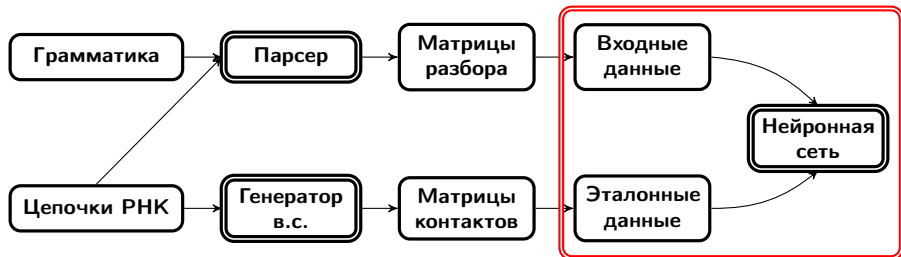
Архитектура решения



Архитектура решения

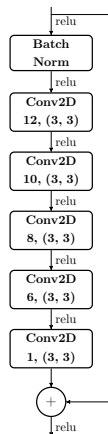


Архитектура решения



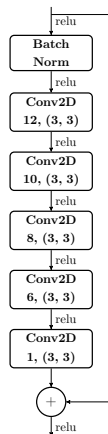
Нейронная сеть

- Convolutional residual neural network
 - ▶ 10 residual блоков
 - ▶ После них — слой, эмулирующий алгоритм локального выравнивания последовательностей
- Loss — взвешенная попиксельная разница
- train/valid/test = 70%/10%/20%



Нейронная сеть

- Convolutional residual neural network
 - ▶ 10 residual блоков
 - ▶ После них — слой, эмулирующий алгоритм локального выравнивания последовательностей
- Loss — взвешенная попиксельная разница
- train/valid/test = 70%/10%/20%
- Метрики
 - ▶ *Precision* — сколько из предсказанных контактов действительно являются контактами в эталоне
 - ▶ *Recall* — сколько из требуемых контактов было найдено
 - ▶ *F1 score* — объединяющая метрика



Задачи

- Предсказание вторичных структур цепочек тРНК с различными интервалами длин
- Предсказание вторичных структур цепочек тРНК с псевдоузлами

Данные

- RNACentral (последовательности тРНК)
- CentroidFold (эталонные структуры)
- Pseudobase (цепочки и эталонные структуры с псевдоузлами)

Технологии

- Платформа YaccConstructor
- Библиотека Keras и фреймворк Tensorflow

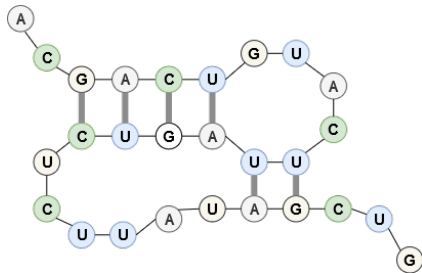
Предсказание вторичных структур для цепочек с разными интервалами длин

Length	Samples	Alignment	Precision	Recall	F1 score
90	26511	×	67%	75%	68%
		✓	80%	66%	70%
88-90	77976	×	66%	78%	69%
		✓	81%	62%	68%
50-90	141835	×	60%	72%	63%
		✓	71%	61%	63%

Средние значения метрик на тестовых выборках

Расширение обученных моделей на данные с псевдоузлами

Псевдоузел состоит из двух шпилек, где половина стебля одной шпильки располагается между двумя половинами стебля другой шпильки (невыразим средствами КС грамматики)



Length	Samples	Alignment	Precision	Recall	F1 score
50-90	266	×	74%	73%	71%

Средние значения метрик на тестовой выборке

- Разработана общая архитектура решения
- Проведены экспериментальные исследования на различных наборах данных
- Подана статья на конференцию Biata-2020

- Предсказание вторичных структур для цепочек различных РНК любой длины
- Выбор оптимального источника эталонных данных
- Более тщательная разработка модели, применяющей адаптивное выравнивание.
- Реализация более развернутой системы тестирования результатов работы нейронных сетей
- Поиск новых средств, а также более тонкая настройка параметров всех моделей для улучшения результатов.