Разработка системы предсказания вторичной структуры РНК с использованием синтаксического анализа и искусственных нейронных сетей

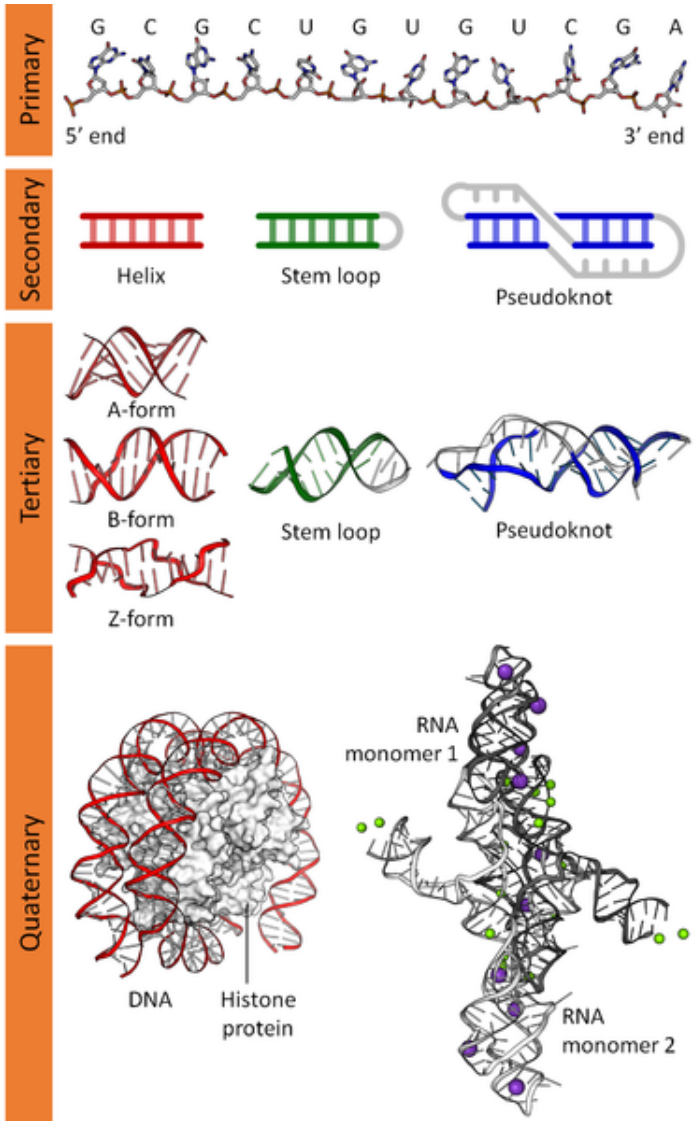
Автор: Кутленков Дмитрий Александрович, 371 группа(17.Б11-мм)
Научный руководитель: к.ф.-м.н., доцент Григорьев С.В.

Санкт-Петербургский государственный университет
Кафедра системного программирования

23 мая 2020г.

Введение

- РНК — биологическая последовательность
- Ее первичная структура — последовательность нуклеотидов, которые задаются алфавитом из 4 букв
- Вторичная структура — то, как нуклеотиды образуют связи



Существующие решения

- Методы сравнительного анализа
- Метод минимальной свободной энергии (MFE) — *RNAfold*, *CentroidFold*, *HotKnots*, *IPknot*
- Иерархическая свертка — *HFold*, *Iterative HFold*
- Исследования с использованием машинного обучения

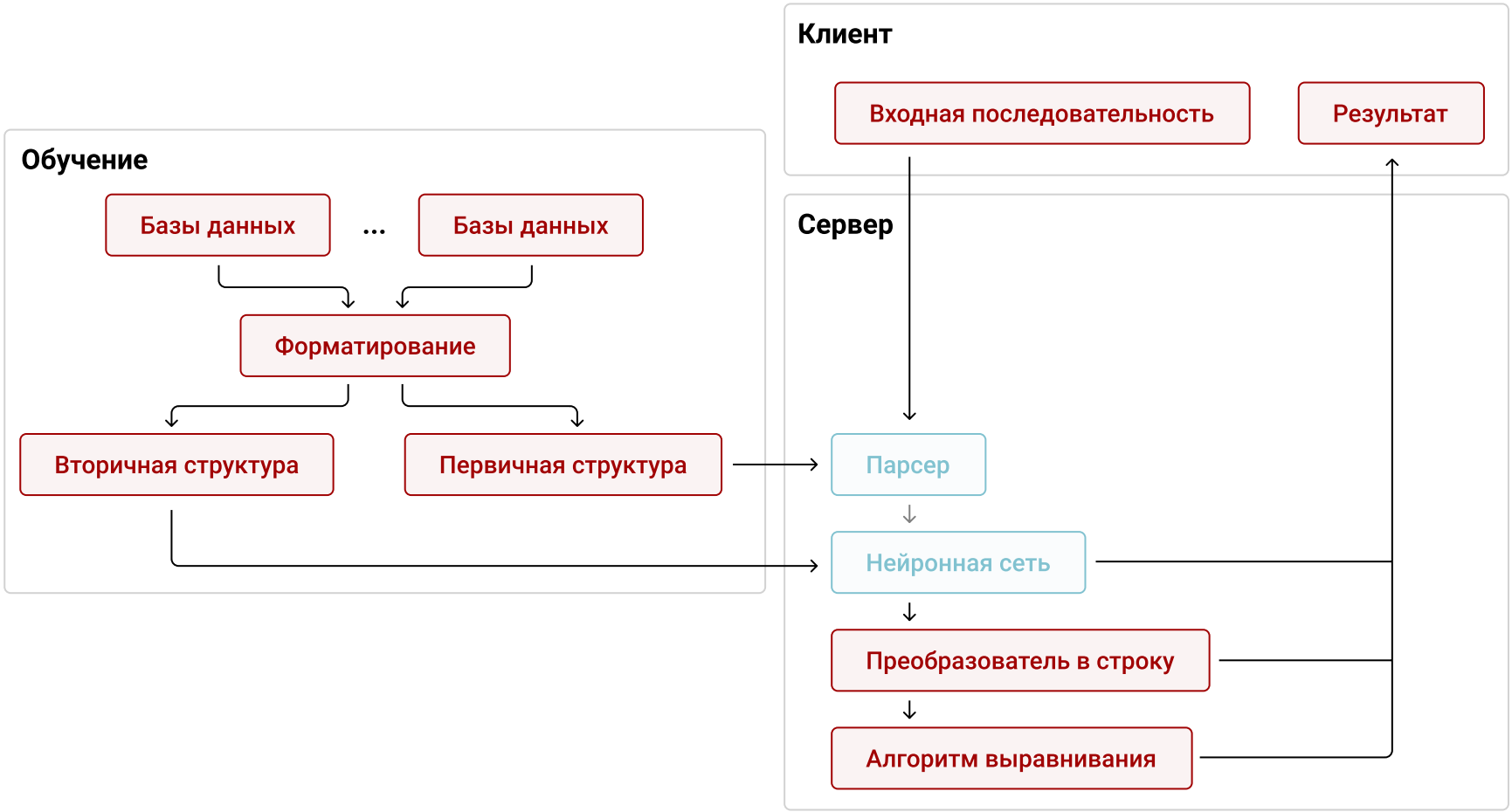
Не существует оптимального метода.

Постановка задачи

Целью данной работы является разработка приложения, способного предсказывать вторичную структуру РНК.
Задачи:

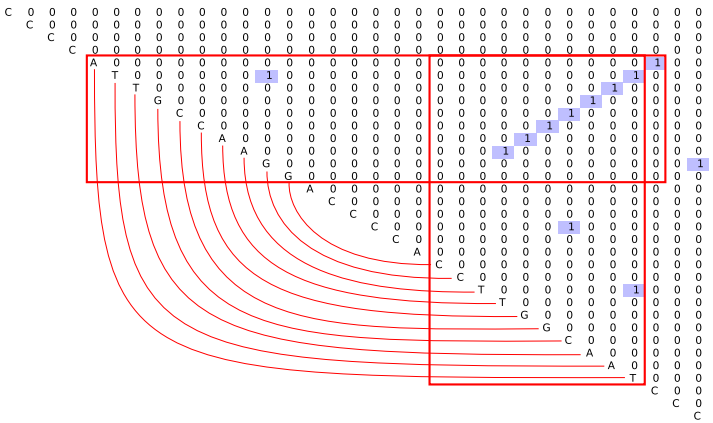
- Создать инструмент для обработки данных. Собрать, проанализировать и обработать с применением созданного инструмента данные для обучения нейронной сети
- Разработать адаптацию алгоритма выравнивания для имеющейся задачи, который будет встроен в итоговое приложение
- Разработать клиент-серверное приложение для предсказания вторичной структуры РНК

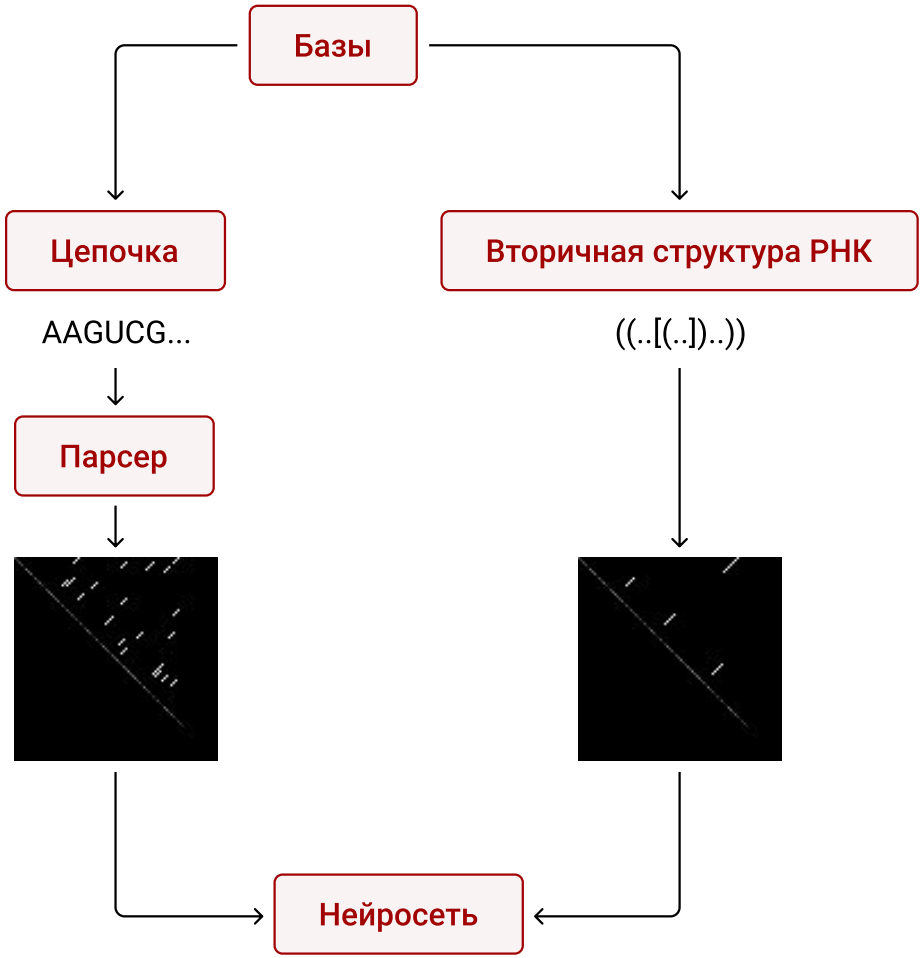
Архитектура всего решения



Подготовка данных

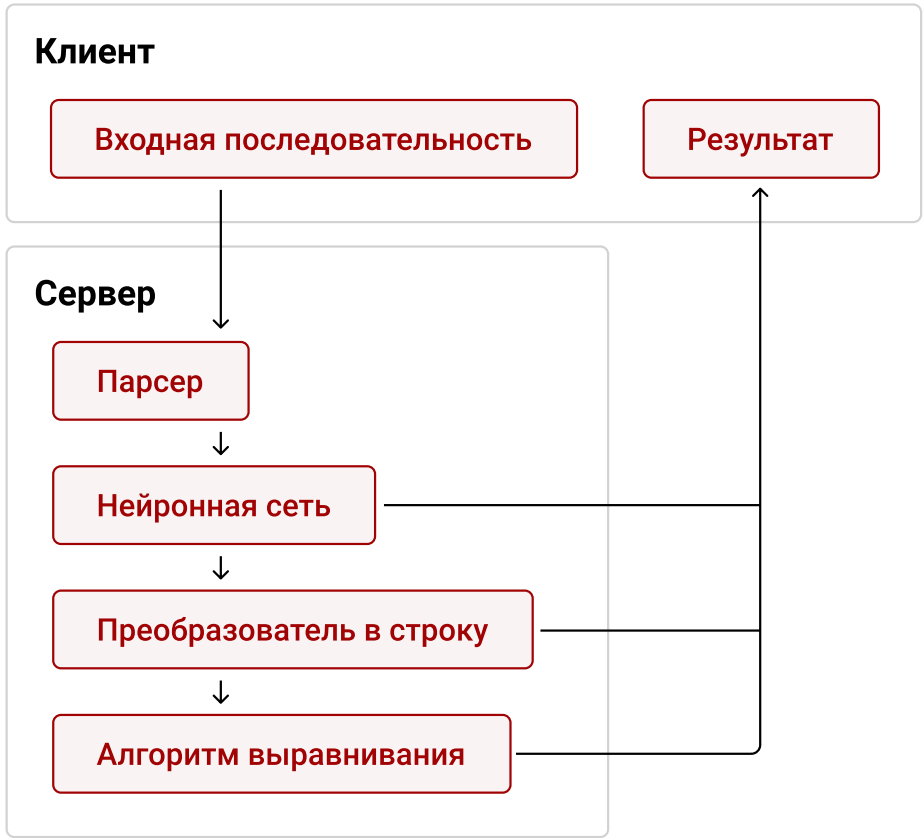
- Парсер — распознает места возможных связей
- Нейросеть — учится очищать результат работы парсера
- Представление данных в виде изображений





Архитектура конечного приложения

- Клиент-серверное приложение
- Пользователь может видеть промежуточные этапы работы системы
- Результат выравнивается, чтобы соответствовать биологическим законам



Система доступна по адресу `http://www.secondarystructure.tk/`

Используемые технологии

- Связь через *REST API*
- Сервер — *Python3*, *Flask*, *Waitress*, *Biopython*
- Клиент — *Bulma.io*, *Vue.js*, *axios*

Результаты

- Создан инструмент для обработки данных для обучения нейронной сети. Собраны, проанализированы и обработаны данные из нескольких источников — *RNA STRAND*, *Pseudobase++*, *RNACentral*
- Разработан алгоритм перевода полученных последовательностей в биологически возможные
- Разработано клиент-серверное приложение, позволяющее предсказывать вторичную структуру РНК последовательностей
- По результатам работы поданы тезисы на постерную секцию международной конференции BiATA 2020