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

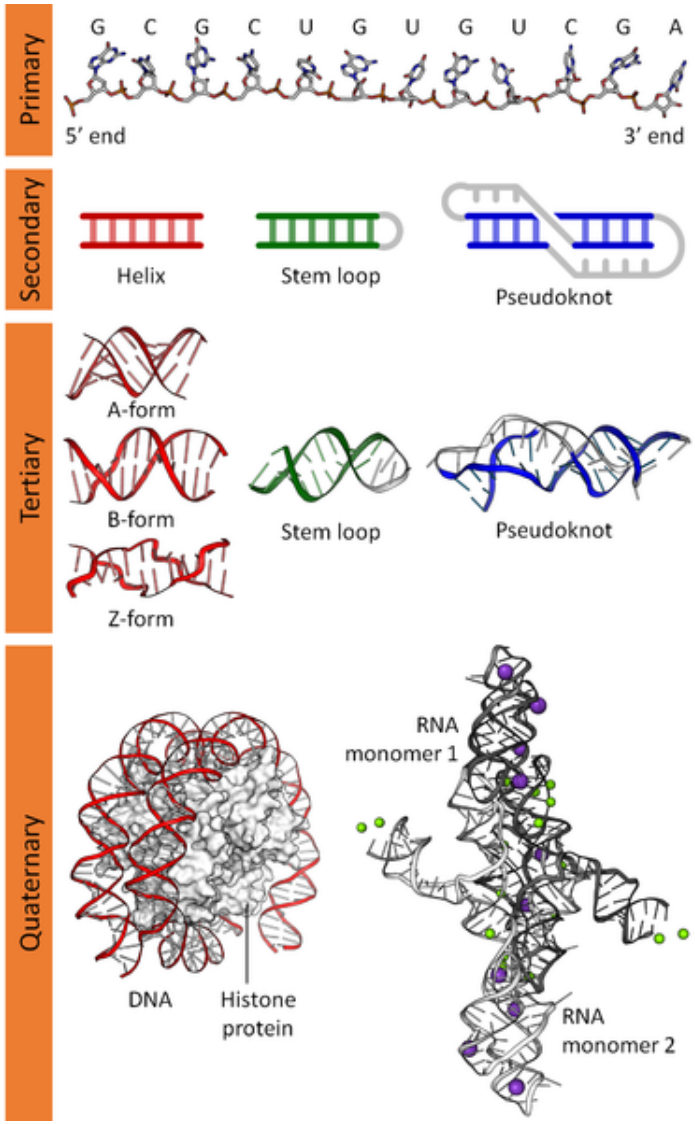
Автор: Кутленков Дмитрий Александрович, 371 группа(17.Б11-мм)
Научный руководитель: к.ф.-м.н., доцент Григорьев С.В.

Санкт-Петербургский государственный университет
Кафедра системного программирования

2 мая 2020г.

Введение

- РНК - биологическая последовательность
- Ее первичная структура - последовательность нуклеотидов, которые задаются алфавитом из 4 букв
- Вторичная структура - то, как нуклеотиды образуют связи
- Псевдоузел - новая петля начинается до конца предыдущей



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

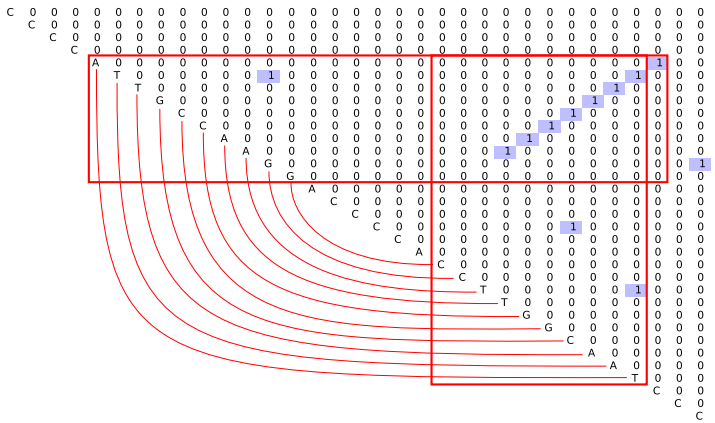
Целью данной работы является разработка системы, способной с достаточной степенью точности предсказывать вторичную структуру РНК

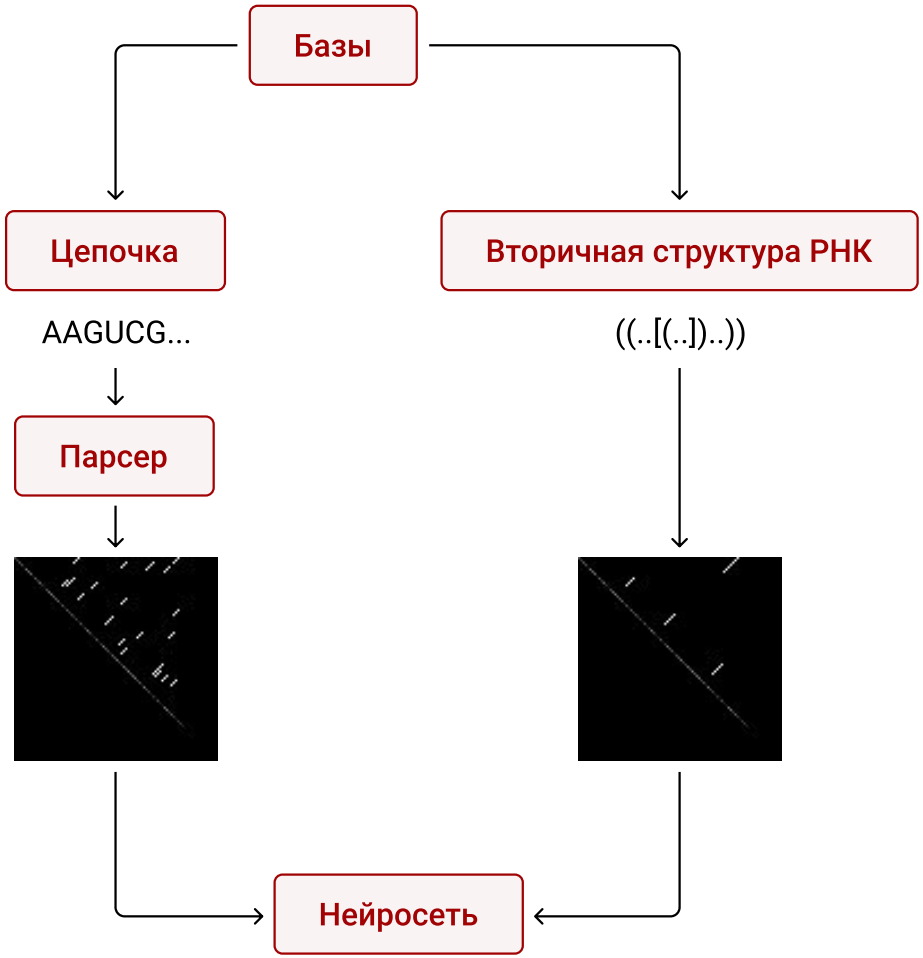
Задачи:

- Изучить предметную область
- Проанализировать существующие решения
- Спроектировать систему на основе формальных грамматик и нейронных сетей
- Собрать и обработать данные для обучения нейронной сети
- Создать систему для подготовки данных
- Обработать результат нейронной сети для получения биологически возможного результата
- Собрать составные части в единую систему, с которой будет удобно работать целевой аудитории, то есть биологам и биоинформатикам

Архитектура системы подготовки данных

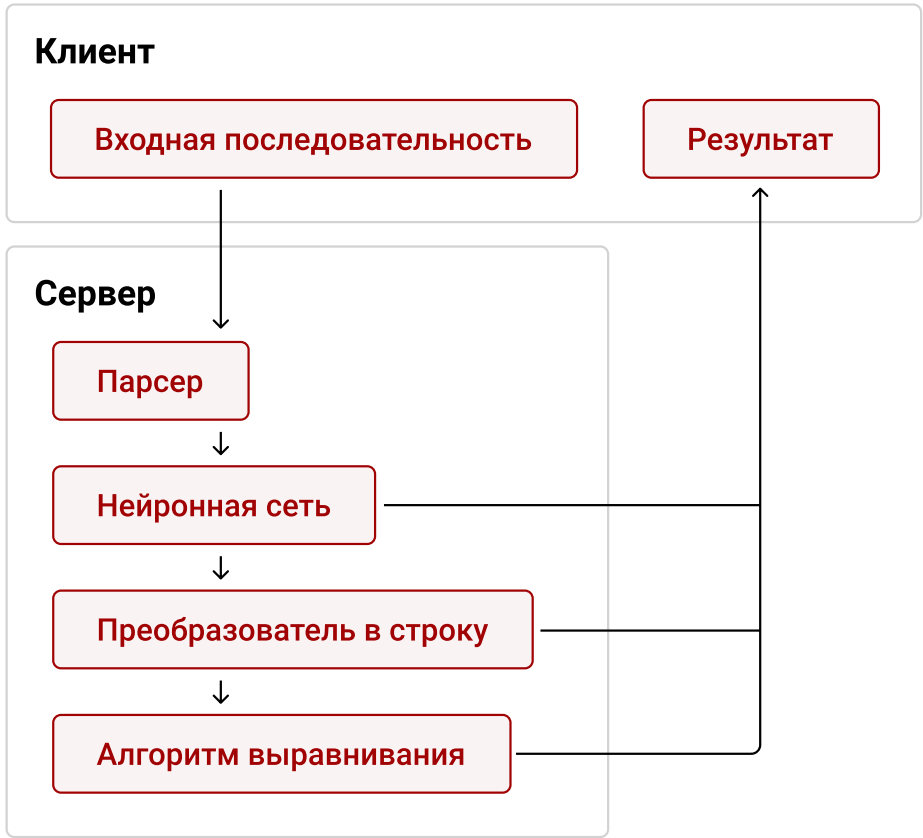
- Парсер - распознает места возможных связей
- Нейросеть - учится очищать результат работы парсера
- Представление данных в виде изображений





Архитектура конечной системы

- 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

- Изучена предметная область
- Проведен анализ уже существующих решений
- Разработана архитектура системы
- Собраны, проанализированы и обработаны данные из нескольких источников - *RNA STRAND*, *Pseudobase++*, *RNACentral*
- Создана система подготовки данных
- Разработан алгоритм перевода полученных последовательностей в биологически возможные
- Разработана система предсказания вторичной структуры РНК последовательностей
- Создано клиент-серверное приложение, предоставляющее доступ к системе