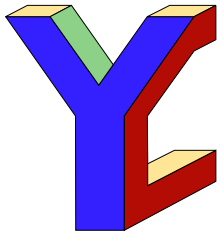


Формальные грамматики и искусственные нейронные сети для предсказания вторичной структкры РНК

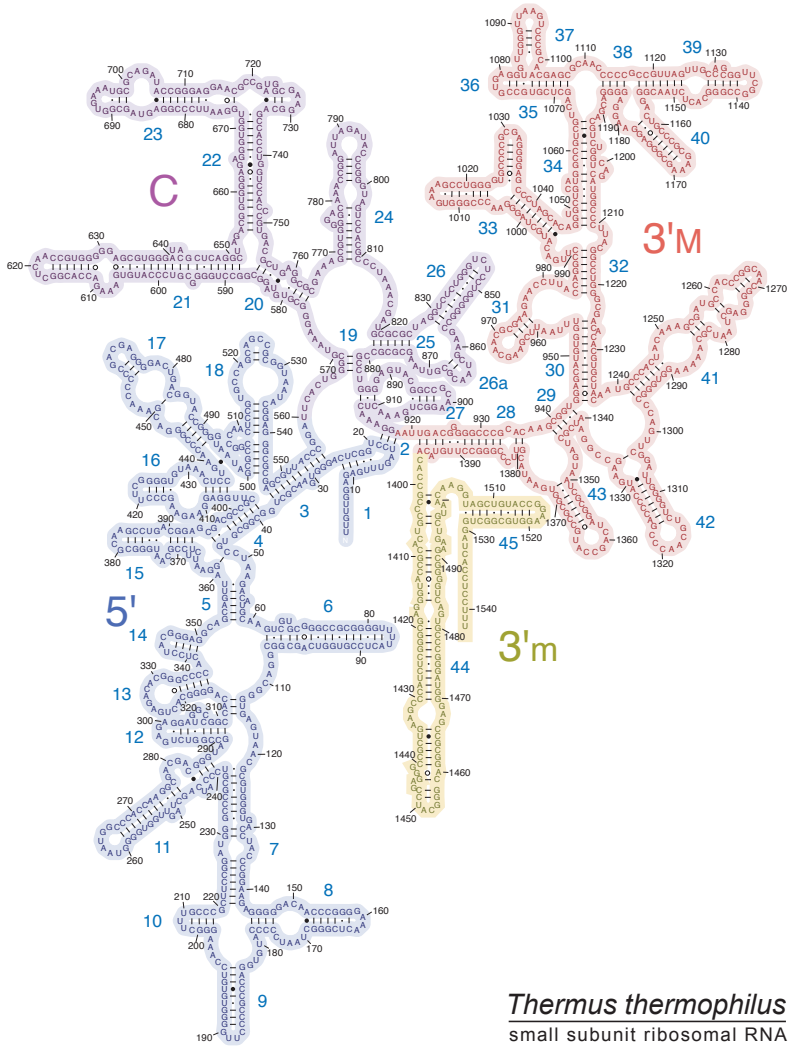
Полина Лунина

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

19 декабря 2020г.

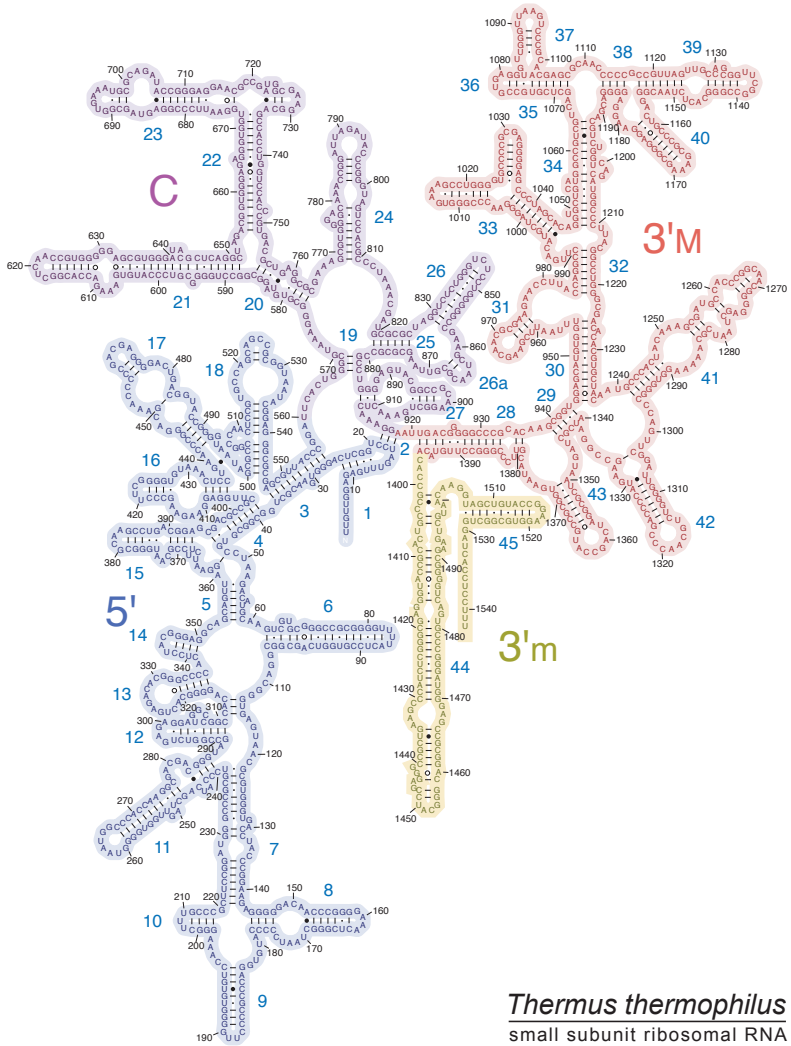
Биоинформатика

- Задачи
  - Распознавание
  - Классификация
  - Предсказание вторичных структур
  - ...



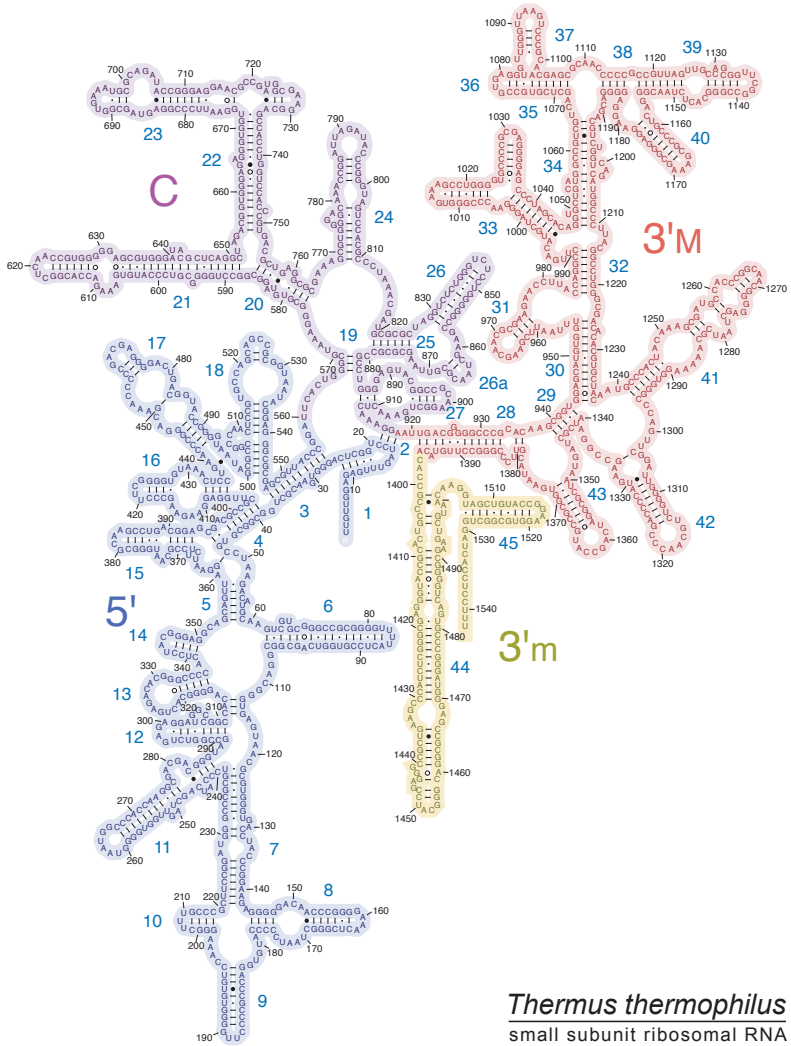
Биоинформатика

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



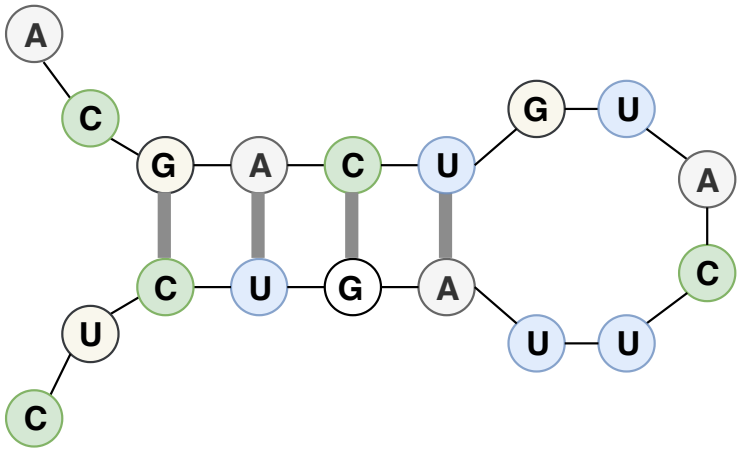
Биоинформатика

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



Подход: синтаксический анализ + машинное обучение

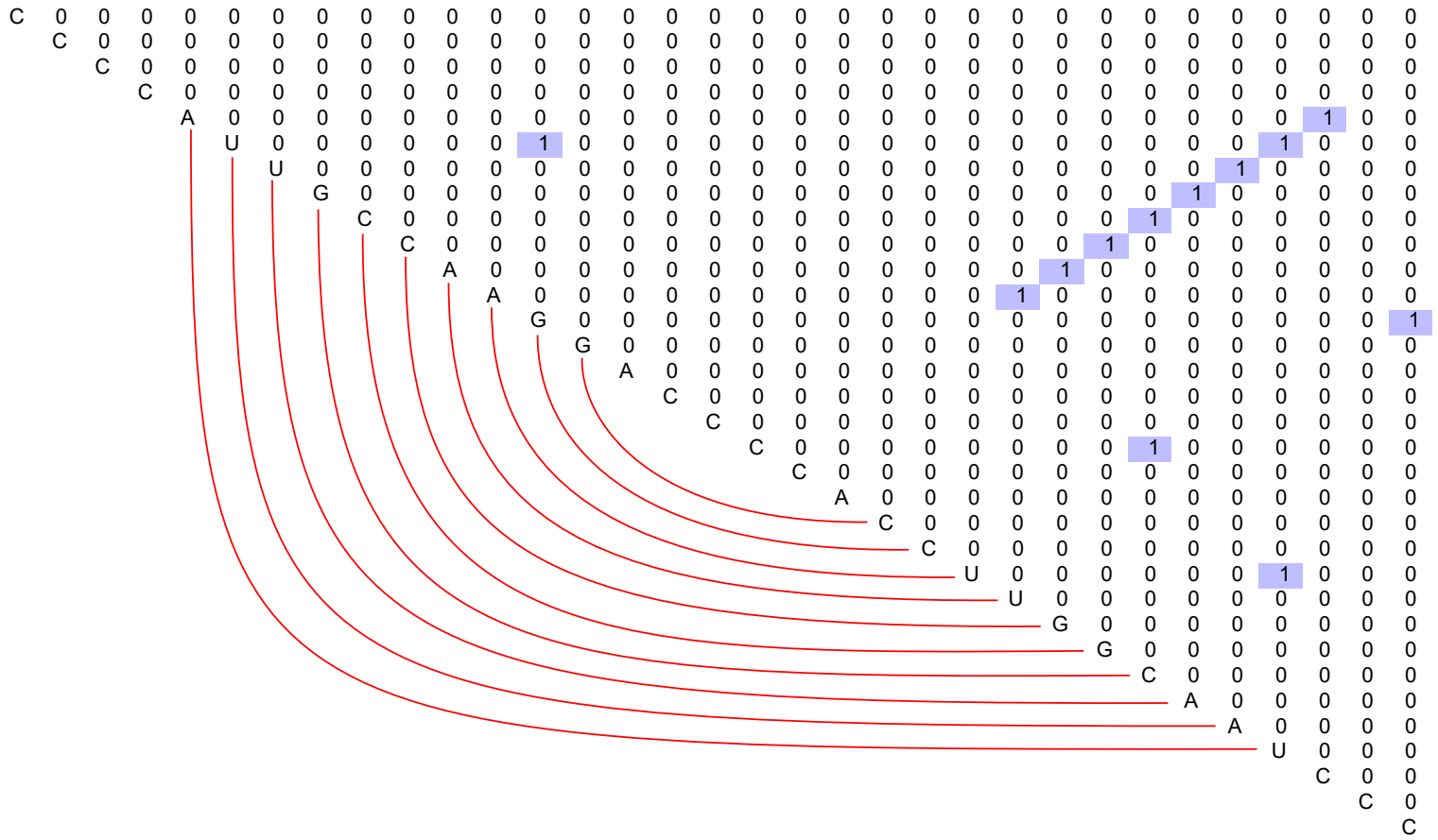
- Задать основные элементы вторичной структуры (стемы) с помощью грамматики
- Искать стемы во входных цепочках при помощи парсера
- Для дальнейшей обработки и вероятностной оценки использовать нейронные сети



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

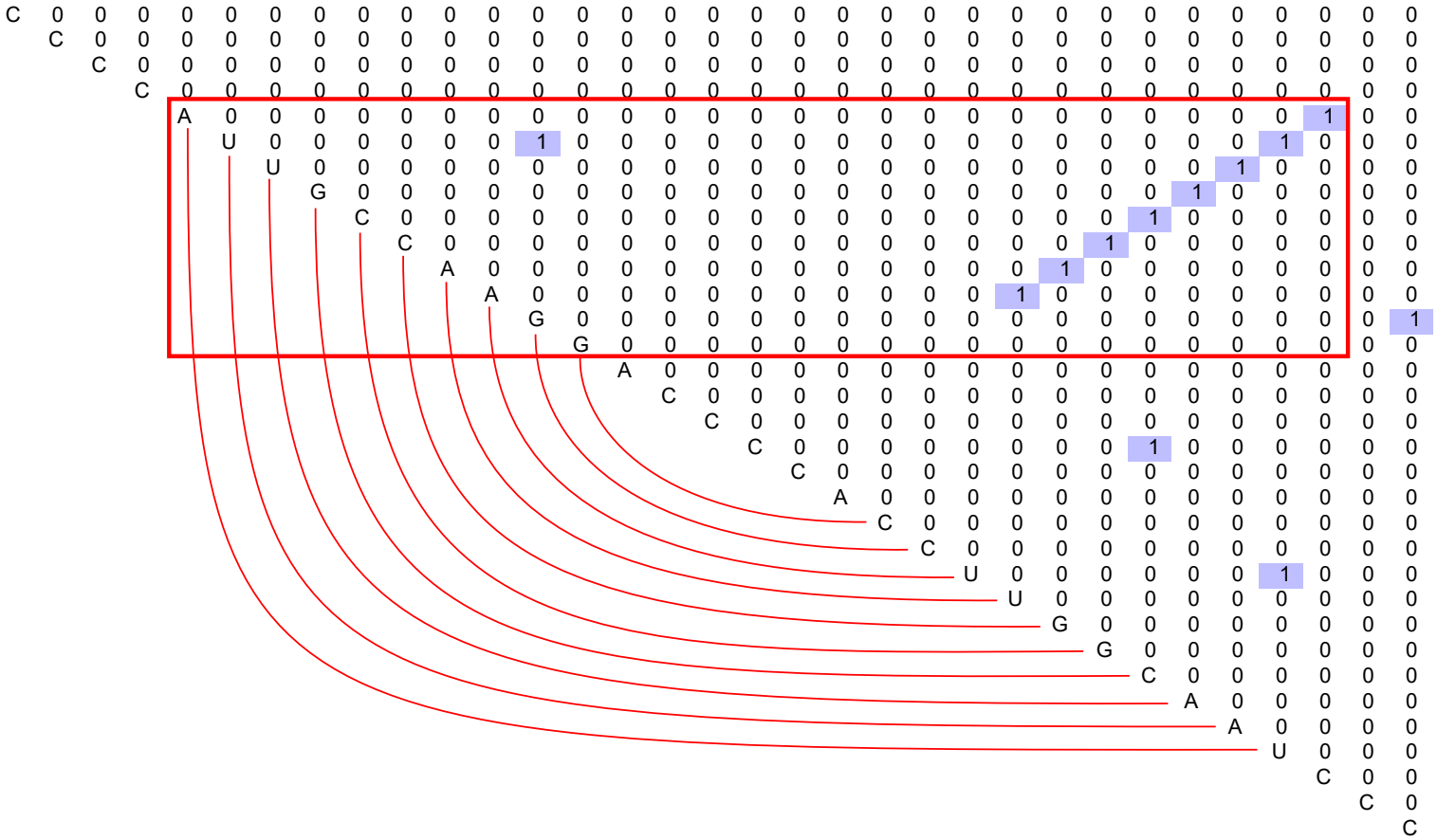
Пример

CCCCAUUGCCAAGGACCCCACCUUGGCAAUCCC



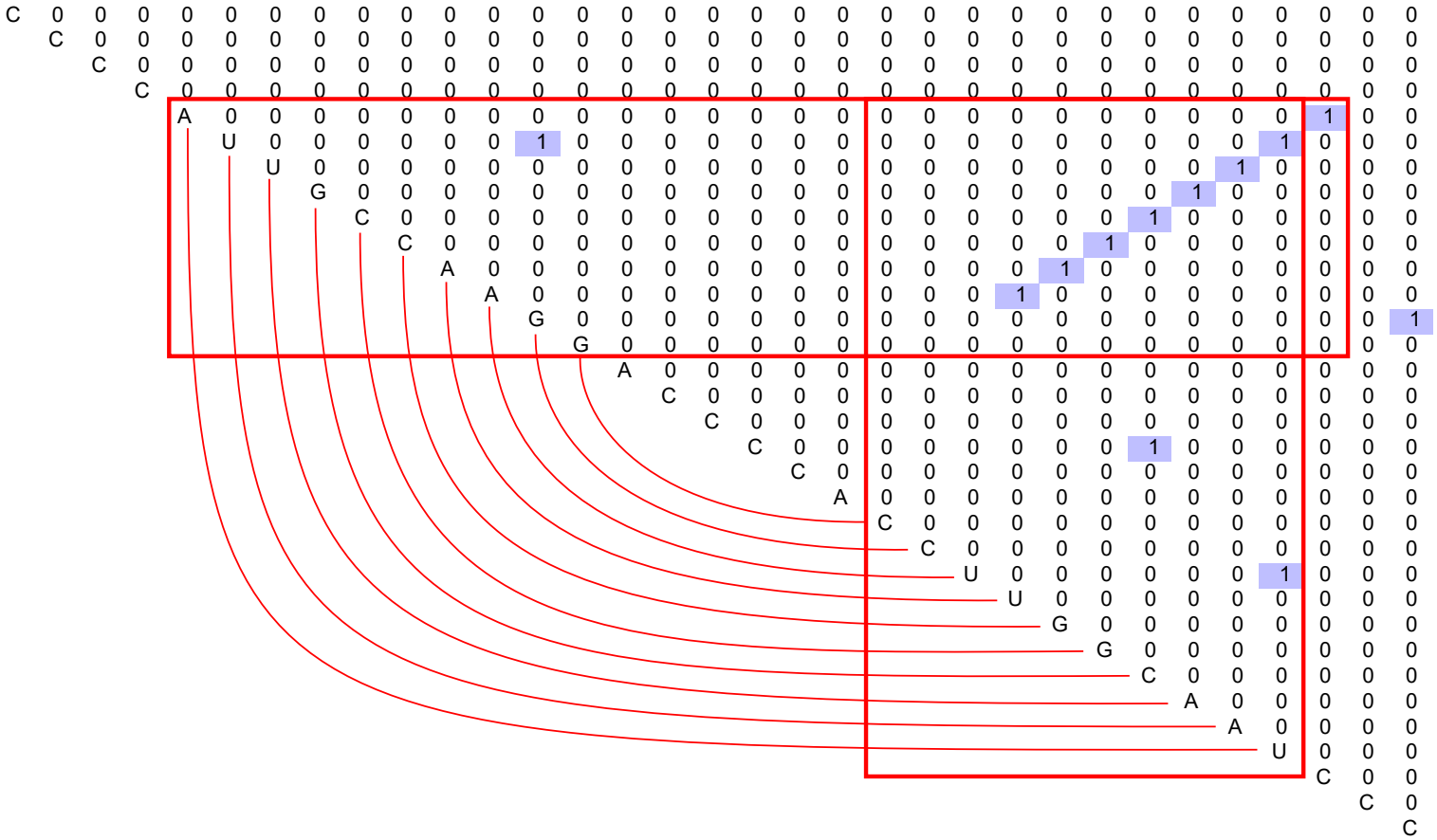
Пример

CCCCAUUGCCAAGGACCCCACCUUGGCAAUCCC



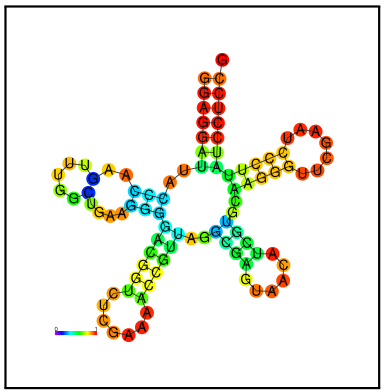
Пример

CCCCAUUGCCAAGGACCCCACCUUGGCAAUCCC

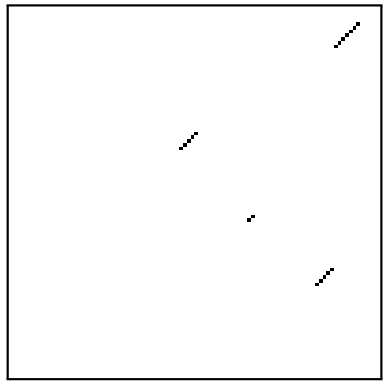


Предсказание вторичной структуры РНК

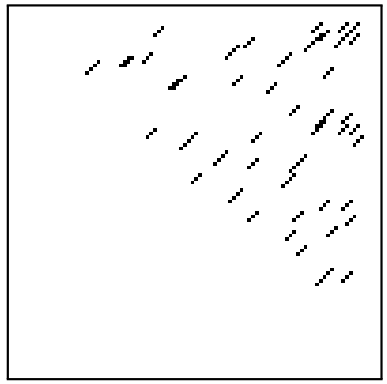
- Парсер находит в цепочке все возможные стемы, однако не все они действительно будут входить в состав вторичной структуры
- Хотим отбработать матрицу разбора нейронной сетью и предсказать вторичную структуру цепочки



Вторичная структура



Матрица контактов



Матрица разбора

Предсказание вторичной структуры РНК

Где найти эталонные вторичные структуры?

- Выгрузить из биологических баз данных
- Сгенерировать некоторым тулом

Проблема: в базах слишком мало данных для обучения

Решение: transfer learning — обучить нейросеть на сгенерированных данных, а затем дообучить ее на реальных вторичных структурах

Проблема: не хотим эмулировать работу уже существующего тула и повторять его ошибки

Решение: обучить n сетей для n тулов, а при дообучении на реальных данных соединить результаты в общую модель

Предсказание вторичной структуры РНК

Задачи:

- Предсказание вторичных структур тРНК по сгенерированным различными инструментыми данным
- Предсказание реальных вторичных структур цепочек тРНК

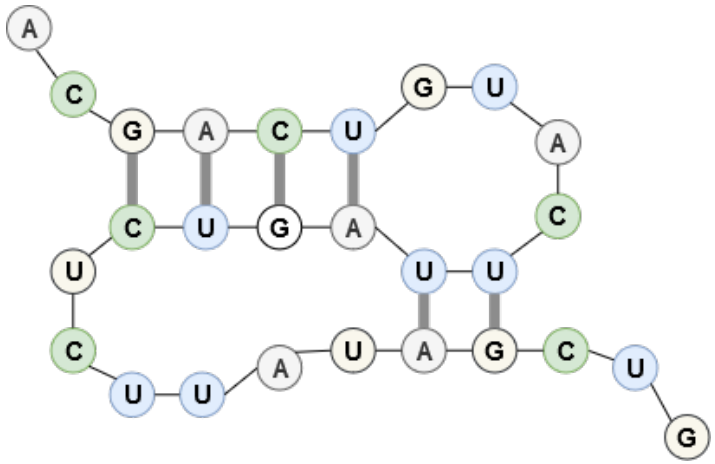
Технологии и данные

- Платформа YaccConstructor
- Библиотека Keras и фреймворк Tensorflow
- Инструменты HotKnots, pknotsRG, RNAstructure и SPOT-RNA
- Базы данных RNAcentral, Pseudobase и RNAstrand

Используемые инструменты

Требования

- Основаны на разных алгоритмах
- Результаты различаются, но все имеют высокую точность
- Удобство и скорость работы
- Предсказывают псевдоузлы

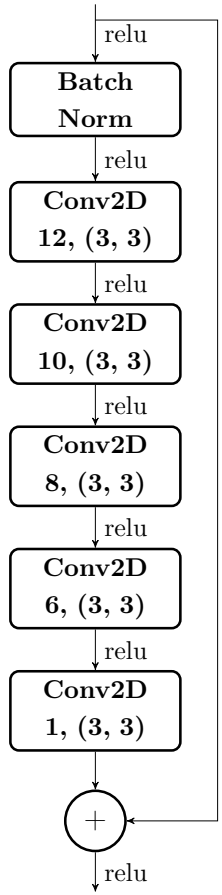


Выбрали

- HotKnots — эвристический алгоритм + MFE
- SPOT-RNA — deep learning
- pknotsRG — Turner energy rules + MFE
- RNAstructure — динамическое программирование + MFE

Нейронная сеть: этап 1

- ResNet из десяти блоков для каждого из четырех тулов
- Loss — взвешенная попиксельная разница
- Метрики
  - Precision — сколько из предсказанных контактов действительно являются контактами в эталоне
  - Recall — сколько из требуемых контактов найдено
  - F1 score — объединяющая метрика
- Длина цепочки от 1 до 100, около 18000 образцов на каждую сеть, train:test = 80%:20%

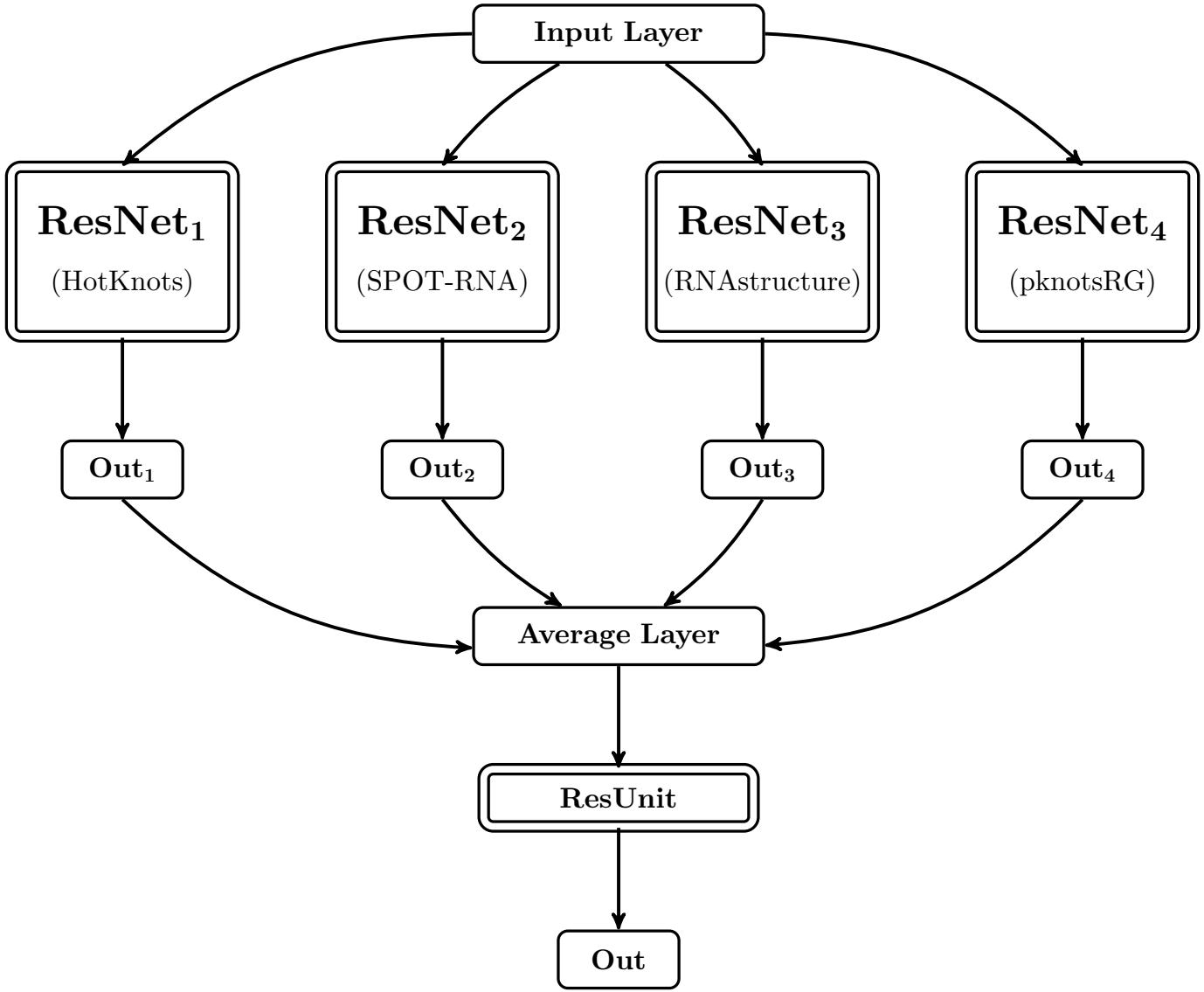


Результаты: этап 1

Средние значения метрик на тестовых выборках для каждой модели

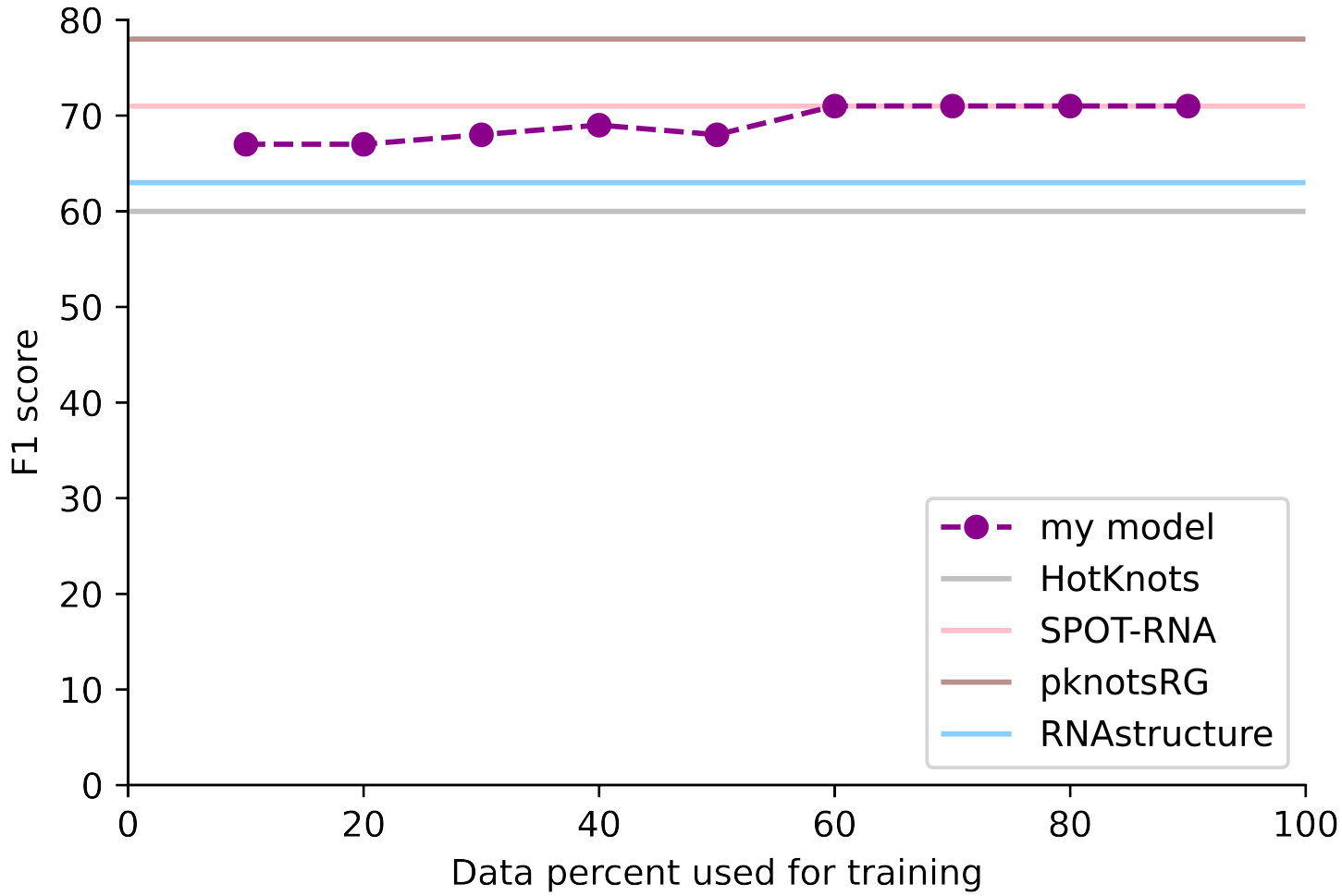
Base tool	Precision	Recall	F1 score
HotKnots	38%	44%	39%
pknotsRG	40%	45%	40%
RNAstructure	41%	48%	42%
SPOT-RNA	41%	50%	42%

Нейронная сеть: этап 2



Результаты: этап 2

База Pseudobase — 255 структур, все с псевдоузлами



Результаты: этап 2

База RNAstrand — 819 структур, из них 74 с псевдоузлами

TODO!!!!!!!

Итоги

Публикации

- Semyon Grigorev, Dmitry Kutlenkov, Polina Lunina. Secondary structure prediction by combination of formal grammars and neural networks. BMC Bioinformatics, Scopus
- Polina Lunina, Semyon Grigorev. On Secondary Structure Analysis by Using Formal Grammars and Artificial Neural Networks. LNBI, Scopus

Планы на будущее

- Улучшение полученных результатов
- Подготовка к конференции AICoB 2020&2021