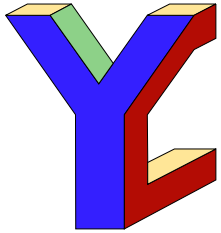


Комбинирование нейронных сетей и синтаксического анализа для предсказания вторичной структуры генетических цепочек

Лунина Полина Сергеевна
Научный руководитель: Григорьев С.В.
Рецензент: Малыгина Т.С.

Санкт-Петербургский государственный университет

9 июня 2021г.

Введение

Геномные последовательности

- РНК
- ДНК
- Белки

Уровни молекулярной организации

- Первичная структура (линейная)
- Вторичная структура (пространственная)

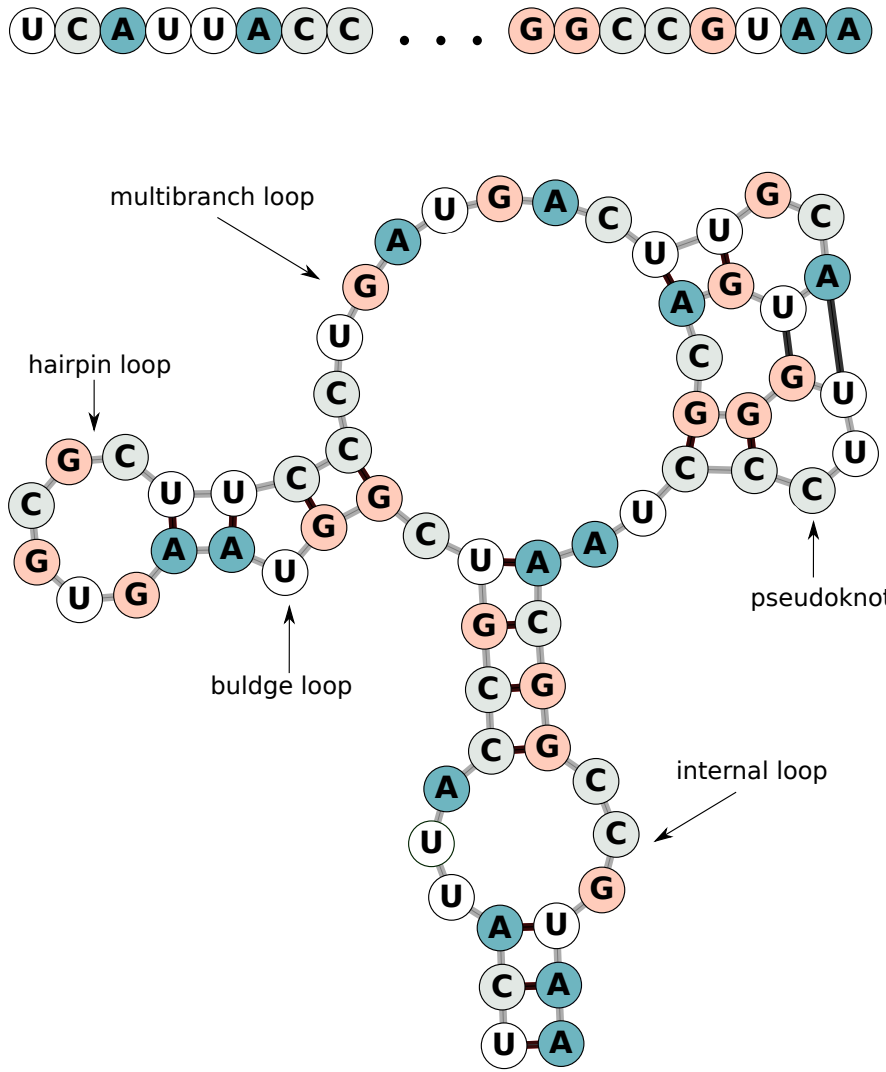
Задачи

- Распознавание последовательностей
- Классификация организмов
- Предсказание вторичных структур
- ...

Вторичная структура РНК

Значение

- Транскрипция, трансляция
- Филогенетика, таксономия



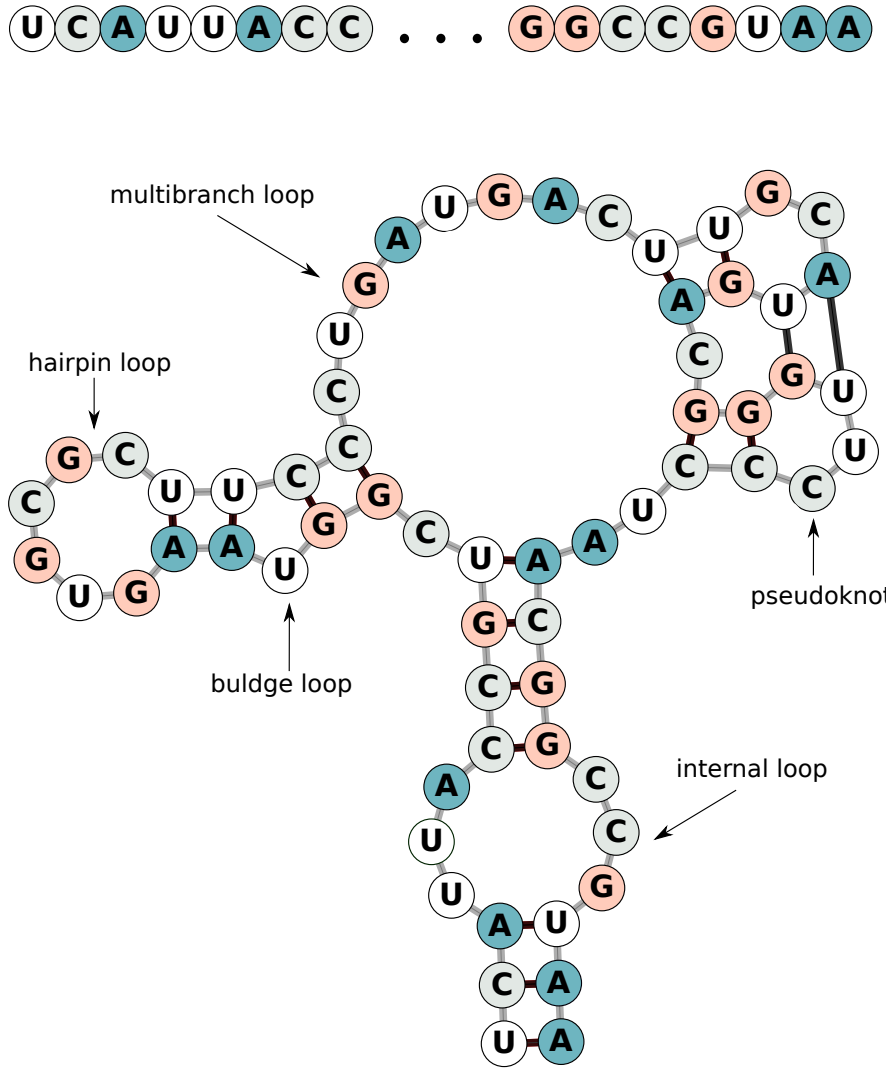
Вторичная структура РНК

Значение

- Транскрипция, трансляция
- Филогенетика, таксономия

Методы предсказания

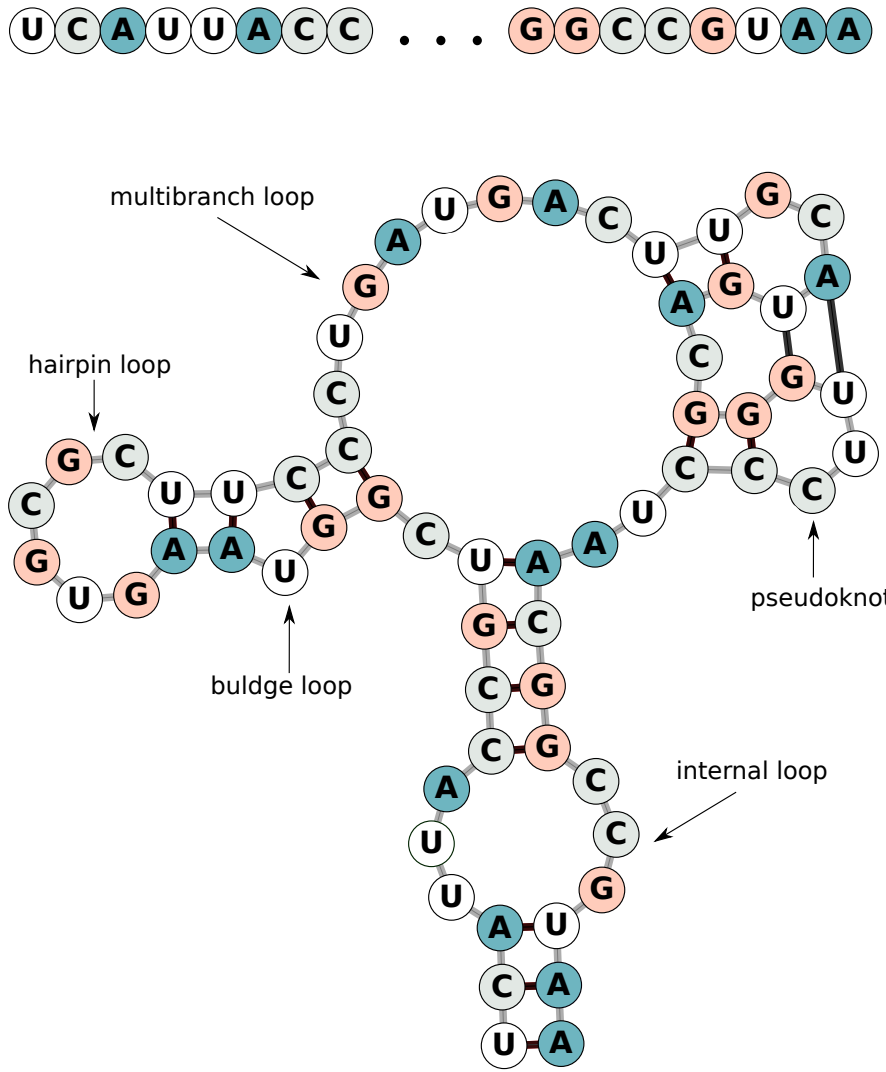
- Лабораторные
- Сравнительные
- Вычислительные
  - Минимизация свободной энергии
  - Стохастические модели и грамматики
  - Машинное обучение
  - ...



Вторичная структура РНК

Проблемы

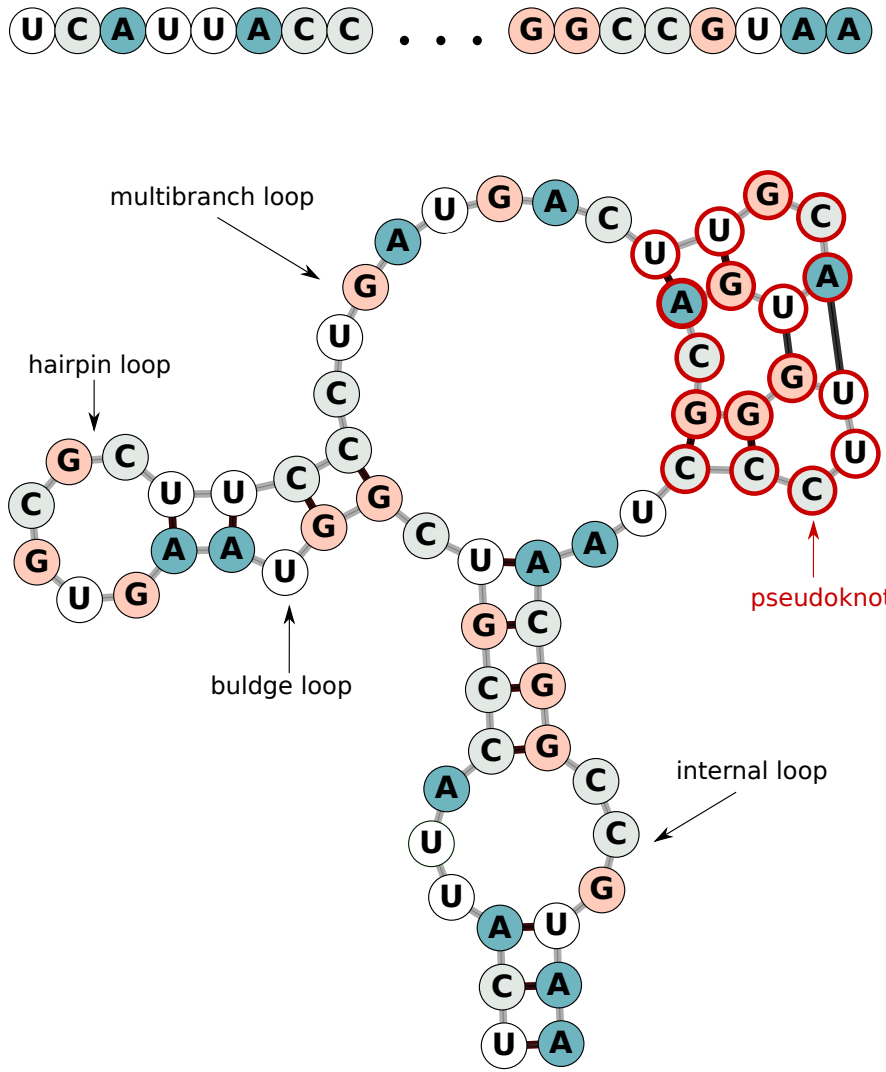
- Сложность формализации
- Предсказание псевдоузлов
- Вариативность элементов
- Зашумленность данных



Вторичная структура РНК

Проблемы

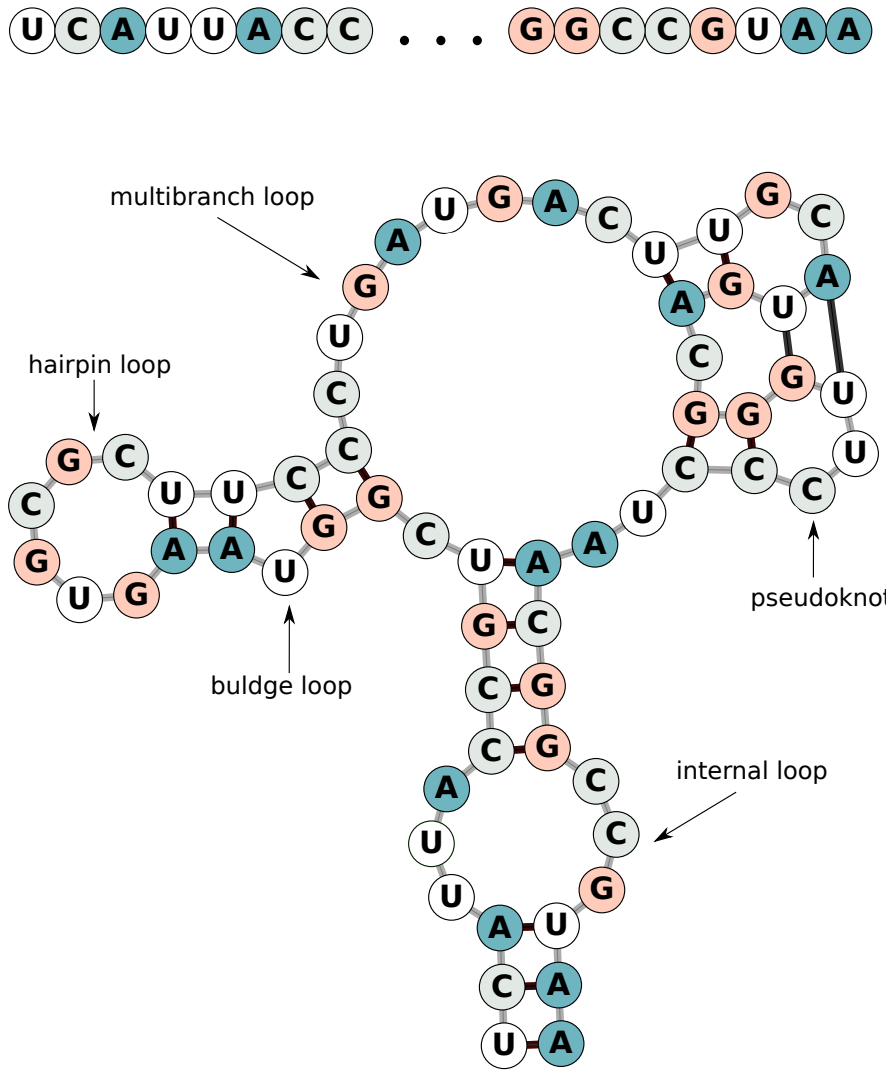
- Сложность формализации
- Предсказание псевдоузлов
- Вариативность элементов
- Зашумленность данных



Вторичная структура РНК

Проблемы

- Сложность формализации
- Предсказание псевдоузлов
- Вариативность элементов
- Зашумленность данных



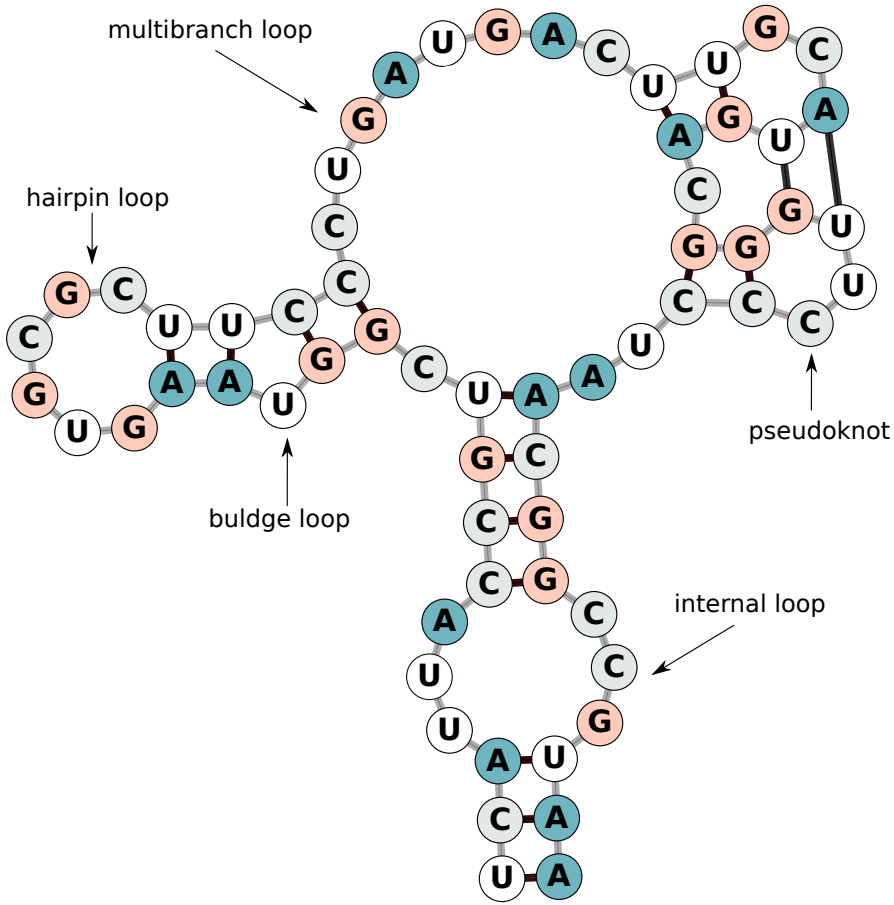
Вторичная структура РНК

Проблемы

- Сложность формализации
- Предсказание псевдоузлов
- Вариативность элементов
- Зашумленность данных

Наше решение

- Формальная грамматика для описания базовых законов
- Нейронная сеть для синтеза вторичной структуры



Постановка задачи

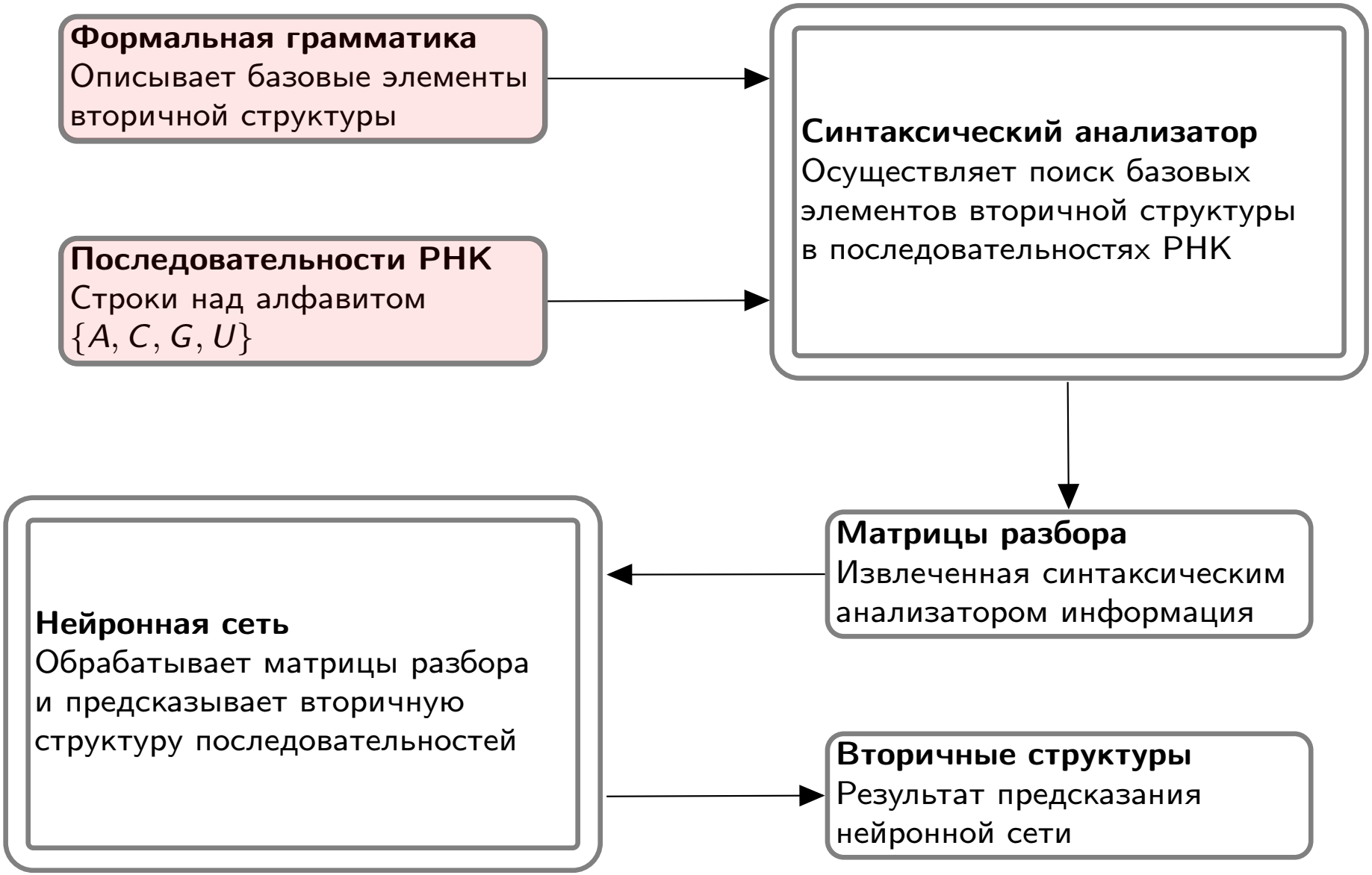
Цель — исследование возможности применения подхода, основанного на комбинировании нейронных сетей и синтаксического анализа, к задаче предсказания вторичной структуры молекулы РНК

Задачи

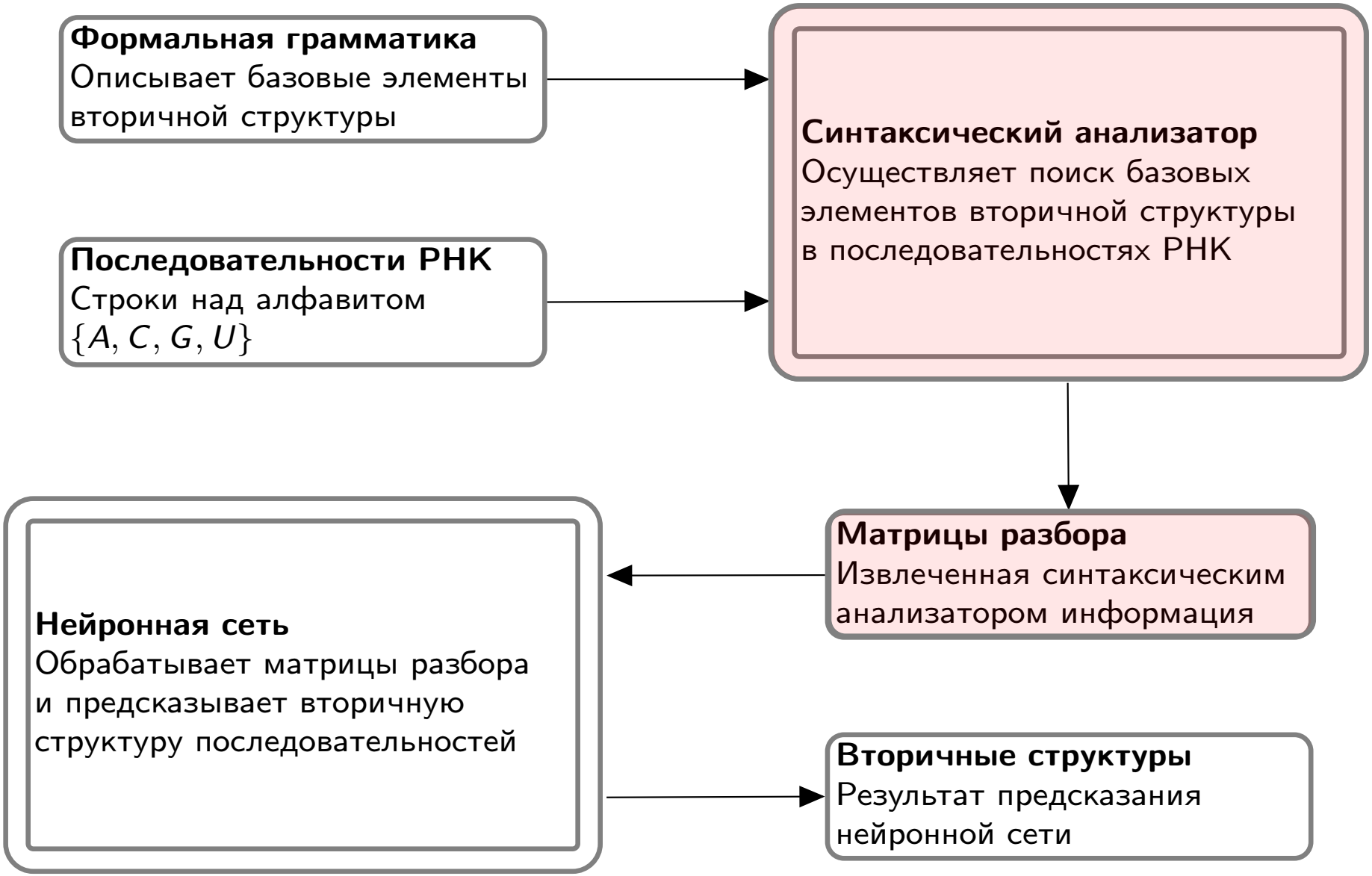
- Разработка архитектуры решения, конкретизирующей форматы данных, используемые грамматики и нейронные сети
- Проведение экспериментальных исследований, сравнение с аналогами

Архитектура решения

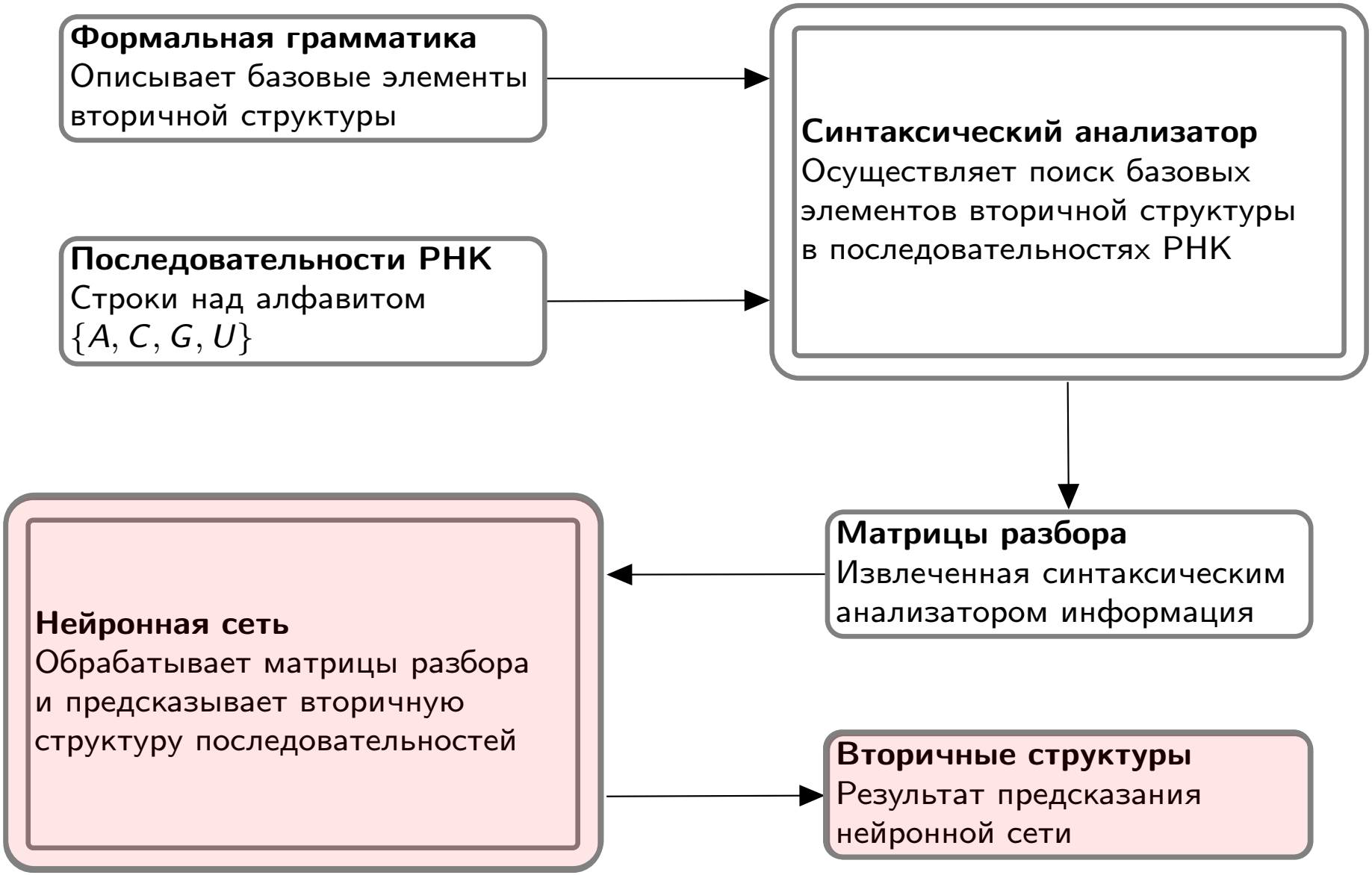
Архитектура решения



Архитектура решения

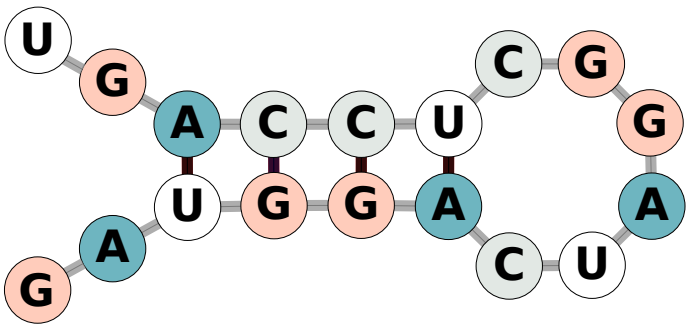


Архитектура решения



Формальная грамматика

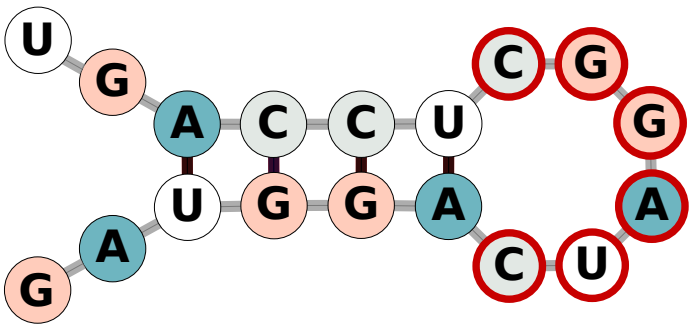
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
      stem1<stem2<s>>
    | A stem3<s> U
    | U stem3<s> A
    | C stem3<s> G
    | G stem3<s> C
```

Формальная грамматика

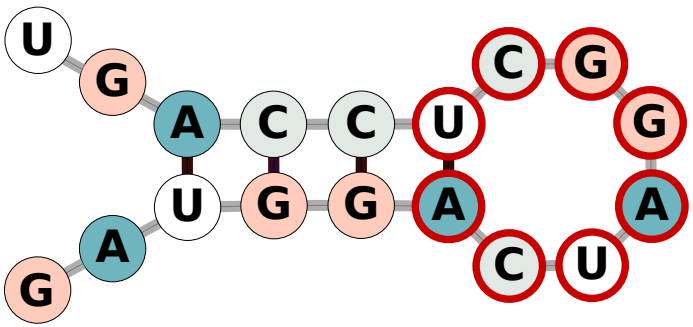
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
        stem1<stem2<s>>
      | A stem3<s> U
      | U stem3<s> A
      | C stem3<s> G
      | G stem3<s> C
```

Формальная грамматика

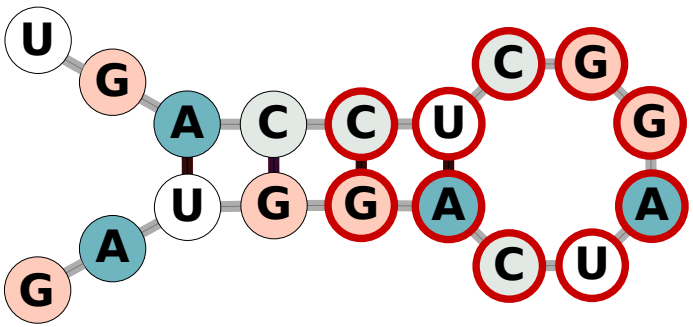
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
        stem1<stem2<s>>
      | A stem3<s> U
      | U stem3<s> A
      | C stem3<s> G
      | G stem3<s> C
```

Формальная грамматика

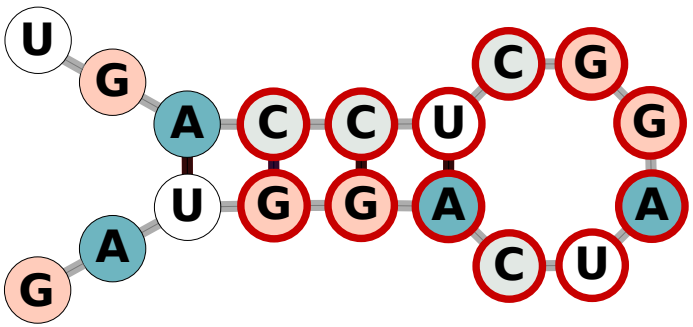
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
        stem1<stem2<s>>
      | A stem3<s> U
      | U stem3<s> A
      | C stem3<s> G
      | G stem3<s> C
```

Формальная грамматика

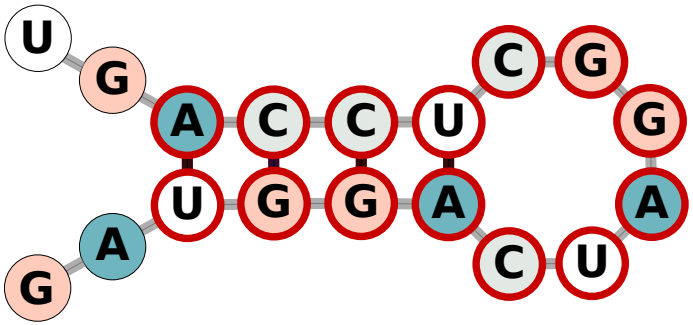
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
        stem1<stem2<s>>
      | A stem3<s> U
      | U stem3<s> A
      | C stem3<s> G
      | G stem3<s> C
```

Формальная грамматика

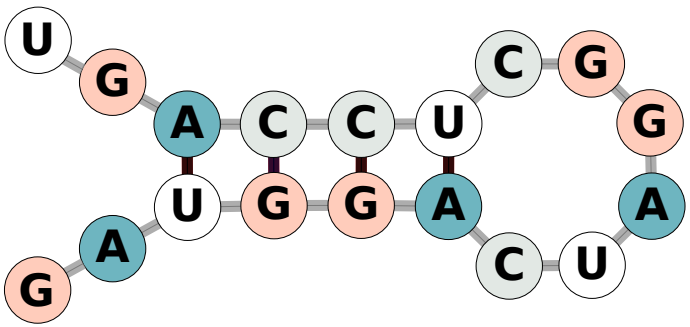
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
        stem1<stem2<s>>
      | A stem3<s> U
      | U stem3<s> A
      | C stem3<s> G
      | G stem3<s> C
```

Формальная грамматика

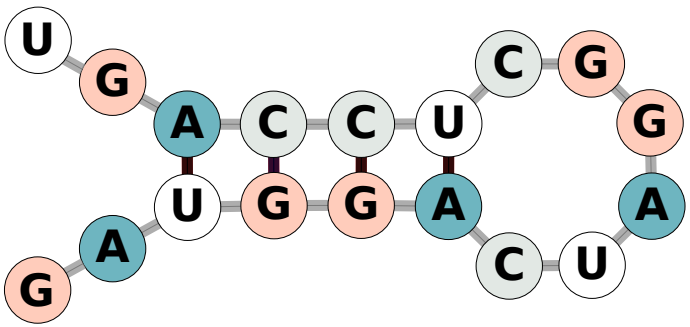
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
      stem1<stem2<s>>
    | A stem3<s> U
    | U stem3<s> A
    | C stem3<s> G
    | G stem3<s> C
```

Формальная грамматика

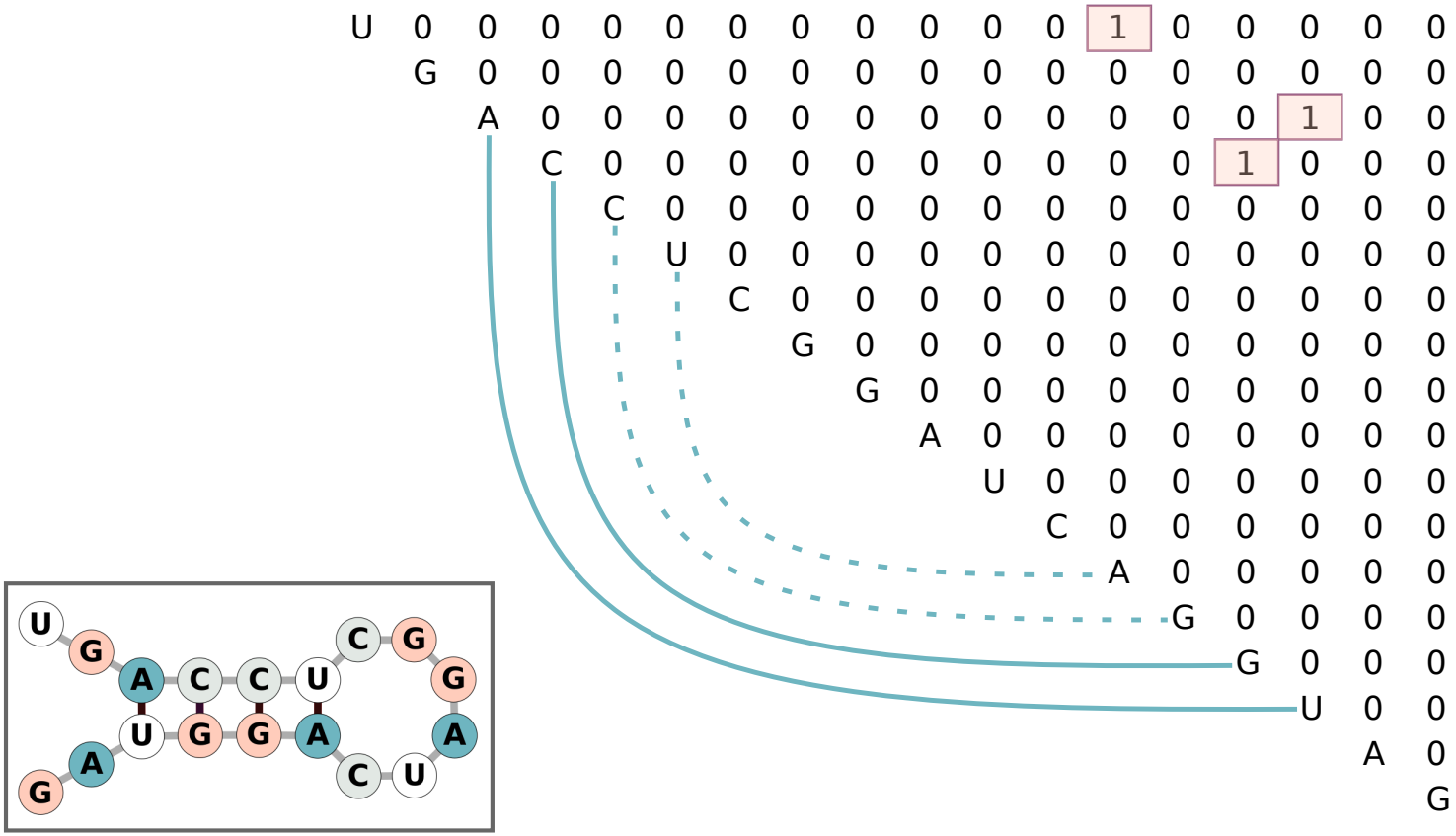
- Вторичная структура как рекурсивная композиция базовых элементов — шпилек (stem-loop)
- KC грамматика для описания общего вида шпильки
- Ограничения: размер петли (1-20), высота (от 3), канонические пары (A-U, C-G)




```
start: stem3<s0>
s0: loop | loop stem3<s0> s0
loop: nucl*[1..20]
nucl: A | U | C | G
stem1<s>: A s U | G s C | U s A | C s G
stem2<s>: stem1<stem1<s>>
stem3<s>:
      stem1<stem2<s>>
    | A stem3<s> U
    | U stem3<s> A
    | C stem3<s> G
    | G stem3<s> C
```

Синтаксический анализатор

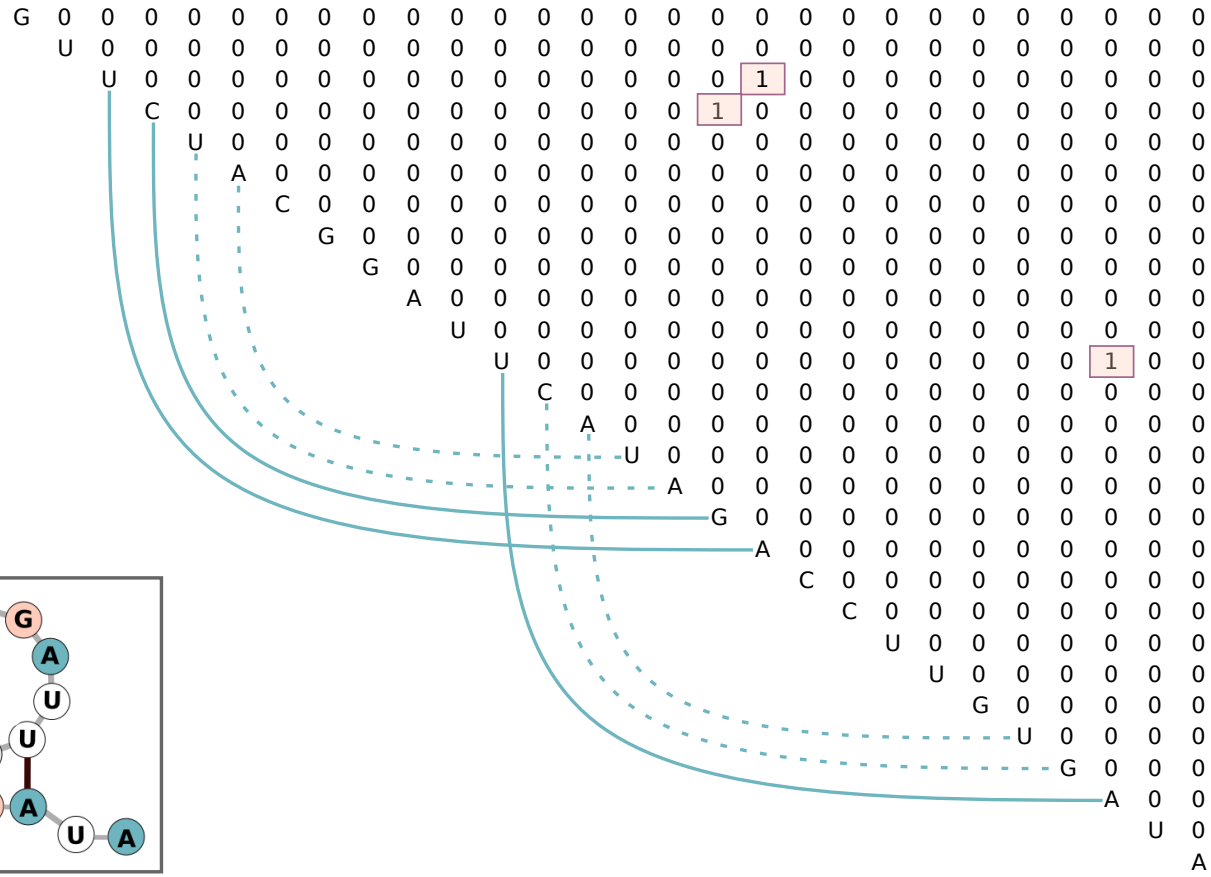
- Задача поиска всех возможных шпилек в последовательности
- Результат работы — матрица разбора
  $M_p(seq)$: $M_p[i,j] = 1 \iff seq[i..j]$ выводима в грамматике
- Шпильке высоты $n \geq 3$ соответствует столбик из $n-2$ единиц

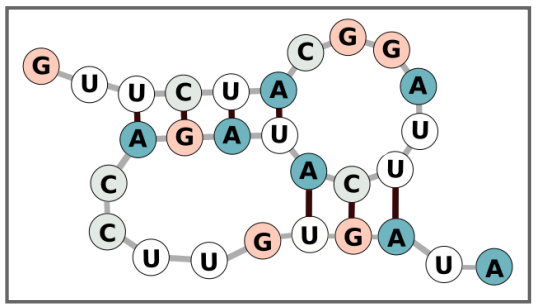


Синтаксический анализатор

- Псевдоузлы не выводимы в КС грамматике
- Шпильки, из которых они состоят, выводимы по отдельности

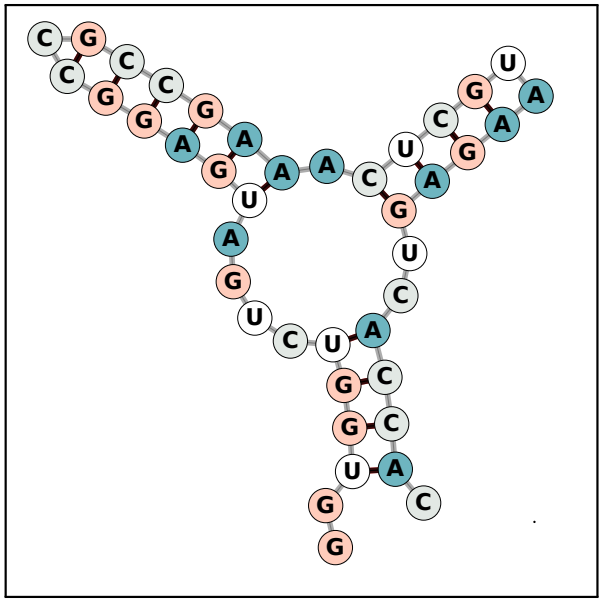
$\Rightarrow$ Наш подход позволяет учитывать псевдоузлы



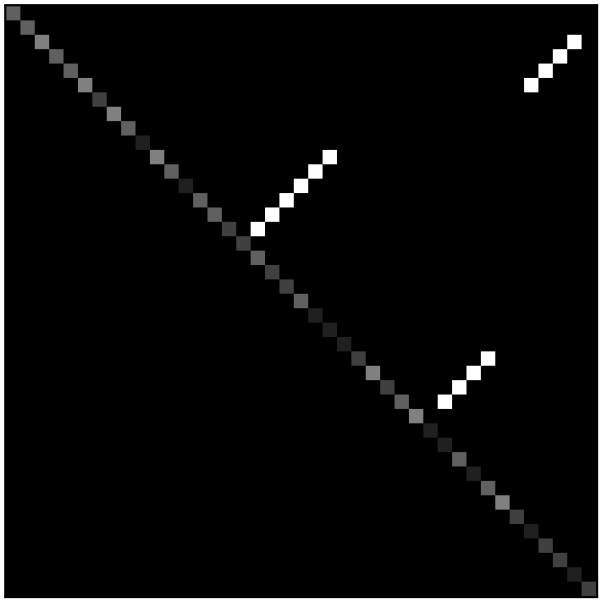


Нейронная сеть

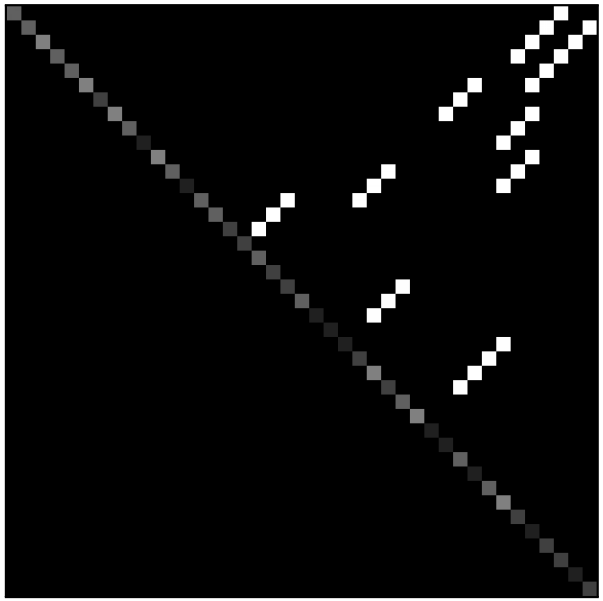
- Парсер находит все возможные шпильки $\Rightarrow$ избыточность M_p
- В грамматике есть ограничения $\Rightarrow$ недостаточность M_p
- Решение — обработка матриц разбора нейронной сетью
- Эталон — матрица контактов реальной вторичной структуры $M_c(seq)$: $M_c[i,j] = 1 \iff seq[i]$ и $seq[j]$ образуют контакт



Вторичная структура



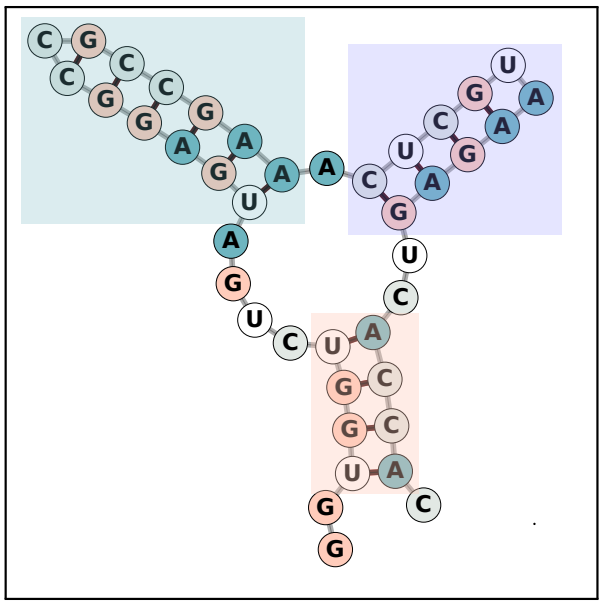
Эталонное изображение (матрица контактов)



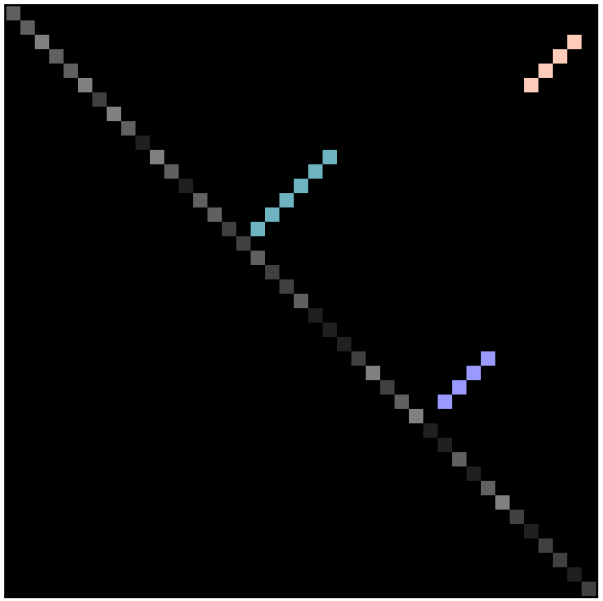
Входное изображение (матрица разбора)

Нейронная сеть

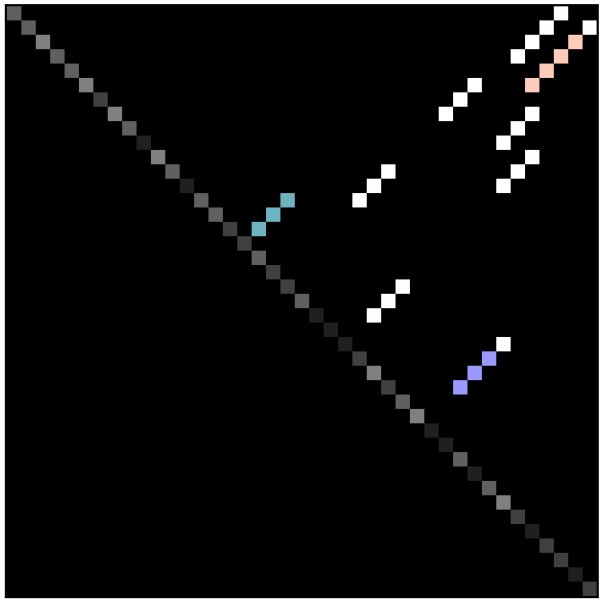
- Парсер находит все возможные шпильки $\Rightarrow$ избыточность M_p
- В грамматике есть ограничения $\Rightarrow$ недостаточность M_p
- Решение — обработка матриц разбора нейронной сетью
- Эталон — матрица контактов реальной вторичной структуры $M_c(seq)$: $M_c[i,j] = 1 \iff seq[i]$ и $seq[j]$ образуют контакт



Вторичная структура



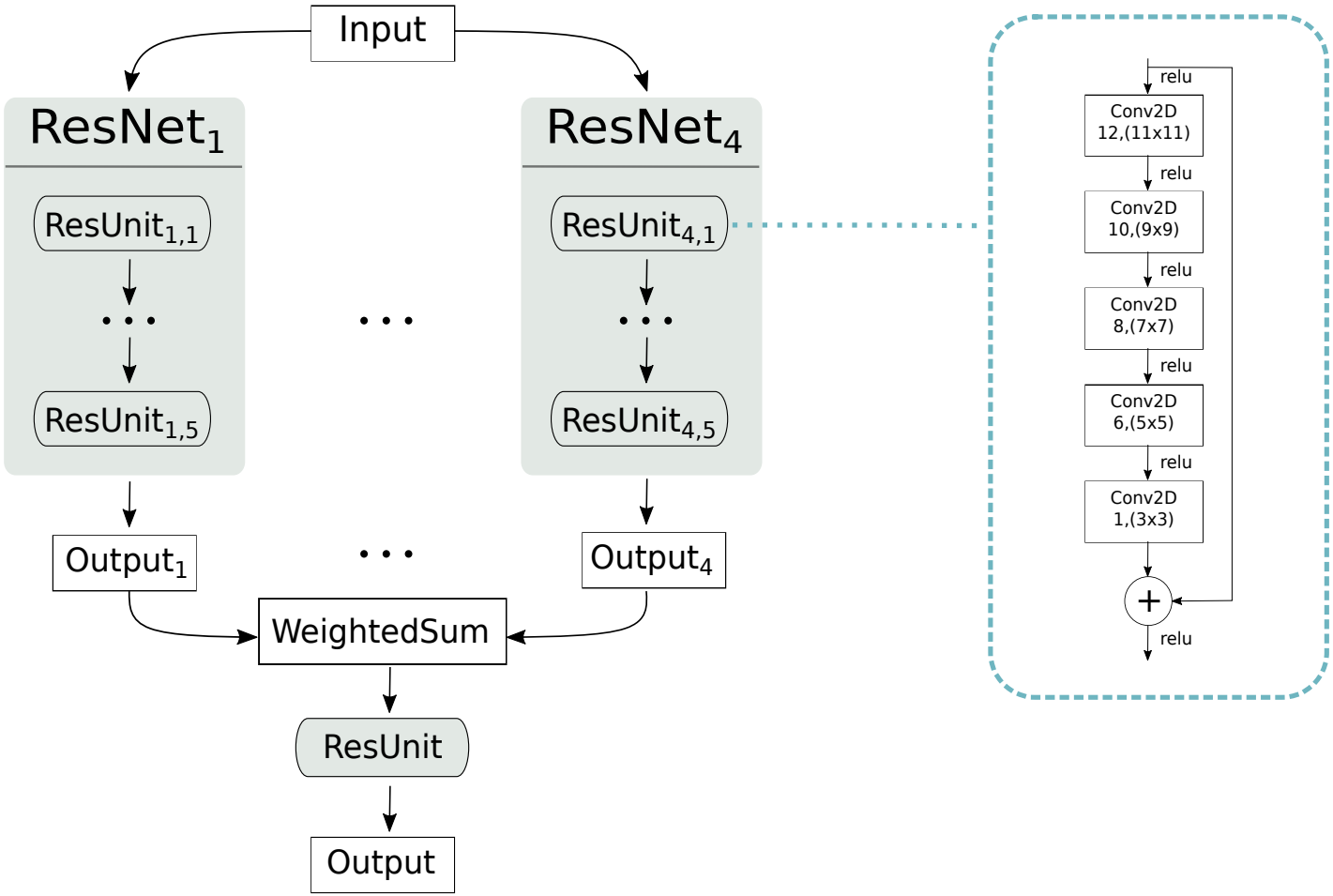
Эталонное изображение (матрица контактов)



Входное изображение (матрица разбора)

Нейронная сеть

- Параллельная остаточная архитектура
- Dropout и L2-регуляризация для снижения переобучения



Эксперименты

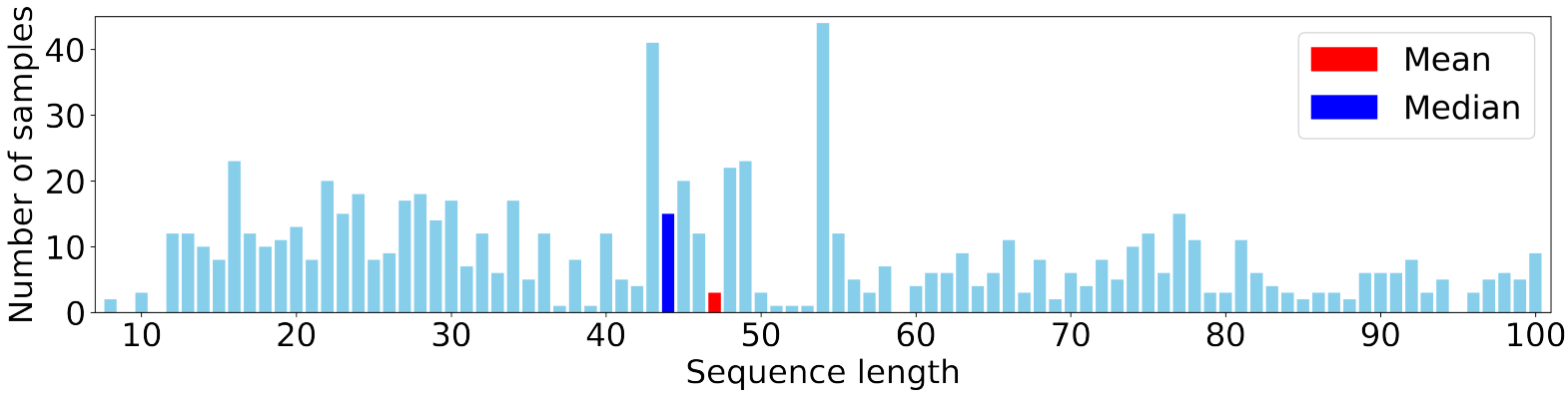
Задача — предсказание вторичных структур для цепочек РНК с имеющимися достоверными эталонными данными

Данные

- База RNAstrand (последовательности + вторичные структуры)
- 800 последовательностей длины до 100 (74 с псевдоузлами)
- Data augmentation — отражение относительно побочной диагонали

Технологии

- Синтаксический анализ: платформа YaccConstructor
- Нейронные сети: библиотека Keras и фреймворк Tensorflow



Эксперименты

Аналоги

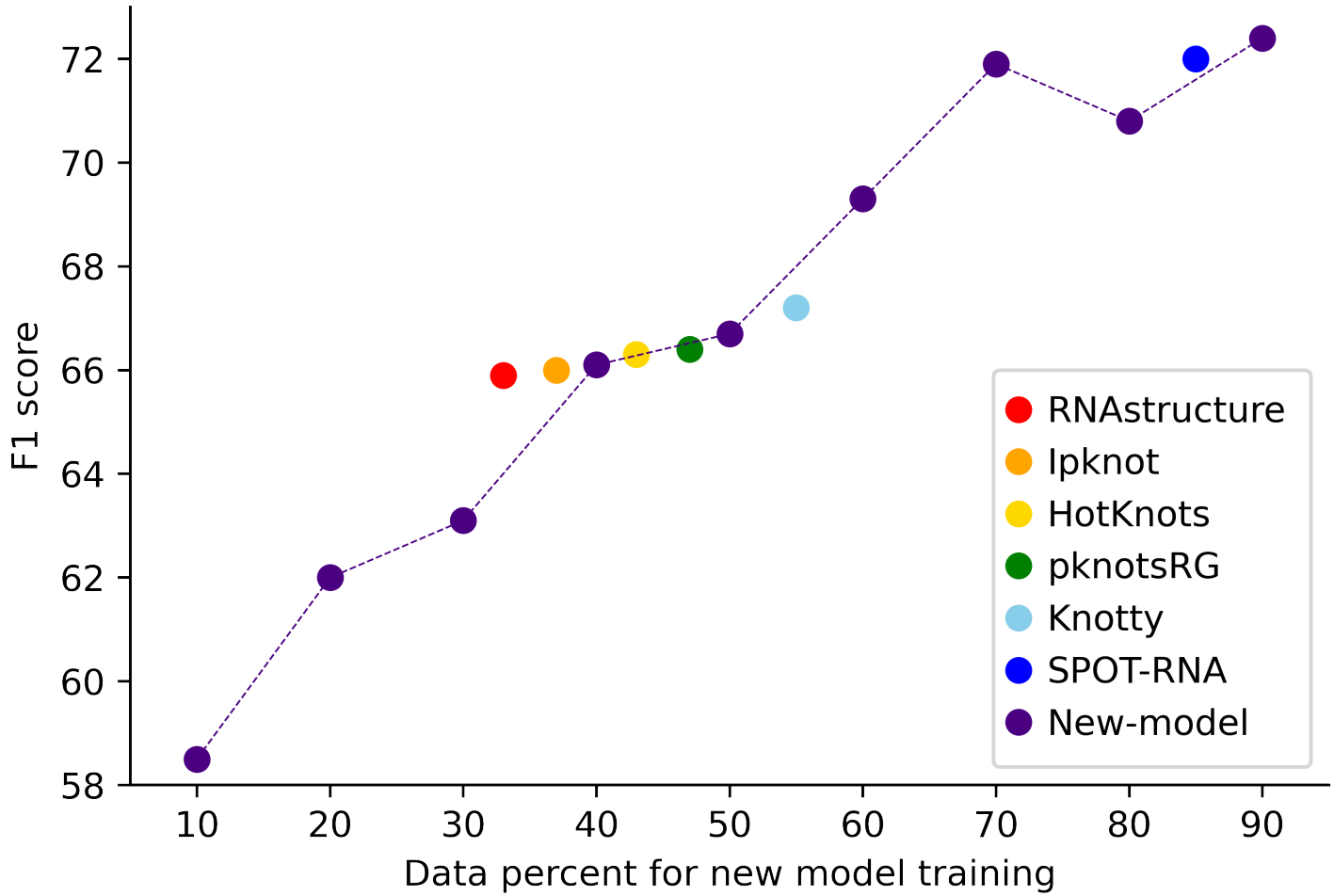
- HotKnots — MFE + эвристический алгоритм
- SPOT-RNA — машинное обучение
- PknotsRG — MFE + Turner energy rules
- RNAstructure — MFE + динамическое программирование
- Ipknot — MEA + целочисленное программирование

Метрики

- $Precision = \frac{TP}{TP+FP}$ (доля верных контактов среди предсказанных)
- $Recall = \frac{TP}{TP+FN}$ (доля предсказанных контактов среди искомых)
- $F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$ (объединяющая метрика)

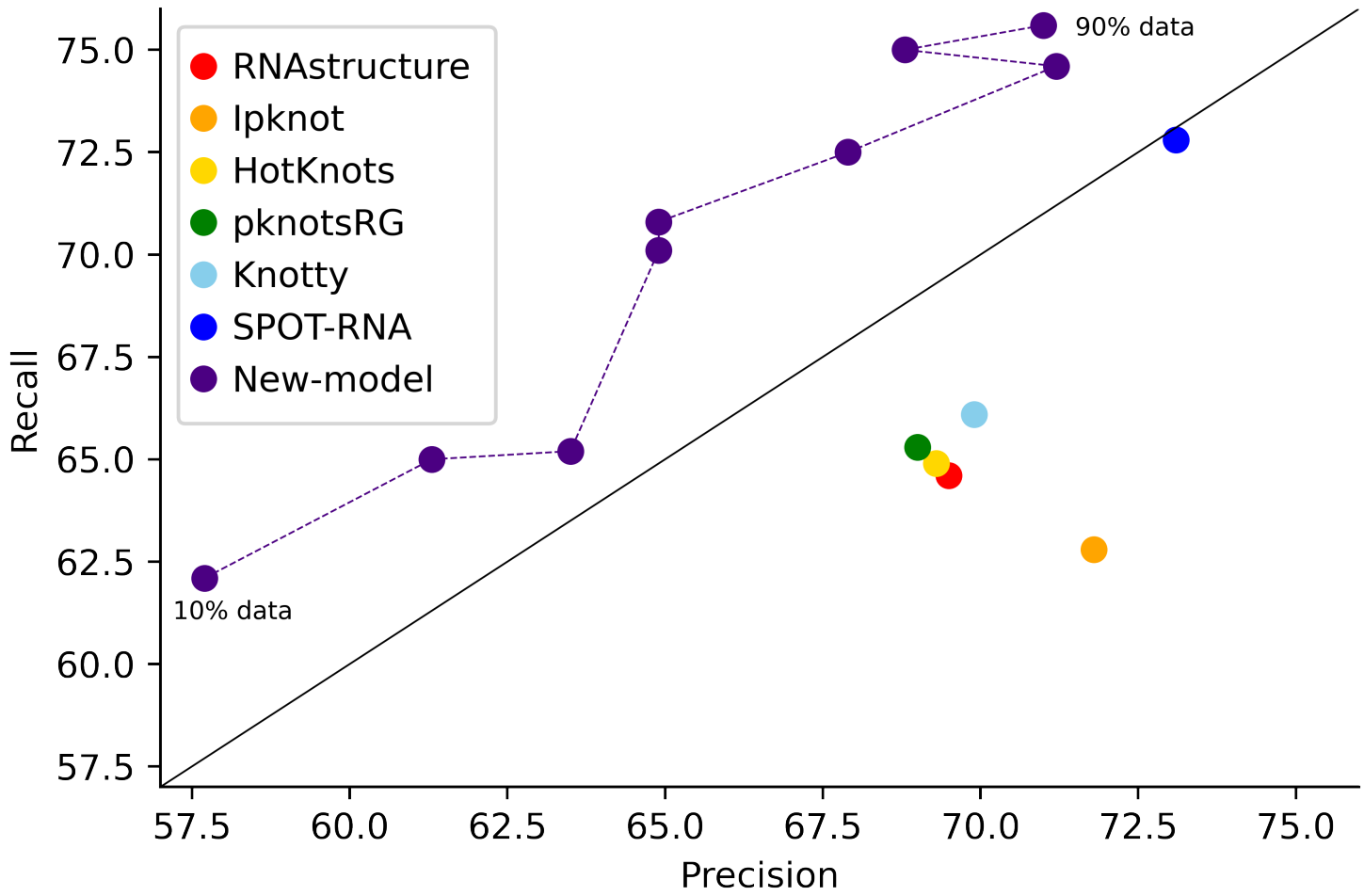
Результаты

Значения метрики $F1$ на тестовых выборках для нашей модели и на всей выборке для остальных инструментов



Результаты

Значения метрик *Precision* и *Recall* на тестовых выборках для нашей модели и на всей выборке для остальных инструментов



Заключение

- Разработана архитектура решения для предсказания вторичной структуры РНК на основе комбинирования методов синтаксического анализа и машинного обучения
- Проведены эксперименты на реальных данных и сравнение полученных результатов с аналогами
- Представлен постер "Secondary structure prediction by combination of formal grammars and neural networks"на конференции Biata 2020 и опубликована одноименная статья (BMC Bioinformatics, Scopus)

Время работы алгоритмов

- Операционная система: Ubuntu 20.04.2 LTS
- Центральный процессор: Intel Core i5-10210U CPU 1.60GHz
- Графический процессор: NVIDIA GeForce MX250
- Объем оперативной памяти: 7.5 GB

Tool	**Time, s**
Ipknot	0.8
RNAstructure	10.3
PknotsRG	14.9
Hotknots	37.0
SPOT-RNA (GPU)	67.8
New-model (PA + NN)	103.1 (80.7 + 22.4)
SPOT-RNA (CPU)	109.7
Knotty	282.8