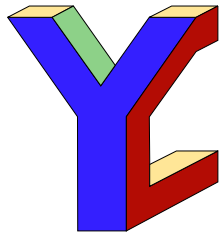


Комбинирование нейронных сетей и синтаксического анализа для предсказания вторичных структур генетических цепочек

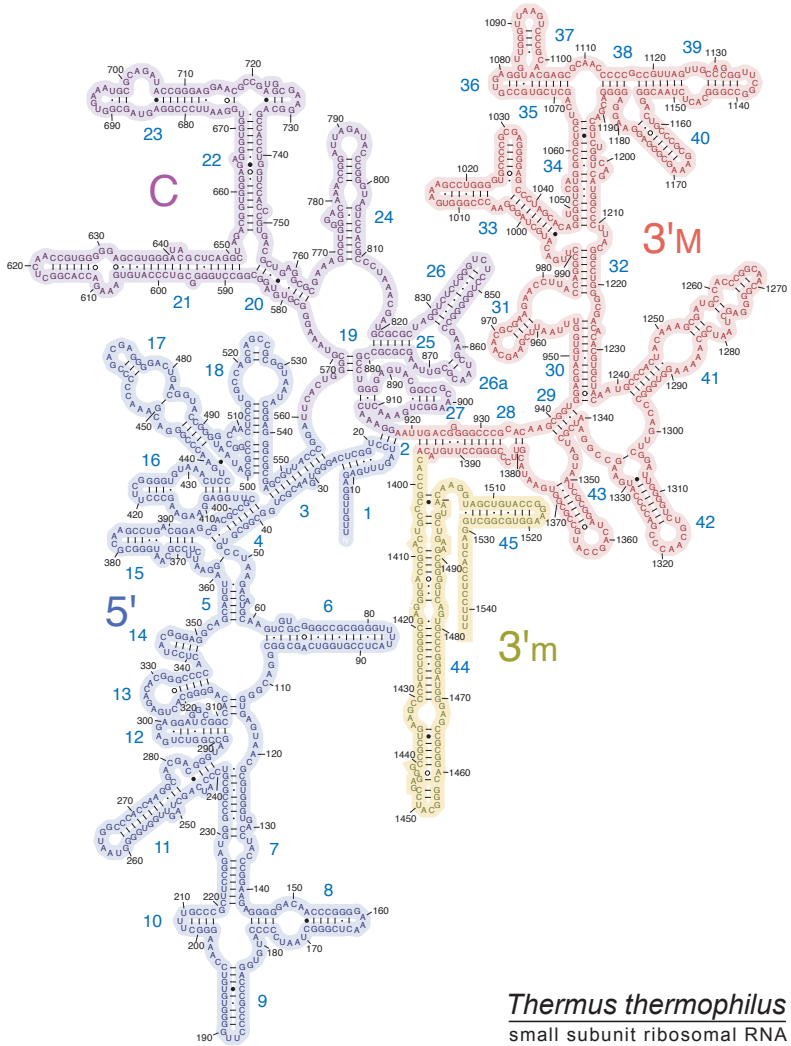
Лунина Полина Сергеевна, 571 группа
Научный руководитель: доцент, к.ф-м.н. Григорьев С.В.

Санкт-Петербургский государственный университет
Кафедра системного программирования

10 июня 2020г.

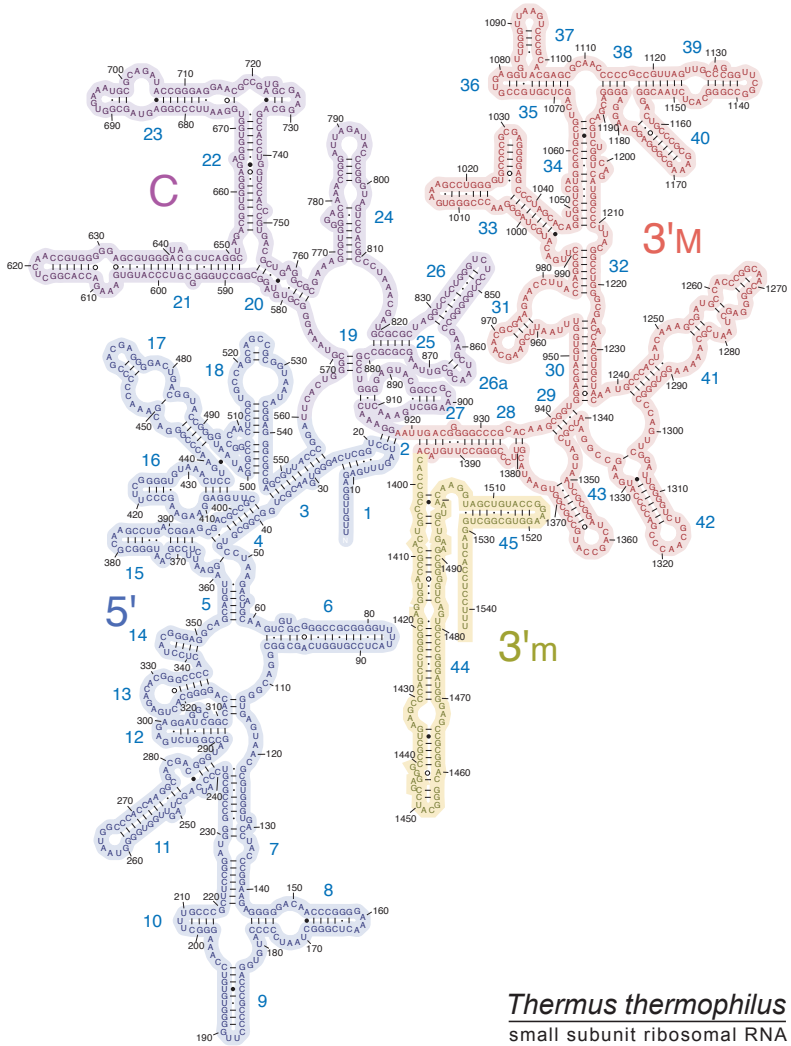
Биоинформатика

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



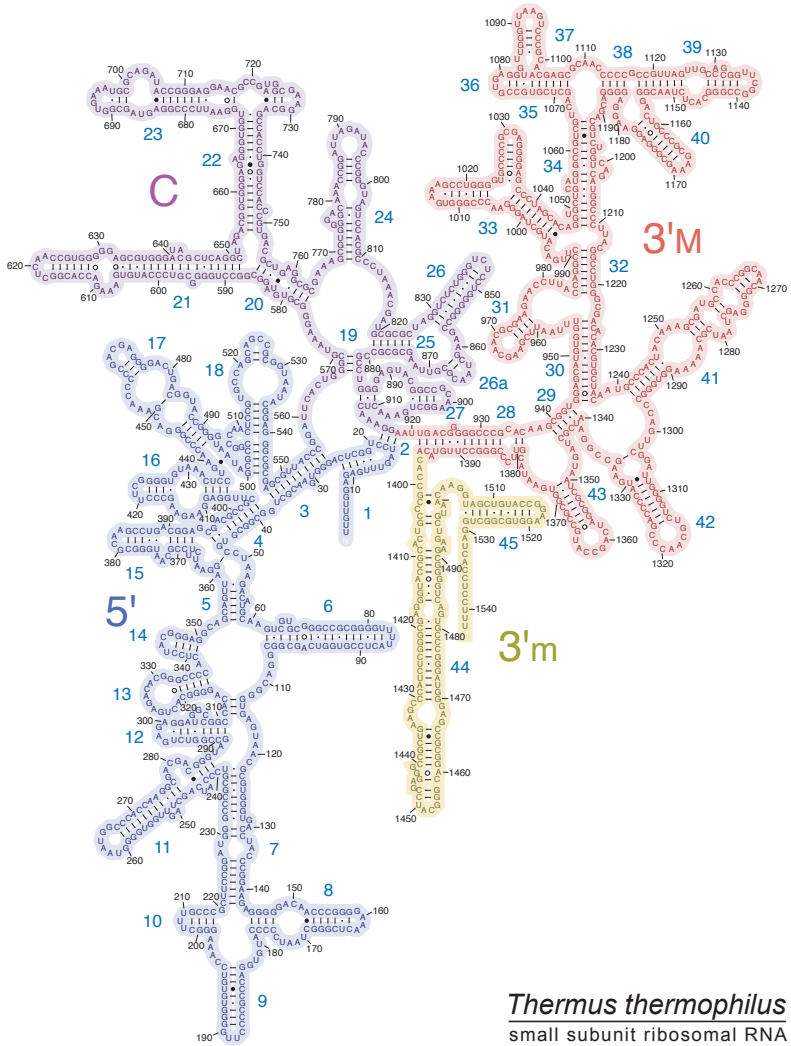
Биоинформатика

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



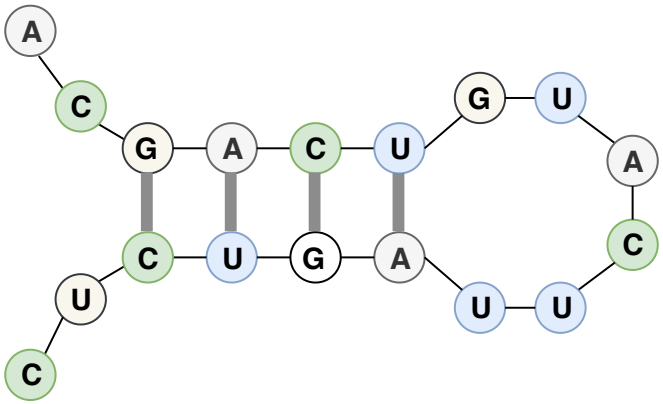
Биоинформатика

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



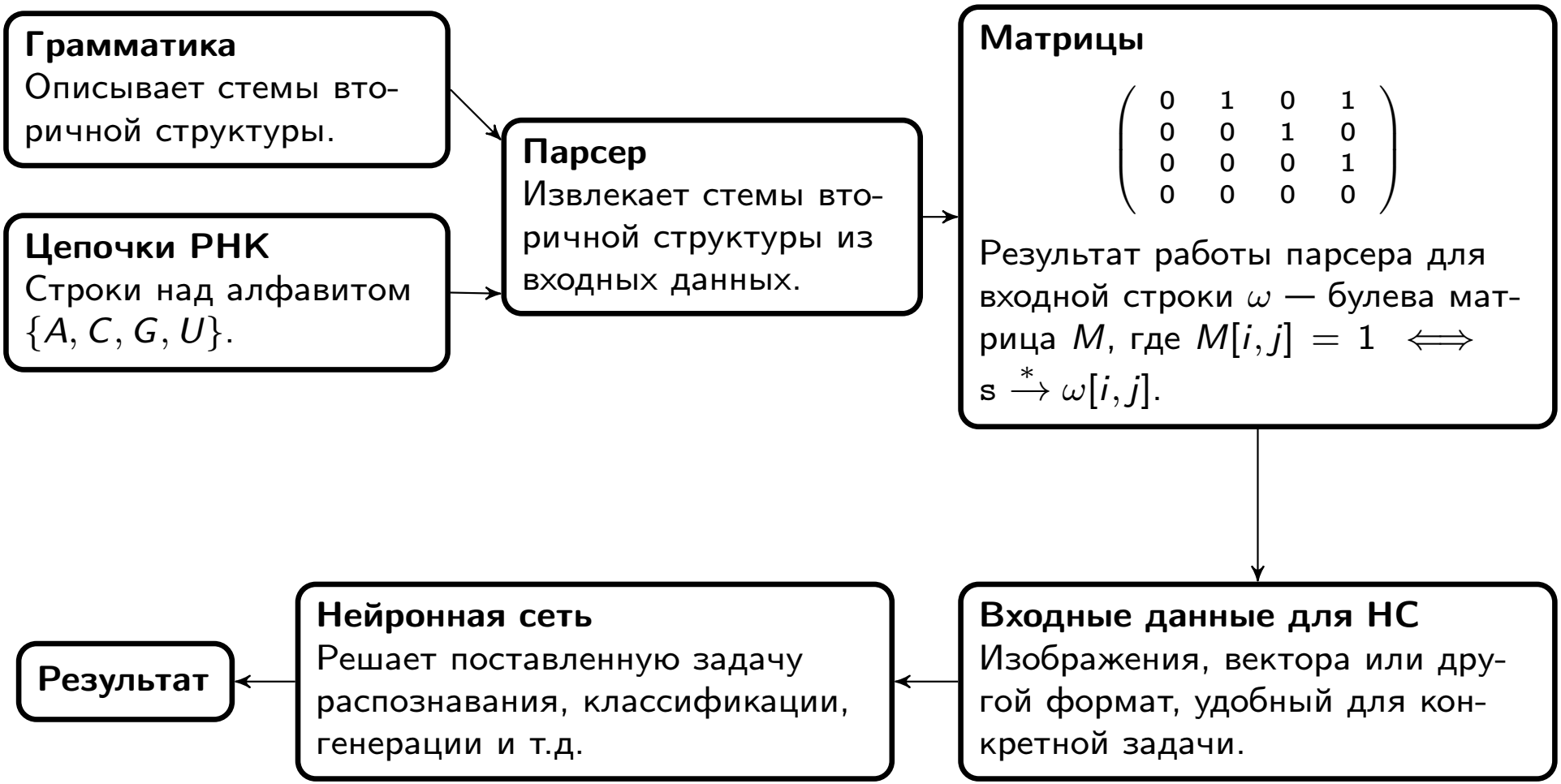
Наш подход

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

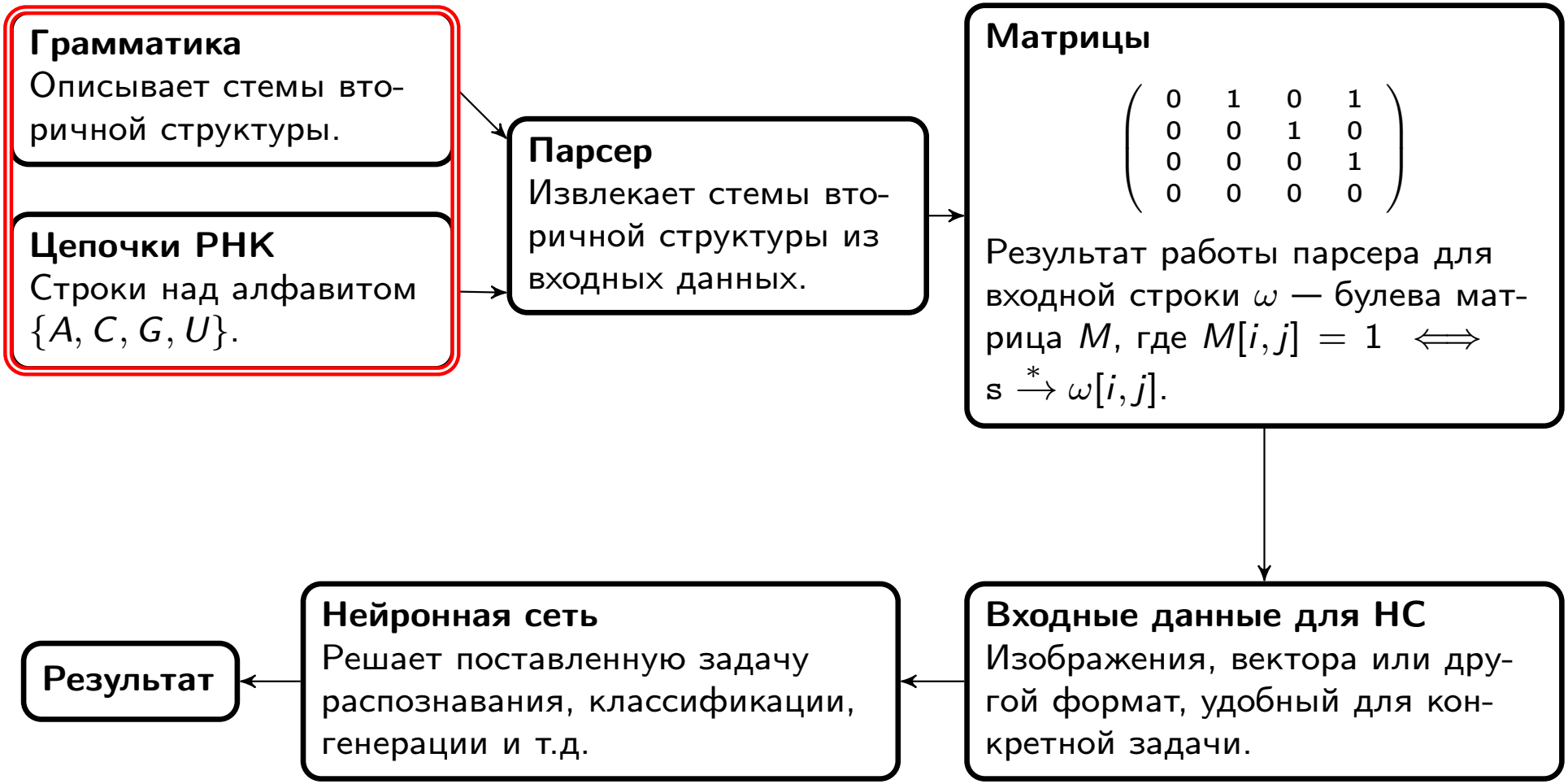



```
s1: stem<s0>
s0: G U A C U U
stem<s>:
      A s U
    | G s C
    | U s A
    | C s G
```

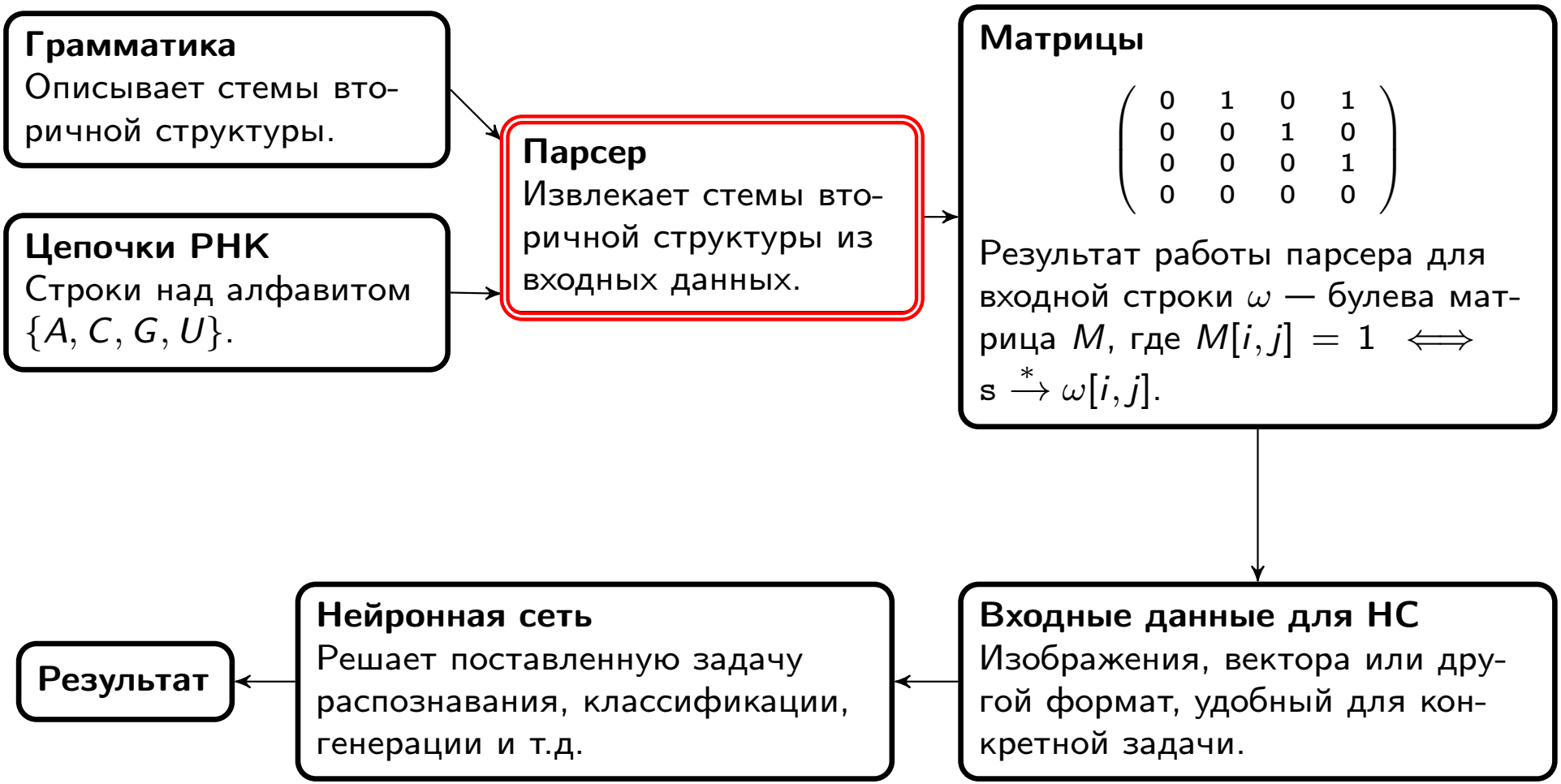
Наш подход


Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

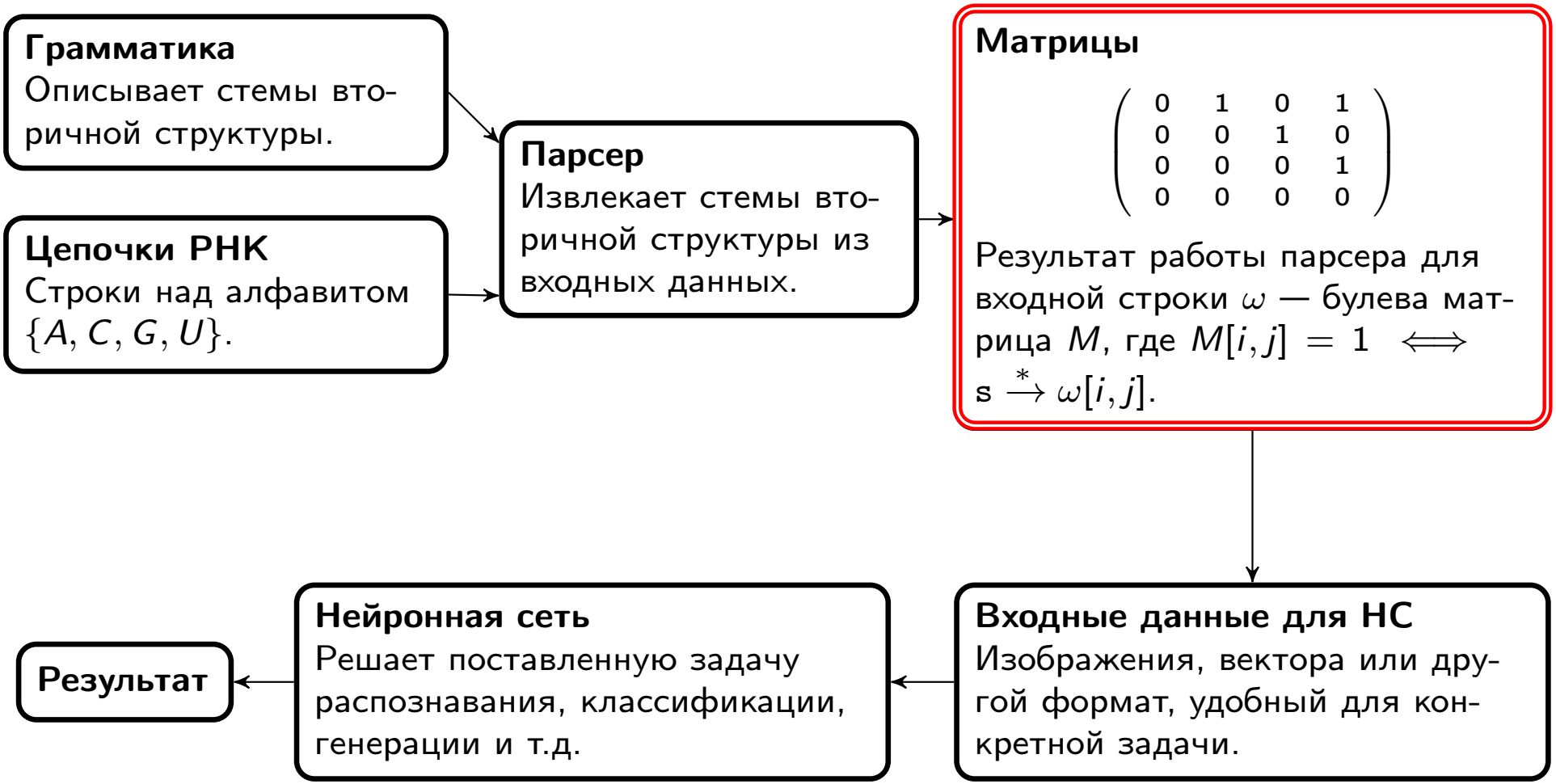
Наш подход


Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

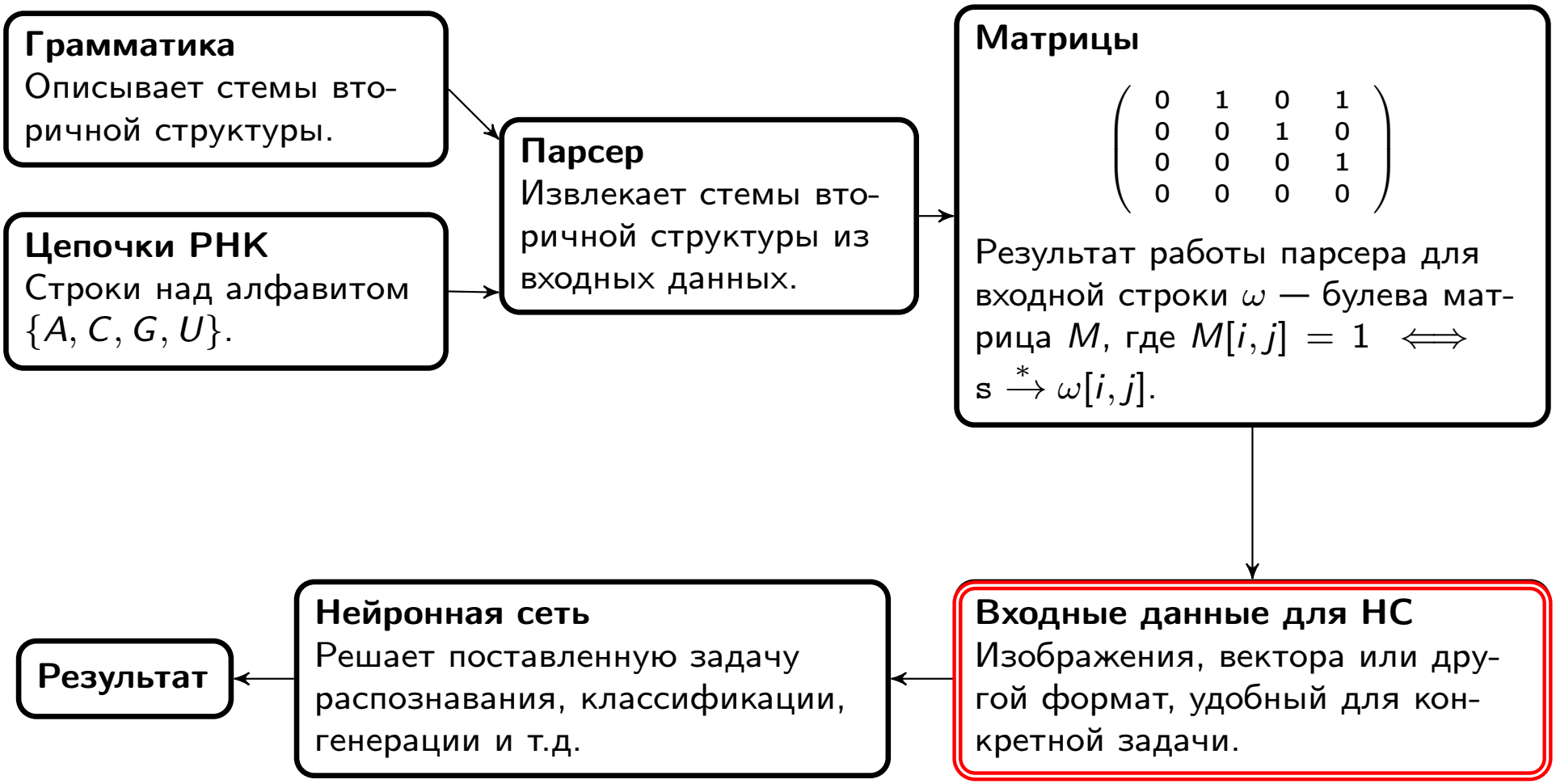
Наш подход


Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

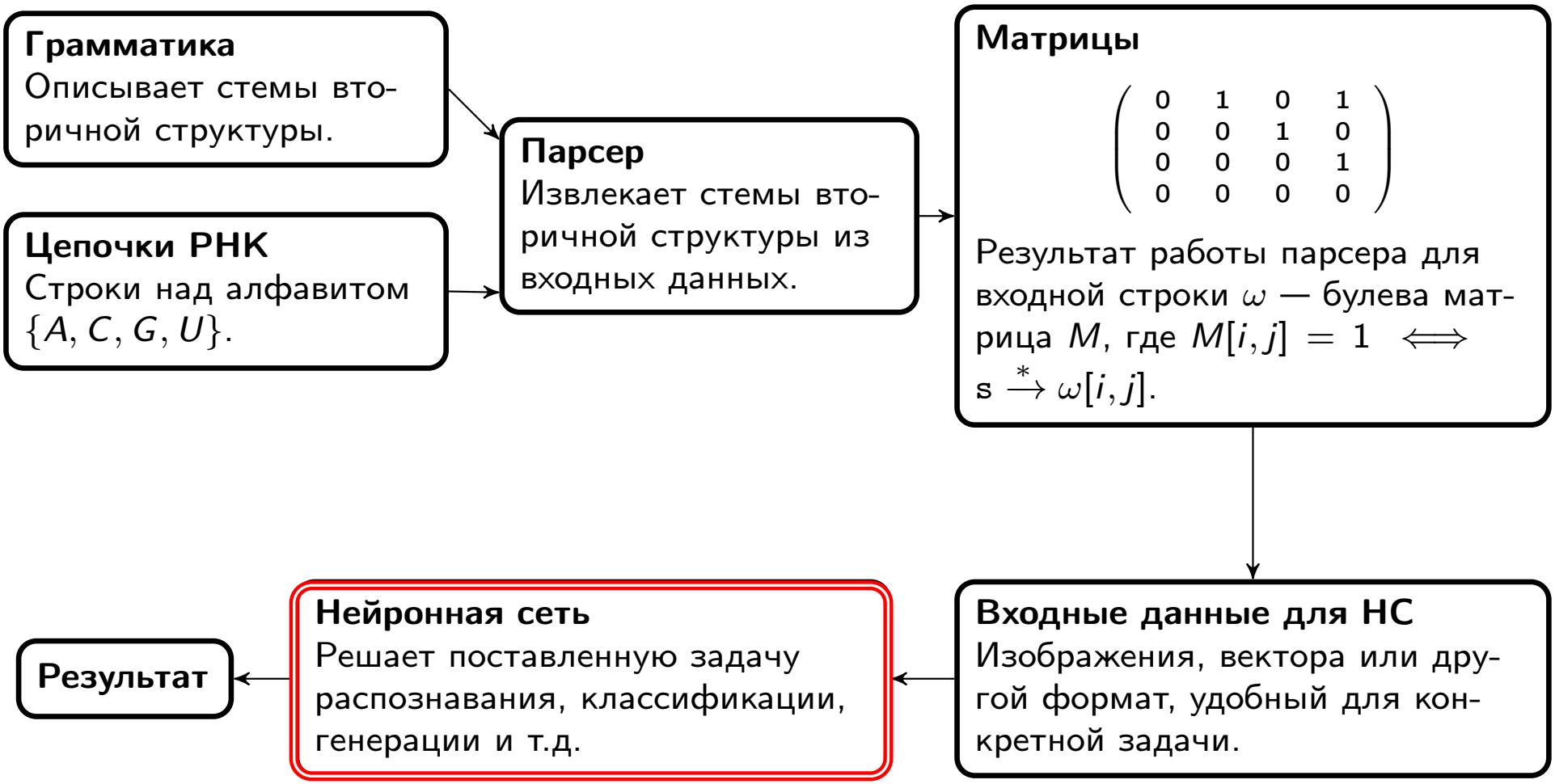
Наш подход


Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

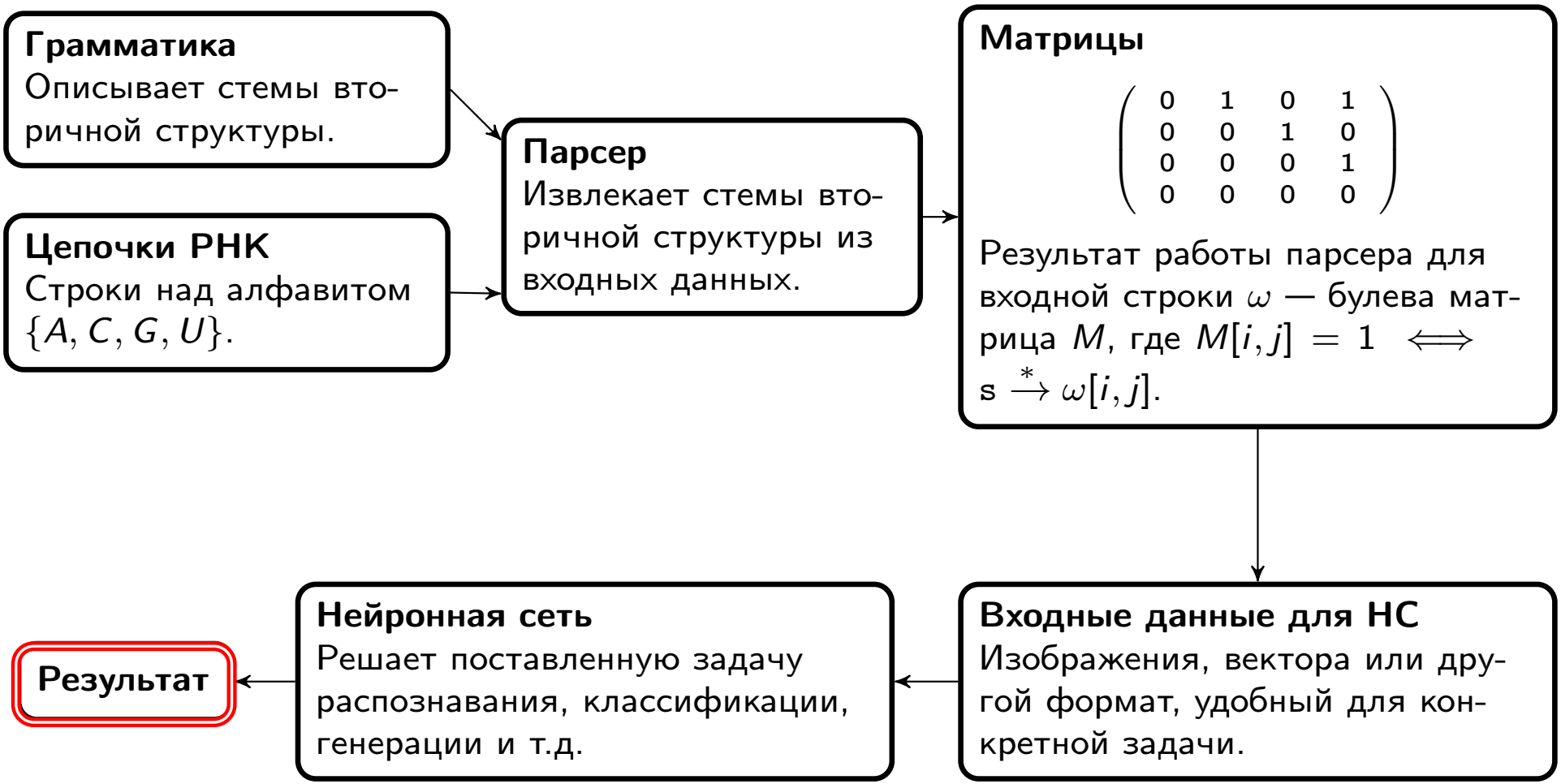
Наш подход


Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

Наш подход


Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

Наш подход

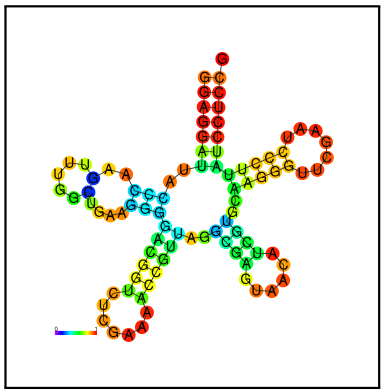

Грамматика
Описывает стемы вторичной структуры.
Цепочки РНК
Строки над алфавитом $\{A, C, G, U\}$.
Парсер
Извлекает стемы вторичной структуры из входных данных.
Матрицы
$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
Результат работы парсера для входной строки ω — булева матрица M, где $M[i,j] = 1 \iff s \xrightarrow{*} \omega[i,j]$.
Входные данные для НС
Изображения, вектора или другой формат, удобный для конкретной задачи.
Нейронная сеть
Решает поставленную задачу распознавания, классификации, генерации и т.д.
Результат

Идея текущего исследования

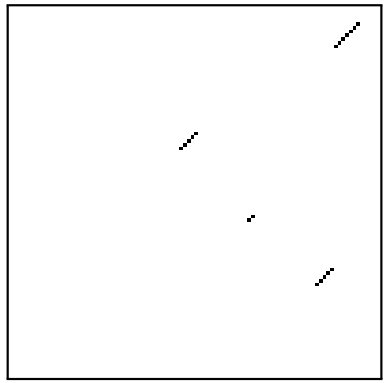
- Парсер находит в цепочке все возможные стемы, однако не все они действительно будут входить в состав вторичной структуры
- Хотим сконструировать нейронную сеть, которая отфильтрует лишние контакты между нуклеотидами и предскажет вторичную структуру цепочки

Идея текущего исследования

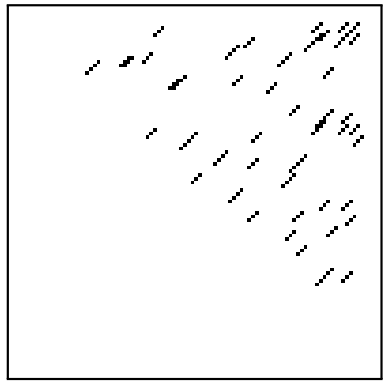
- Парсер находит в цепочке все возможные стемы, однако не все они действительно будут входить в состав вторичной структуры
- Хотим сконструировать нейронную сеть, которая отфильтрует лишние контакты между нуклеотидами и предскажет вторичную структуру цепочки



Вторичная структура



Матрица контактов



Матрица разбора

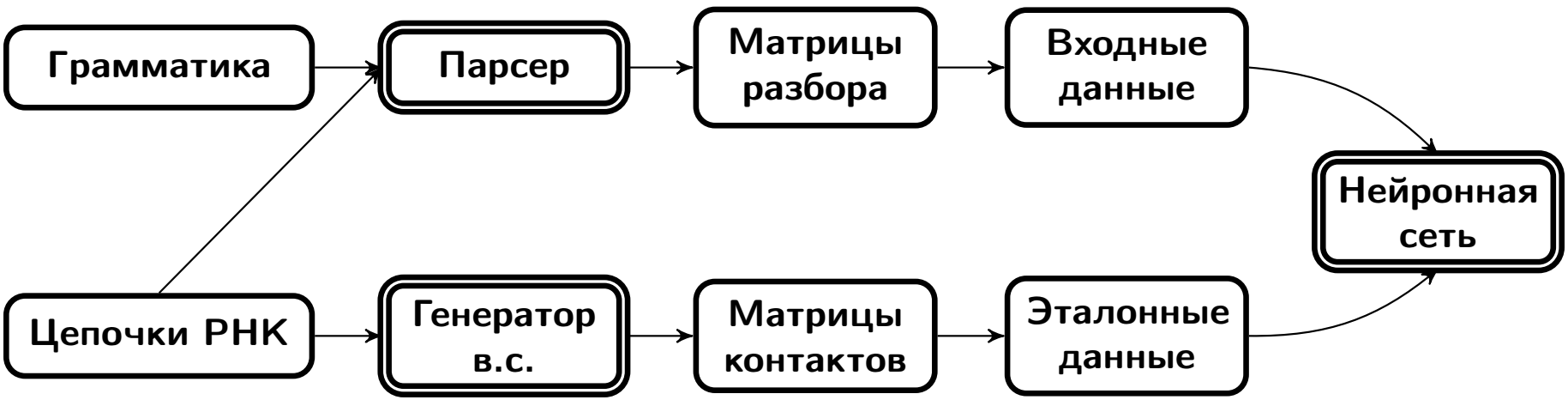
Постановка задачи

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

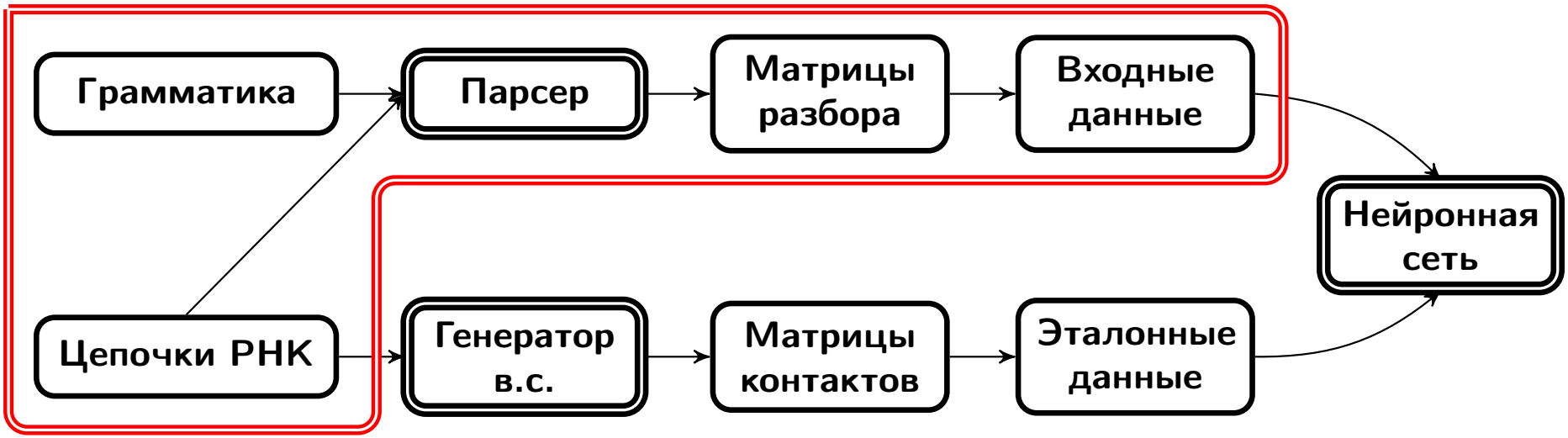
Задачи

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

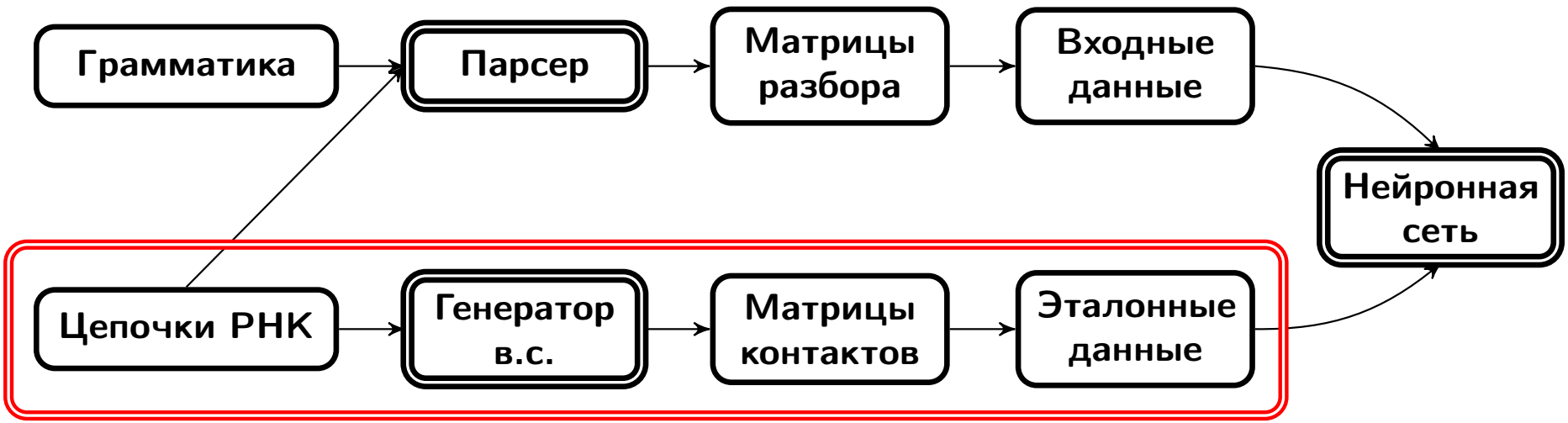
Архитектура решения



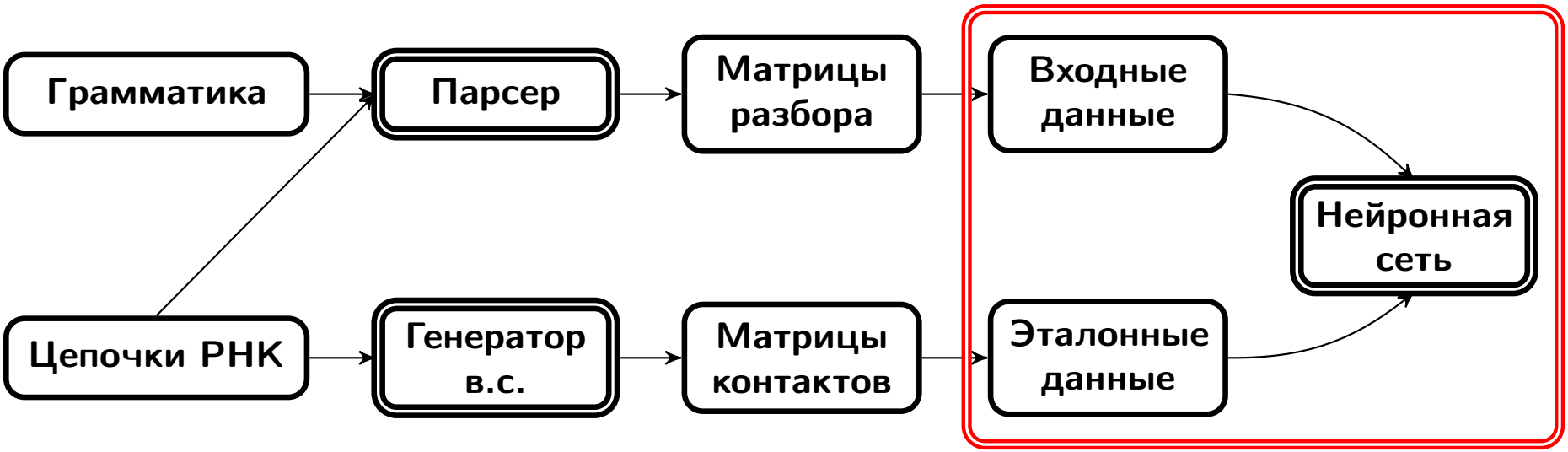
Архитектура решения



Архитектура решения

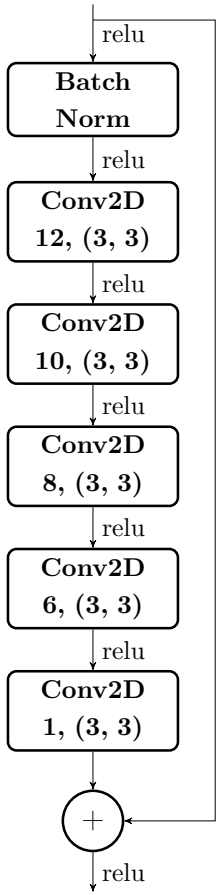


Архитектура решения



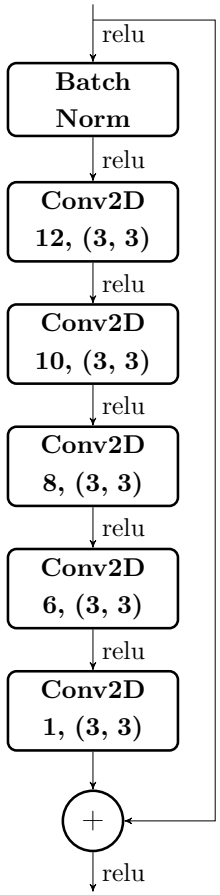
Нейронная сеть

- Fully convolutional residual neural network
- Loss — взвешенная попиксельная разница
- Алгоритм локального выравнивания последовательностей
- train/valid/test = 70%/10%/20%



Нейронная сеть

- Fully convolutional residual neural network
- Loss — взвешенная попиксельная разница
- Алгоритм локального выравнивания последовательностей
- train/valid/test = 70%/10%/20%
- Метрики
  - *Precision* — сколько из предсказанных контактов действительно являются контактами в эталоне
  - *Recall* — сколько из требуемых контактов было найдено
  - *F*1 *score* — объединяющая метрика



Эксперименты

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

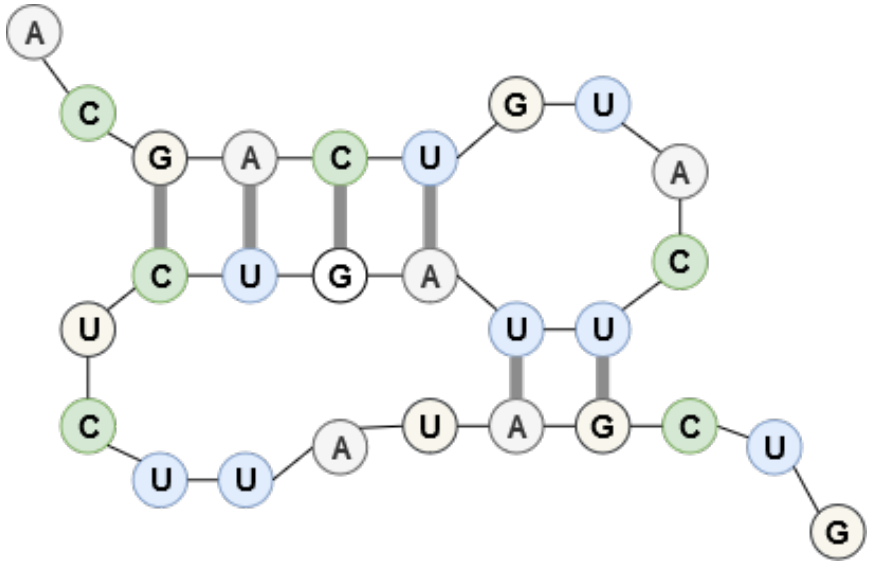
Предсказание вторичных структур для цепочек с разными интервалами длин

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

Результаты

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

Планы

- 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.