Санкт-Петербургский государственный университет

Программная инженерия

Лунина Полина Сергеевна

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

Курсовая работа

Научный руководитель:
к. ф.-м. н., доцент Григорьев С. В.

Санкт-Петербург
2020

SAINT-PETERSBURG STATE UNIVERSITY

Software Engineering

Polina Lunina

The composition of neural networks and parsing for genomic sequences secondary structure prediction

Course Work

Scientific supervisor:
Assistant Professor Semyon Grigorev

Saint-Petersburg
2020

Оглавление

Введение

В биоинформатике существует множество задач анализа генетических цепочек, например, классификация организмов по их генетическим данным, поиск подпоследовательностей и другие. Для решения этих задач требуется разработка новых алгоритмов и совершенствование существующих, и при этом в основе различных методов анализа биологических данных лежат некоторые общие базовые принципы.

Молекула РНК представляет собой цепочку нуклеотидов (первичная структура), и в том случае, когда два фрагмента этой цепи соединяются друг с другом, перегибаясь и образуя на конце неспаренный участок — петлю, формируется элемент, называемый в биологии стемом. И совокупность стемов различных размеров, а также вложенных стемов составляет сложную и стабильную вторичную структуру. Известно, что вторичная структура содержит в себе важную для идентификации организма информацию, поэтому среди алгоритмов для решения различных задач анализа цепочек РНК наибольшим успехом пользуются те, что ее каким-либо образом учитывают. Существуют различные методы формального описания вторичной структуры, например, скрытые марковские модели, ковариационные модели [2] и формальные грамматики [5, 8, 10].

При работе с биологическими данными важно учитывать присутствие в них шумов и мутаций, что делает точные алгоритмы неприменимыми и требует проведения некоторой вероятностной оценки. Популярным способом обработки зашумленных данных являются методы машинного обучения, в частности, нейронные сети, которые в настоящее время успешно используются в биоинформатике [6, 12].

В рамках предыдущей дипломной работы был предложен новый подход для решения задач обработки последовательностей, обладающих некоторой синтаксической структурой. Данный подход основан на комбинировании методов синтаксического анализа и машинного обучения. Предлагается использовать грамматику для кодирования характерных особенностей синтаксической структуры, алгоритм синтаксиче-

ского анализа — для их поиска во входных данных, а обработку информации о наличии и расположении этих особенностей в цепочке и вероятностную оценку провести с помощью нейронной сети, сконструированной для решения конкретной задачи. И, применительно к генетическим данным, синтаксической структурой является вторичная структура РНК, а искомыми особенностями — составляющие ее стемы.

Направлением исследования, представленного в данной работе, является предсказание вторичных структур РНК с использованием разработанного в предыдущей работе подхода. Синтаксический анализатор находит во входной строке все подстроки, которые удовлетворяют правилам образования стемов, описанным в грамматике. Однако такое множество подстрок описывает все теоретически возможные в заданной последовательности стемы, закодированные средствами используемой грамматики, а реальная вторичная структура РНК живого организма содержит только небольшое их подмножество. Кроме того, существуют более сложные особенности вторичной структуры, невыразимые простой контекстно-свободной грамматикой, такие как псевдоузлы и неклассические пары комплиментарности оснований. Поэтому для генерации чистой вторичной структуры из результата работы парсера в рамках предложенного подхода предлагается использовать нейронную сеть, задача которой в данном случае — отфильтровать лишние стемы и достроить невыразимые в грамматике элементы.

1. Постановка задачи

Целью данной работы является исследование возможности применения предложенного в предыдущей дипломной работе подхода к задаче предсказания вторичных структур геномных последовательностей. Для реализации данной цели были поставлены следующие 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.

2. Обзор используемого подхода

Предложенный в предыдущей дипломной работе подход для анализа вторичной структуры последовательностей основан на комбинировании нейронных сетей и синтаксического анализа. Основная идея заключается в том, что характерные элементы вторичной структуры последовательностей необходимо формально описать средствами простой (контекстно-свободной) грамматики, а затем для поиска в некоторой строке подстрок, подходящих под это описание (формально — выводимых из стартового нетерминала грамматики), использовать матричный алгоритм синтаксического анализа. Результат работы такого алгоритма — матрица разбора — будет хранить информацию о выводимости всех подстрок данной строки в используемой грамматике, что с практической точки зрения означает наличие и взаимное расположение искомых характерных элементов. Завершающим этапом в рамках предложенного подхода является обработка матриц разбора, преобразованных в некоторый удобный формат, с помощью искусственных нейронных сетей для решения конкретной поставленной задачи (например, классификация генетических цепочек).

Опишем более формально описанные выше концепции применительно к задачам в области биоинформатики, где рассматриваемыми последовательностями являются цепочки РНК различных организмов, а кодируемым в грамматике элементом является стем вторичной структуры. Первичная структура молекулы РНК представляет собой последовательность нуклеотидов, т.е. символов алфавита {A, C, G, U}. Некоторые участки такой последовательности могут соединяться между собой, образуя характерные элементы — стемы, состоящие из спаренного участка (стебля) и неспаренного (петли), как показано на рис. 1. Вторичная структура молекулы РНК состоит из комбинации стемов различного размера и степени вложенности, и общий вид таких конструкций необходимо формализовать в грамматике.

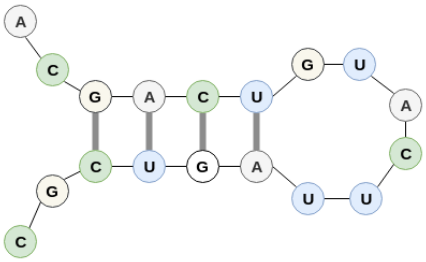


Рис 1: Образование вторичной структуры молекулы РНК

2.1. Грамматика

В данной работе для описания стемов вторичной структуры была использована контекстно-свободная грамматика G_0 (рис. 2), предложенная в предыдущей работе. Эта грамматика учитывает классические правила формирования нуклеотидных пар $A-U$, $C-G$ (строка **5**) и описывает рекурсивные композиции стемов высоты от трех и более (строки **7-12**). Размер петли внутри стема лежит в пределах от двух до десяти нуклеотидов, такую же длину имеют последовательности, расположенные между любыми двумя стемами (строка **2**). Эти числа были получены путем экспериментов и теоретических исследований реальных вторичных структур РНК и могут быть изменены в рамках решения конкретной задачи.

```
1  s1: stem<s0>
2  any_str : any_smb*[2..10]
3  s0: any_str | any_str stem<s0> s0
4  any_smb: A | U | C | G
5  stem1<s>: A s U | G s C | U s A | C s G
6  stem2<s>: stem1< stem1<s> >
7  stem<s>:
8        A stem<s> U
9      | U stem<s> A
10     | C stem<s> G
11     | G stem<s> C
12     | stem1< stem2<s> >
```

Рис 2: Контекстно-свободная грамматика G_0 для описания стемов вторичной структуры РНК

2.2. Синтаксический анализатор

Синтаксический анализ в данном случае используется для поиска всех подстрок некоторой строки, выводимых из стартового нетерминала грамматики G_0, т.е. тех участков этой строки, которые в терминах G_0 являются стемами вторичной структуры. Формально, для входной строки w парсер сформирует верхнетреугольную булеву матрицу разбора M, где $M[i, j] = 1$, тогда и только тогда, когда подстрока $w[i, j]$ выводима из стартового нетерминала грамматики.

В данной работе был использован разработанный в рамках проекта YaccConstructor [14] в лаборатории JetBrains [7] алгоритм, основанный на матричных операциях [1], который демонстрирует высокую производительность на практике в связи с использованием параллельных вычислений.

2.3. Искусственные нейронные сети

Для анализа данных, проведения вероятностной оценки и генерации конечного результата в рамках некоторой задачи далее используются нейронные сети. Входными данными здесь являются матрицы, полученные парсером (при необходимости подвергнутые некоторой постобработке и приведенные в удобный формат), а выходные данные зависят от особенностей конкретного исследования, например, класс, к которому принадлежит каждая цепочка для проблемы классификации организмов. Архитектура сетей, их количество и шаги, предпринятые для промежуточной обработки всех данных, также зависят от специфики поставленной задачи.

3. Архитектура решения

3.1. Формальное представление вторичной структуры молекулы РНК

Вторичная структура молекулы РНК представляет собой комбинацию вложенных стемов различной высоты и размера петли. Пример вторичной структуры транспортной РНК (тРНК) представлен на рис. 3а. Формально такой объект можно представить разными способами, но самыми популярными являются следующие два.

- Матрица контактов, описывающая наличие или отсутствие связи между каждыми двумя нуклеотидами — верхнетреугольная матрица M, где $M[i, j] = 1$ тогда и только тогда, когда между символами цепочки в позициях i и j существует контакт во вторичной структуре.

- Строка в формате dot-bracket, где каждому символу цепочки сопоставлен один из символов алфавита {.(){}<>}. Здесь точка обозначает неспаренный нуклеотид, открытая скобка указывает на то, что нуклеотид сопряжен с некоторым другим впереди него, а закрытая — на то, что нуклеотид сопряжен с некоторым другим позади него. Другие виды скобок предназначены для представления более сложных элементов, в частности, псевдоузлов.

Такие форматы описания вторичных структур повсеместно используются как в базах данных вторичных структур живых организмов, так и в различных инструментов для их предсказания.

3.2. Задача предсказания вторичной структуры РНК в рамках предложенного подхода

В данном разделе рассмотрим общую мотивацию применения подхода, основанного на комбинации методов синтаксического анализа и

машинного обучения, к задаче предсказания вторичных структур РНК различных организмов.

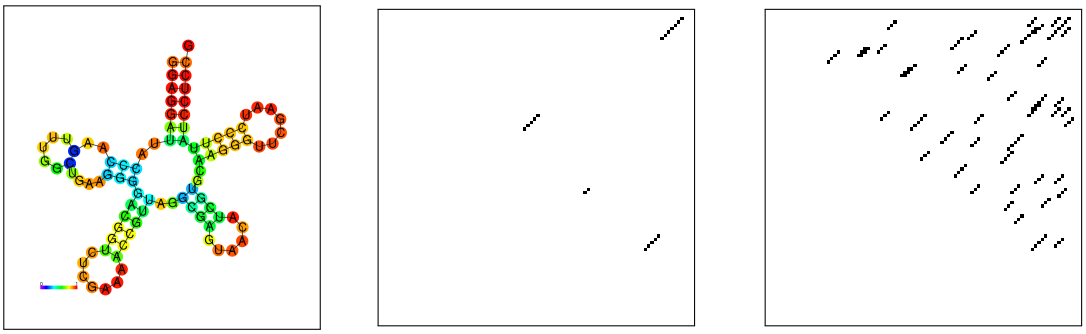
Синтаксический анализатор находит все теоретически возможные в правилах заданной грамматики стемы, однако реальная вторичная структура содержит далеко не все из них, так как грамматикой мы описали только общий вид стема, но не учли особенности их взаимного расположения и различные термодинамические и биологические законы, которые тяжело поддаются формализации. Поэтому для использования предложенного подхода в рамках задачи генерации чистой вторичной структуры живого организма необходимо обработать полученные матрицы разбора и отфильтровать лишние стемы и, возможно, достроить те элементы вторичной структуры, которые невыразимы средствами данной грамматики. Например, важным элементом вторичной структуры РНК являются так называемые псевдоузлы, которые состоят из двух шпилек, где половина стебля одной шпильки располагается между двумя половинами стебля другой шпильки. Однако псевдоузлы невыразимы средствами контекстно-свободных грамматик. Кроме того, в реальных вторичных структурах иногда встречаются неклассические пары нуклеотидов (например, $U - G$), но включение их в грамматику приведет к кратному увеличению количества правил и, как следствие, к существенным затратам по времени на синтаксический анализ.

Таким образом, для преобразования результата работы парсера в корректную вторичную структуру предлагается использовать нейронную сеть, для которой входным слоем будет матрица разбора в некотором формате, а выходным — сгенерированная по ней вторичная структура.

Как было описано в разделе 2.2, результат работы парсера на входной строке — верхнетреугольная булева матрица, где единица в позиции $[i, j]$ означает тот факт, что подстрока этой строки от i до j свернется в стем по правилам грамматики. В то же время, в разделе 3.1 было рассмотрено матричное представление вторичной структуры, в котором единица в позиции $[i, j]$ указывает на наличие контакта между нуклеотидами i и j. Несложно заметить, что данные обозначения эк-

виваленты, поэтому, обрабатывая матрицы разбора, мы имеем дело с данными в классическом формате представления вторичной структуры. Примеры таких матриц для реальной вторичной структуры тРНК продемонстрированы на рис. 3.

Удобным форматом данных для обучения нейронной сети являются изображения, поэтому мы предлагаем привести матрицу разбора и матрицу контактов к черно-белым изображениям следующим образом: единицы заменить на белые пиксели, а нули на черные. И такие изображения использовать для обучения и тестирования нейронных сетей.



(a) Вторичная структура (b) Матрица контактов (c) Матрица разбора

Рис 3: Пример представления вторичной структуры тРНК

3.3. Схема решения

Общая схема решения для предсказания вторичной структуры РНК в рамках данной работы представлена на рис. 4 и состоит из следующих ключевых этапов.

- Подготовка эталонных данных.
  - Поиск базы последовательностей РНК, на которой будут проводиться экспериментальные исследования.
  - Сбор эталонных вторичных структур для всех последовательностей с помощью некоторого выбранного инструмента или базы реальных вторичных структур.
  - Промежуточная обработка, преобразование структур в матрицы контактов, а затем в черно-белые изображения.

  - Подготовка выборок train, valid и test для нейронных сетей.

- Подготовка анализируемых данных.
  - Применение алгоритма синтаксического анализа на выбранной базе цепочек РНК.
  - Преобразование полученных матриц разбора в черно-белые изображения.
  - Подготовка выборок train, valid и test для нейронных сетей.

- Анализ данных с помощью нейронных сетей.
  - Разработка архитектуры сети, генерирующей по входному изображению максимально близкое к эталонному.
  - Обучение и тестирование по различным метрикам.

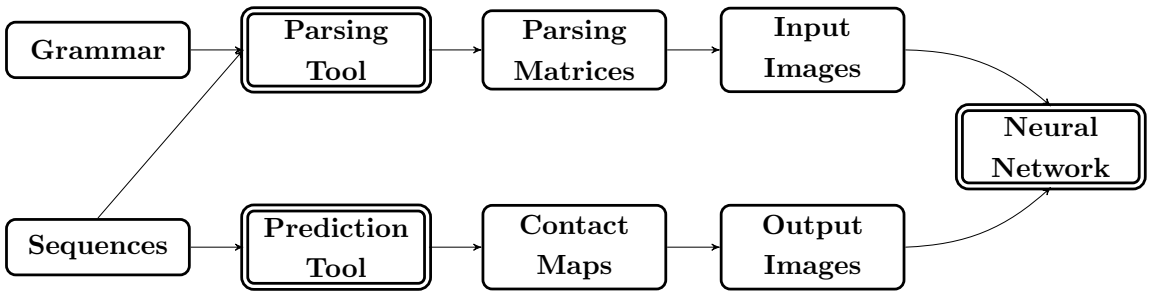


Рис 4: Схема решения для предсказания вторичной структуры РНК

3.4. Нейронные сети

Для создания и обучения нейронных сетей в данной работе были использованы библиотека Keras [4] и фреймворк Tensorflow [13].

Рассмотрим общую модель генеративной нейронной сети, разработанной в рамках данной работы. Входными и выходными данными являются изображения, и в рамках поставленной задачи необходимо найти достаточно сложные закономерности между элементами данных, на-

ходящимися на большом расстоянии друг от друга, поэтому была использована глубокая сверточная нейронная сеть. Для оптимизации процесса обучения и повышения скорости сходимости была использована технология остаточных нейронных сетей (residual networks), которая основана на добавлении дополнительных связей между отделенными друг от друга слоями. Типичный блок (residual unit) сети, используемой в описанных ниже экспериментах представлен на рис. 5.

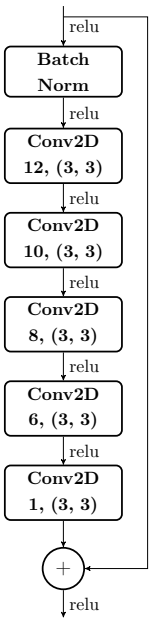


Рис 5: Residual unit

Для вычисления ошибки нейронной сети при обучении была использована просуммированная попиксельная разница эталона и предсказанного изображения с некоторым коэффициентом $k > 1$ при белых пикселях в эталонных изображениях. Наличие коэффициента было обусловлено тем, что белые пиксели, обозначающие контакт в матрице, составляют малую часть изображения, следовательно, взятие чистой попиксельной разницы не было бы чувствительно к небольшим ошибкам в белых пикселях, которые, тем не менее, привели бы к существенным ошибкам в плане предсказания вторичной структуры.

Формально, если p_e — нормированное значение пикселя эталонного изображения, а p_p — нормированное значение пикселя предсказанного

изображения, то

$$WeightedLoss = \frac{|\sum \omega - \sum \omega * \delta|}{\sum \omega},$$

$$\omega = (k-1) * p_e + 1, \delta = 1 - |p_e - p_p|.$$

3.4.1. Метрики

Для тестирования обученных нейронных сетей были выбраны следующие метрики относительно оценки попиксельной разницы между предсказанным и эталонным изображениями. Далее TW (true white), TB (true black), FW (false white) и FB (false black) — информация о том, сколько раз нейронная сеть приняла верное и сколько раз неверное решение по каждому пикселю каждого изображения выборки.

- $Precision = \frac{TW}{TW+FW}$ (какая доля предсказанных контактов действительно является контактами в эталоне).
- $Recall = \frac{TW}{TW+FB}$ (какая доля искомых контактов была найдена).
- $FMera = 2 * \frac{Precision*Recall}{Precision+Recall}$ (гармоническое среднее $Precision$ и $Recall$ — объединяющая метрика).

4. Эксперименты

4.1. Предсказание вторичных структур тРНК без псевдоузлов

Для экспериментального исследования применимости предложенной архитектуры к реальным биологическим данным были поставлены задачи предсказания вторичной структуры тРНК одинаковой длины и тРНК, длины которых находятся в некотором фиксированном интервале.

Задачи, связанные с подготовкой данных, были выполнены в рамках курсовой работы Кутленкова Дмитрия Александрович на кафедре системного программирования. Цепочки тРНК были взяты из базы данных RNAcentral [11], а эталонные структуры были получены при помощи инструмента CentroidFold [3]. В данном инструменте не реализована возможность предсказания псевдоузлов, однако простота и удобство использования позволили выбрать его в качестве источника данных для первых экспериментов. Были рассмотрены три варианта длин последовательностей: 90, 88-90, 50-90 и для каждого эксперимента собранные цепочки были поделены на выборки в соотношении train:valid:test = 70%:10%:20%. В экспериментах с вариативной длиной цепочки изображения при обучении распределялись на батчи так, чтобы в каждом батче присутствовали изображения только одного размера. Использованная во всех экспериментах модель состояла из десяти residual блоков, конструкция которых описана в разделе 3.4.

Кроме того, была исследована возможность применения алгоритма локального выравнивания последовательностей, разработанного и реализованного в рамках курсовой работы Кутленкова Дмитрия Александровича на кафедре системного программирования. Данный алгоритм на основании сгенерированного нейронной сетью изображения получает вторичную структуру, удовлетворяющую биологическим законам. Однако непосредственное применение алгоритма к результатам работы нейронной сети привело к падению точности, поэтому было принято

решение сделать его адаптивным, т.е. встроить выравниватель как финальный слой нейронной сети и дообучить оригинальную модель с выравнивающей надстройкой. При дообучении была использована другая функция loss: линейная комбинация $WeightedLoss$ и $1 - FMera$.

Обученные нейронные сети были протестированы в соответствии с метриками, заданными в разделе 3.4.1 данной работы. Метрики $Precision$, $Recall$ и $FMera$ были вычислены для каждого изображения, а затем взяты средние значения по выборке. Результаты тестирования оригинальной и комбинированной нейронных сетей по данным метрикам для каждой задачи представлены в таблице 1.

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%

Таблица 1: Результаты тестирования моделей для данных без псевдоузлов

Как видно по таблице 1, предложенный подход применим к задаче предсказания вторичной структуры тРНК и можно сделать следующие выводы.

- Чем больше разброс длин, тем менее точным получается результат и тем больше требуется данных для обучения. Поэтому вероятно, что понадобится некоторая архитектура, позволяющая предобучать нейронные сети на цепочках близкой или равной длины.

- Оригинальная нейронная сеть дает $Recall$, более высокий, чем $Precision$, в то время как нейронная сеть с выравниванием — наоборот, что говорит о том, что первая хорошо предсказывает контакты, но также находит много лишних, а вторая достаточно точно отсеивает эти лишние контакты, повышая биологическую

корректность результата, при этом, однако, удаляя и часть ценной информации. Поэтому следует найти некоторый баланс между возможностями двух реализованных моделей.

Таким образом, полученные результаты говорят о том, что теоретически предсказание вторичной структуры для цепочек тРНК в рамках предложенного подхода возможно, однако необходимы дальнейшие эксперименты, изучение техники выравнивания и более тонкая настройка параметров модели.

4.2. Предсказание вторичных структур тРНК с псевдоузлами

Одной из задач, поставленных в данной работе, было изучение возможности предсказания вторичной структуры для цепочек, содержащих псевдоузлы. Псевдоузлы невыразимы средствами контекстно-свободных грамматик, поэтому было необходимо проверить, насколько точно можно достроить их по матрице разбора с помощью нейронной сети.

В данном эксперименте были взяты цепочки со вторичными структурами из базы Pseudobase [9], train:valid:test = 70%:10%:20. Цепочки имели различные длины, поэтому были использованы веса модели для длин 50-90 из раздела 4.1 с последующем дообучением. Тестирование проводилось по метрикам $Precision$, $Recall$ и $FMera$ и результаты продемонстрированы в таблице 2

Length	Samples	Alignment	Precision	Recall	F1 score
50-90	266	×	74%	73%	71%

Таблица 2: Результаты тестирования моделей для данных с псевдоузлами

Можно увидеть, что данная модель показала лучший результат, чем та, на которой она была основана, что позволяет полагать, что у задачи предсказания псевдоузлов есть потенциал.

Заключение

В данной работе было проведено исследование возможности применения подхода, основанного на комбинировании формальных грамматик и нейронных сетей, к задаче предсказания вторичных структур РНК. Были получены следующие результаты.

- Разработана архитектура решения.
- Проведены экспериментальные исследования применительно к задачам предсказания вторичных структур транспортных РНК.

Направлениями дальнейших исследований являются следующие задачи.

- Более тщательная разработка модели, применяющей адаптивное выравнивание.
- Повышение точности нейронной сети, обучающейся на изображениях произвольного размера.
- Реализация системы тестирования результатов работы нейронных сетей на данных с псевдоузлами.
- Поиск новых средств, а также более тонкая настройка параметров всех моделей для улучшения результатов.

Список литературы

[1] Azimov Rustam, Grigorev Semyon. Context-free Path Querying by Matrix Multiplication // Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA). — GRADES-NDA '18. — New York, NY, USA : ACM, 2018. — P. 5:1–5:10. — Access mode: http://doi.acm.org/10.1145/3210259.3210264.

[2] Biological sequence analysis: probabilistic models of proteins and nucleic acids / Richard Durbin, Sean R Eddy, Anders Krogh, Graeme Mitchison. — Cambridge university press, 1998.

[3] CENTROIDFOLD: a web server for RNA secondary structure prediction / Kengo Sato, Michiaki Hamada, Kiyoshi Asai, Toutai Mituyama // Nucleic acids research. — 2009. — Vol. 37, no. suppl_2. — P. W277–W280.

[4] Chollet François et al. Keras. — https://keras.io. — 2015.

[5] Dowell Robin D, Eddy Sean R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction // BMC bioinformatics. — 2004. — Vol. 5, no. 1. — P. 71.

[6] Higashi Susan, Hungria Mariangela, Brunetto MADC. Bacteria classification based on 16S ribosomal gene using artificial neural networks // Proceedings of the 8th WSEAS International Conference on Computational intelligence, man-machine systems and cybernetics. — 2009. — P. 86–91.

[7] JetBrains Programming Languages and Tools Lab [Электронный ресурс]. — Access mode: https://research.jetbrains.org/groups/plt_lab (online; accessed: 05.05.2019).

[8] Knudsen Bjarne, Hein Jotun. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history //

Bioinformatics (Oxford, England). — 1999. — Vol. 15, no. 6. — P. 446–454.

[9] Pseudobase database [Электронный ресурс]. — Access mode: http://pseudobaseplusplus.utep.edu/ (online; accessed: 13.04.2020).

[10] Rivas Elena, Eddy Sean R. The language of RNA: a formal grammar that includes pseudoknots // Bioinformatics. — 2000. — Vol. 16, no. 4. — P. 334–340.

[11] RnaCentral database [Электронный ресурс]. — Access mode: https://rnacentral.org/ (online; accessed: 13.04.2020).

[12] Sherman Douglas. Humidor: Microbial Community Classification of the 16S Gene by Training CIGAR Strings with Convolutional Neural Networks. — 2017.

[13] Abadi Martín, Agarwal Ashish, Barham Paul et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. — 2015. — Software available from tensorflow.org. Access mode: http://tensorflow.org/.

[14] YaccConstructor [Электронный ресурс]. — Access mode: https://github.com/YaccConstructor (online; accessed: 05.05.2019).