

Исследование рекуррентной нейронной сети для векторного представления нуклеотидных последовательностей

Д. О. Комаровских, В. Л. Литвинов

Санкт-Петербургский государственный
электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)

Санкт-Петербург, Россия

komarovskih.d@yandex.ru, vlad.litvinov61@gmail.com

И. А. Киселев, А. М. Панюков

EPAM Systems

Санкт-Петербург, Россия

ikiselev7@gmail.com, chgk1101@gmail.com

Н. И. Трофимов

Санкт-Петербургский государственный экономический университет

Санкт-Петербург, Россия

studunec@gmail.com

Аннотация. В работе исследуются возможности машинного обучения, в частности рекуррентных нейронных сетей, для выявления эволюционных связей между организмами. Для того, чтобы выявлять эволюционные взаимосвязи между различными организмами, используется метод множественного выравнивания нуклеотидных последовательностей. Множественное выравнивание имеет высокую вычислительную сложность, поиск глобального оптимума относят к *NP*-полным задачам. Предлагается использовать такой подход, как векторное представление (эмбединги) слов, для дальнейшего выявления закономерностей с помощью LSTM-сети. Обучение сети проведено на датасете из 1.5 миллионов пар нуклеотидных последовательностей, принадлежащих к виду *Escherichia coli*.

Ключевые слова: метод множественного выравнивания последовательностей; LSTM рекуррентная нейронная сеть; обработка естественного языка; эволюционная дистанция

I. ВВЕДЕНИЕ

В обработке естественного языка широко используется подход векторного представления слов (word embeddings). Для того, чтобы компьютер мог работать с текстом, отдельные его части (слова, предложения или целый текст) кодируют по специальным правилам в вектора (эмбединги). С помощью эмбедингов можно решать задачи классификации текстов, определять являются ли слова похожими и т.д. В данной работе рассматривается применение такого подхода для выявления эволюционных связей между организмами, принадлежащими к одному виду, по их нуклеотидным последовательностям.

В биоинформатике для того, чтобы выявлять эволюционные связи между организмами используется метод множественного выравнивания последовательностей. Алгоритмы, реализующие данный метод,

имеют высокую вычислительную сложность, так как поиск глобального оптимума относится к *NP*-полным задачам [1].

Для того, чтобы получить вектор последовательности гена, в работе предлагается использовать рекуррентную LSTM нейронную сеть [2]. Из множества достижений LSTM сетей можно выделить наилучшие результаты в распознавании несегментированного слитного рукописного текста, а также в задаче распознавания фоном на классическом корпусе естественной речи. Пример фрагмента сети показан на рис. 1 [2].

Обучение сети проводится на датасете из 1.5 миллионов пар нуклеотидных последовательностей, принадлежащих к виду *Escherichia coli* (кишечная палочка) с известными для них эволюционными расстояниями.

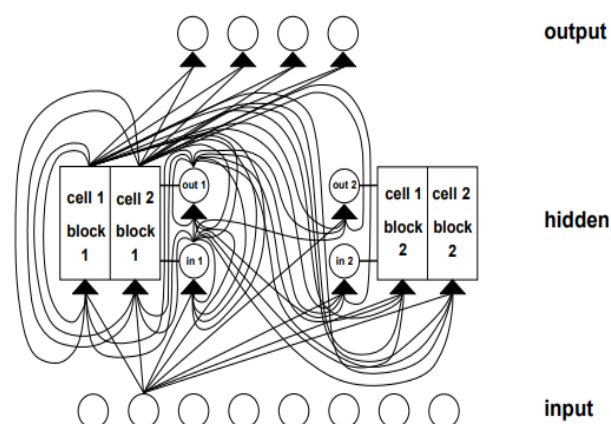


Рис. 1. Пример LSTM сети с 8 входами, 4 выходами и 2 ячейками размера 2

II. ПОДГОТОВКА ДАННЫХ

Молекула ДНК состоит из нуклеотидов, последовательно идущих друг за другом. Всего различают четыре нуклеотида в ДНК – аденин, гуанин, тимин, цитозин. Таким образом, ген представляет собой некоторую строку, в которой может быть не более 4 различных символов.

Последовательность нуклеотидов предварительно нужно разбить на токены, для того чтобы работать с ней дальше. Для этого используем два подхода – разбиение на k -меры длины шесть и *byte pair encoding* [3].

Нуклеотидные последовательности были сгруппированы по группам генов, к которым они принадлежат и разбиты на батчи по 30 последовательностей. По выравненным генам в каждом батче (для выравнивания использовался прогрессивный метод MAFFT [4]) были посчитаны эволюционные расстояния следующими способами:

- Метод Джукса–Кантора (JC69) [5].
- Метод Кимуры (K81) [6].
- Метод Тамуры–Нея (TN93) [7].

Датасет был разбит на *train* датасет (70% от общего числа пар последовательностей) и *test* датасет (30% от общего числа последовательностей).

III. ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТОВ

Построенная модель состояла из двух слоев. Первый – векторное представление, поданных на вход токенов последовательности размерности 300. Второй – рекуррентная нейронная сеть LSTM с размерностью выходного вектора равной 300. Мы рассматривали однослойную и двухслойную LSTM. На выход модели также накладывалась функция среднего, для того чтобы получить один общий вектор для всей последовательности токенов.

В обучении модели использовалась следующая функция потерь:

$$\text{loss}(x_1, x_2, y) = \left| 1 - \frac{x_1 \cdot x_2}{\max(\|x_1\|_2 \cdot \|x_2\|_2, \varepsilon)} - y \right|,$$

где x_1 и x_2 – два вектора, y – эволюционное расстояние, ε – малое положительное число (порядка $1e-8$), нужное для того, чтобы избежать деления на ноль.

На вход функции потерь передаются два посчитанные моделью вектора и эволюционное расстояние между парой последовательностей генов. Таким образом, модель учится располагать вектора в векторном пространстве так, чтобы косинусное расстояние между ними соответствовало эволюционному.

Для валидации обученной модели использовалась метрика *accuracy* – доля пар последовательностей, для

которых разница между косинусными и эволюционными расстояниями была меньше 0.01. Валидация проводилась на тестовом датасете.

В таблице 1 представлены результаты обучения однослойной LSTM на разных данных. Видно, что только для эволюционных расстояний, посчитанных методом Тамуры–Нея, модель показала результат чуть хуже. Также нет разницы, как разбивались последовательности генов на токены. В обоих случаях показатель *accuracy* получился равным 0.87.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ОДНОСЛОЙНОЙ LSTM

Методы разбиения	Методы эволюционных расстояний		
	JC69	K81	TN93
BPE	0.87	0.87	0.85
k-mer	0.87	-	-

Обучая модель с двухслойной LSTM на эволюционных расстояниях, посчитанных методом Джукса–Кантора, и разбиением на токены при помощи BPE, удалось добиться *accuracy*, равного 0.9. Для более глубокого анализа результатов обучения последней модели было построено два графика. На рис. 2. изображен график соответствия косинусных расстояний эволюционным. Зеленым цветом отмечены точки, где разница между косинусными и эволюционными расстояниями меньше 0.01. Также можно наблюдать некоторые выбросы, для установления причин которых необходимо дальнейшее исследование.

На рис. 3 представлен график распределения значений функции потерь. Наибольшую долю (31.5 %) пар векторов составили те, у которых разница между косинусными и эволюционными расстояниями меньше 0.001.

Для визуализации полученных результатов также можно использовать алгоритм *t-SNE* (стохастическое вложение соседей с *t*-распределением) [8], хорошо подходящий для вложения данных высокой размерности в пространство низкой размерности. На рис. 4. видно, что последовательности генов, принадлежащие к одной группе генов, располагаются ближе друг к другу.

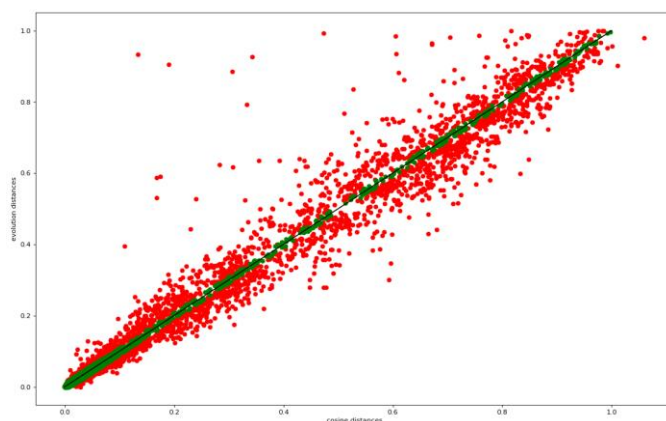


Рис. 2. График соответствия косинусных расстояний эволюционным

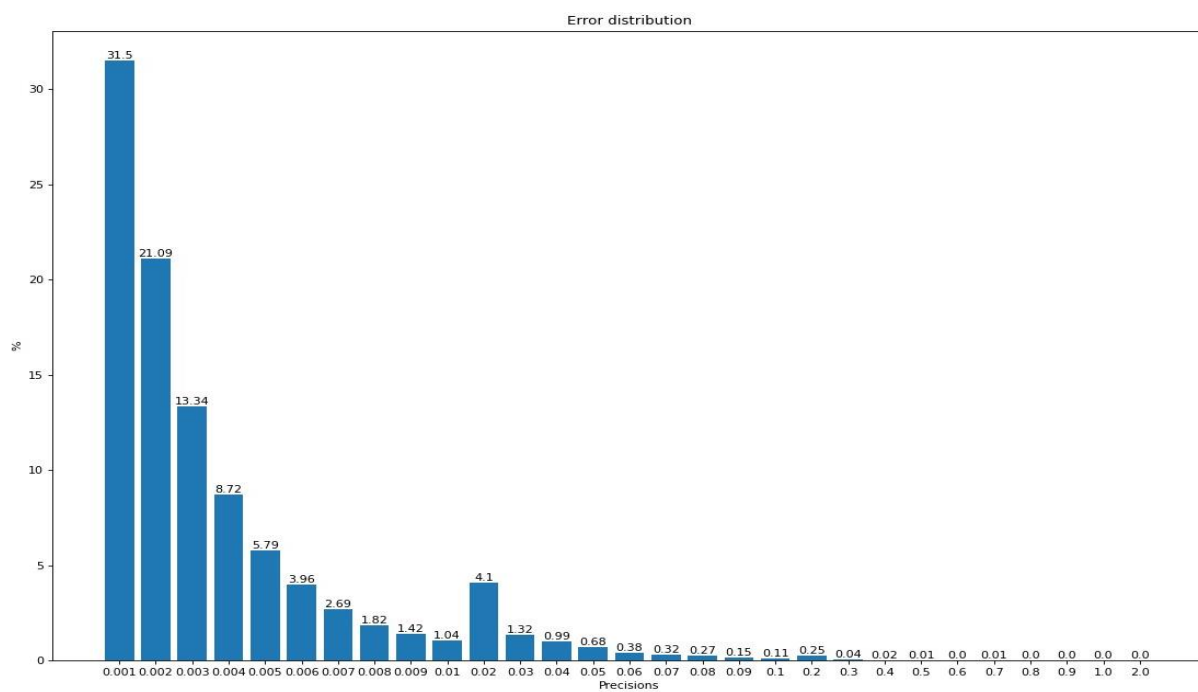


Рис. 3. График распределения значений функции потерь

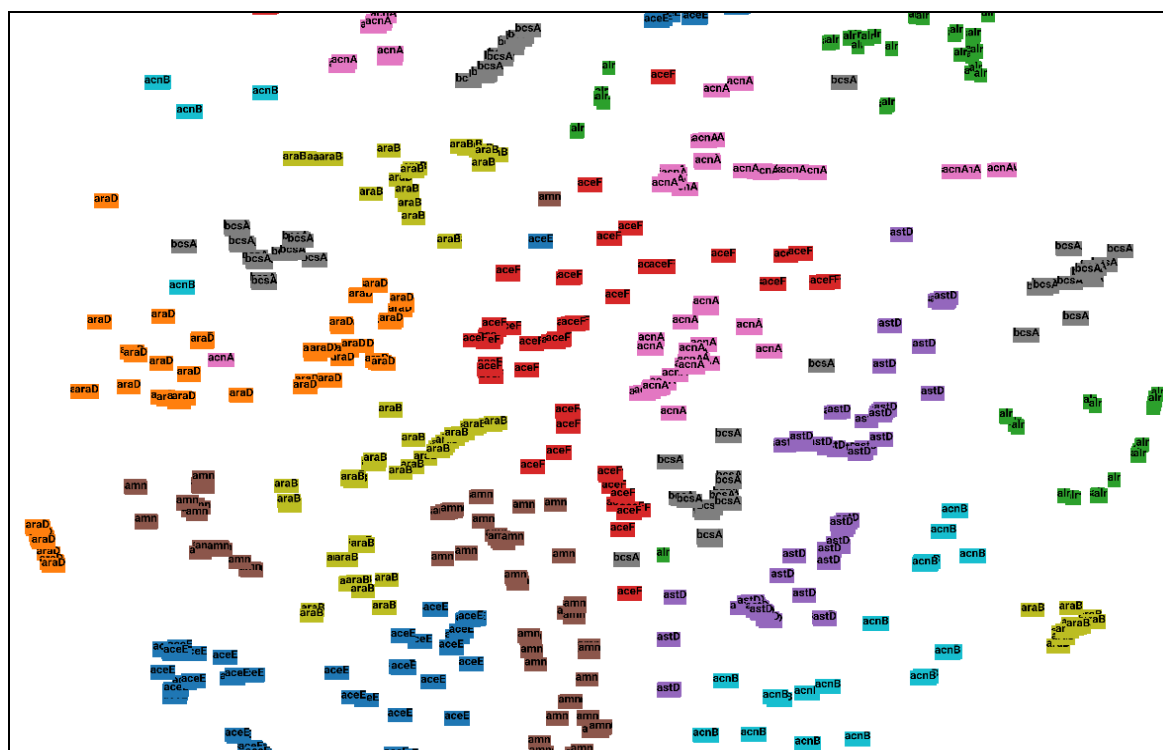


Рис. 4. Визуализация векторов последовательностей генов в 2D пространстве при помощи алгоритма *t-SNE*

IV. ЗАКЛЮЧЕНИЕ

Таким образом, в работе рассмотрено применение рекуррентной LSTM нейронной сети для векторного представления нуклеотидных последовательностей. Модель, обученная на последовательностях генов, принадлежащих к виду *Escherichia coli*, показала хороший результат на последовательностях из того же вида. Полученные вектора можно в дальнейшем использовать для решения различных задач, например, для классификации последовательностей по именам генов, определения эволюционных расстояний, построения филогенетических деревьев.

СПИСОК ЛИТЕРАТУРЫ

- [1] Jurate Daugelaite, Aisling O' Driscoll, Roy D. Sleator. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. 2013.
- [2] Sepp Hochreiter, Jurgen Schmidhuber. Long short-term memory // Neural Computation. 1997, №9(8). P. 1735-1780.
- [3] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, №1. P. 1715–1725.
- [4] Kazutaka Katoh, Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability // Molecular Biology and Evolution. 2013, №30(4). P. 772-780.
- [5] Thomas H. Jukes, Charles R. Cantor. Evolution of Protein Molecules // Mammalian Protein Metabolism. 1969, №3. P. 21-132.
- [6] Motoo Kimura. Estimation of evolutionary distances between homologous nucleotide sequences // Proceedings of the National Academy of Sciences of the United States of America. 1981, №78(1). P. 454-458.
- [7] Koichiro Tamura, Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees // Molecular Biology and Evolution. 1993, №10(3). P. 512–526.
- [8] Laurens van der Maaten, Geoffrey Hinton. Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008, №9. P. 2579-2605.