БИОФИЗИКА, 2020, том 65, № 4, с. 676–679

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ С ПАМЯТЬЮ ДЛЯ ПРЕДСКАЗАНИЯ ИНТРОН-ЭКЗОННОЙ СТРУКТУРЫ ГЕНА

© 2020 г. Л.А. Урошлев, Н.В. Баль, Е.А. Чеснокова

Институт высшей нервной деятельности и нейрофизиологии РАН, 117485, Москва, ул. Бутлерова, 5а

E-mail: leoniduroshlev@gmail.com

Поступила в редакцию 21.04.2020 г.
После доработки 21.04.2020 г.
Принята к публикации 29.04.2020 г.

Построены несколько типов нейросетей с памятью. Каждая из них была обучена на полном геноме мыши для предсказания интрон-экзонной структуры гена. Было проведено сравнение нейросетей в работе как на тестовой выборке, так и на экспериментальном материале, полученном после секвенирования культуры мозга крысы, обработанного реагентами, ингибирующими сплайсинг.

Ключевые слова: рекуррентные нейросети, сплайсинг, LSTM-сеть, GRU-сеть, пладиенолид.

DOI: 10.31857/S0006302920040079

Сплайсинг – один из ключевых механизмов в обеспечении белкового разнообразия у эукариотических организмов. Сплайсинг также регулирует стабильность различных вариантов мРНК. Сравнение экспрессии транскриптов в разных тканях человека показало, что мозг, печень и семенники имеют самые высокие уровни альтернативного сплайсинга, при этом в разных тканях могут преобладать разные типы альтернативного сплайсинга [1].

В нервной системе важным фактором, зависящим от альтернативного сплайсинга, является пространственная локализация транскриптов [2]. Кроме того, обнаружено, что активность нейронов может оказывать влияние на вырезание интронов [3], что, в свою очередь, влияет на экспрессию целевых белков и является механизмом тонкой настройки работы нейронов в различных условиях их функционирования [4].

Альтернативный сплайсинг обеспечивает разнообразие транскриптов не только внутри клеток, но и между ними. При формировании нервной системы происходит дифференцировка клеток-предшественников в различные типы нейронов и глии. С помощью выделения рибосомо-ассоциированной РНК из разных типов нейронов было обнаружено достоверное различие наборов альтернативных сплайс-изоформ РНК в разных типах клеток, что может свидетельствовать о том, что альтернативный сплайсинг вовлечен в процесс дифференциации нейронов в ходе развития нервной системы [5].

В настоящее время существует множество алгоритмов для поиска и аннотации сплайс-изоформ *in silico*. В качестве основы алгоритма могут быть использованы графы де Брюина, жадные алгоритмы, графы перекрытий и другие. Однако почти все алгоритмы предназначены для обработки данных, полученных с помощью различных вариаций секвенатора Illumina. Если же для получения данных используется иной секвенатор, алгоритмы могут ошибаться, так как большая часть современных алгоритмов для определения сплайс-изоформ работает с множеством небольших отдельных чтений (100 п.н.), выровненных на референсный геном. Последовательности, получаемые с помощью таких приборов как IonTorrent или Oxford Nanopore, представляют собой небольшой набор достаточно длинных (свыше 500 пар нуклеотидов) контигов, которые должен последовательно обрабатывать алгоритм. Еще одной возможной областью применения алгоритма, работающего с длинными последовательностями нуклеотидов, является поиск и аннотация интронов в NGS-данных различного происхождения, уже собранных в транскрипты, что может быть полезно при аннотации сборок.

Чтобы выстроить достаточно эффективный алгоритм выделения сплайс-изоформ, необходимо воспользоваться методами, которые способны эффективно выделять паттерны из последовательности. К таковым можно отнести методы машинного обучения, которые успешно используются в самых разных областях вычислительной молекулярной биологии. В частности, примерами использования этих методов могут служить классификация событий сплайсинга [6] и влияния полиморфизмов на патогенность [7].

МЕТОДЫ

Для поиска интрон-экзонной структуры генов использовали несколько моделей машинного обучения с памятью, подходящих для обработки одномерных последовательностей. Модель должна запоминать самые устойчивые паттерны и забывать незначимые. Такие задачи очень распространены в различных областях вычислительной лингвистики и обработки изображений. Для этих целей используются различные виды рекуррентных нейронных сетей. Дополнительным аргументом в пользу выбора этих моделей может служить то, что, при правильной стратегии обучения, модель может быть устойчива к ошибкам секвенирования.

В качестве первого слоя нашей нейронной сети (см. рис. 1) мы использовали одномерный сверточный слой нейронов с размером окна в 2 п.н. На втором слое нашей модели мы используем два вида рекуррентных сетей – однонаправленная GRU-сеть [9] и LSTM-сеть [8], двунаправленная GRU и LSTM-сеть. Третий слой используется для предотвращения переобучения модели с помощью дропаута. Все реализации нейросетей строили с помощью пакета keras для языка программирования Python версии 3.8.

Тестирование нейросетей проходило в два этапа – на первом этапе нейросеть обучалась с помощью искусственных данных. Данные об интрон-экзонной структуре генов брали из аннотации. При помощи аннотации генома мыши были выделены полные последовательности генов, для которых были сформированы характеристические векторы по следующему правилу: 0 соответствует нуклеотидам, попавшим в интроны, 1 – в экзоны. Далее выборка делилась на обучающую и тестовую в соотношении 80% (обучающая выборка) и 20% (тестовая). Вычислительные эксперименты по обучению сетей проводили для разных объемов окон последовательности, а именно 600, 700, 800, 1000 п.н. Для поиска коэффициентов моделей использовался метод оптимизации ADAM [10]. Результаты для наших моделей показаны в таблице.

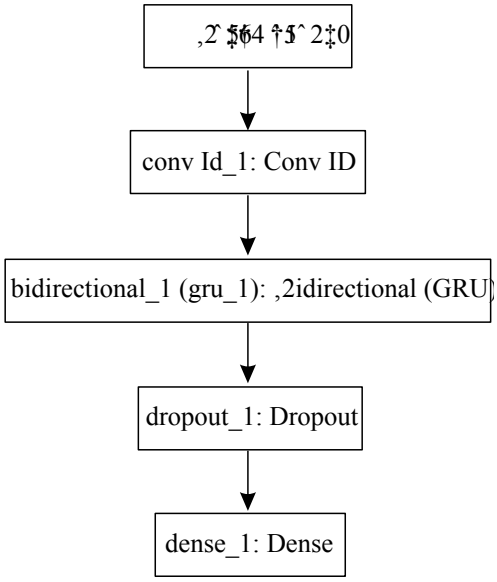

Рис. 1. Схема нейросети, используемой для предсказания интрон-экзонной структуры гена.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В определенных условиях вырезание интронов может быть подавлено, поэтому сети, обученные отличать участки интронов от участков экзонов, могут использоваться для опознавания ситуации, в которой происходит такое подавление. Это дает возможный путь для экспериментальной проверки разработанного алгоритма.

На втором этапе уже предобученную сеть тестировали на наборе экспериментальных данных, полученных с помощью секвенатора IonTorrent. Для этого были отсеквенированы транскриптомы образцов нейроглиальной культуры гиппокампа крысы [11], два из которых были обработаны пладиенолидом – реагентом, который ингибирует процесс сплайсинга, два же были оставлены в качестве контроля. Экспериментальный и контрольный образцы были собраны с помощью сборщика SPAdes [12] в контиги. N50 для получившейся сборки был равен 1022 п.н. для кон-

Таблица 1. Средняя величина корректно предсказанной принадлежности нуклеотидов для разных типов нейронных сетей

Тип сети	600	700	800	1000
Однонаправленная GRU	65%	74%	71%	80%
Двунаправленная GRU	82%	84%	91%	92%
Однонаправленная LSTM	72%	72%	80%	82%
Двунаправленная LSTM	93%	71%	67%	81%

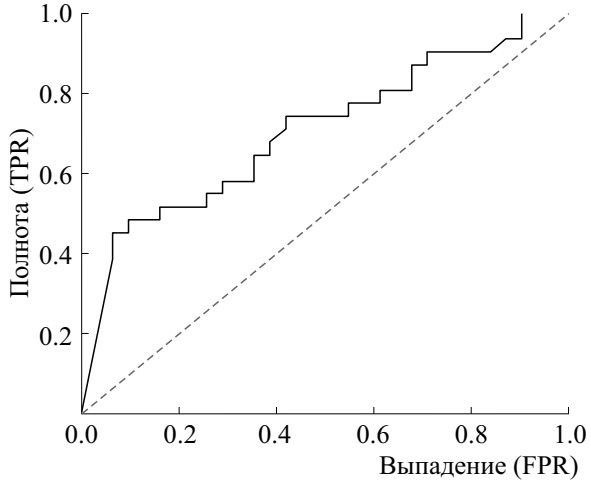

Рис. 2. ROC-кривая нейросетевого классификатора «эксперимент–контроль».

трольных образцов и 913 п.н. для экспериментальных образцов.

Вычислительный эксперимент заключался в следующем: разделить контрольные и экспериментальные (с «застрявшими» интронами) контиги. С этой целью для каждой группы были выбраны 50 случайных контигов, для которых была произведена разметка интрон-экзонной структуры с помощью разработанной программы. Предполагалось, что те последовательности, в которых длина интронов больше пороговой, – экспериментальные. В противном случае контиг определялся как принадлежащий к контрольной группе. Далее на основе количества обнаруженных интронов контиги классифицировались на контрольные и экспериментальные. Для иллюстрации классификации построена ROC-кривая (рис. 2) для нейросети, которая на этапе тестирования показала наилучший результат. Как видно, полученные нейросети обеспечивают достаточно неплохой уровень классификации, даже при использовании обучения на организмах, с близкой, хотя и иной видовой принадлежностью.

Работа большей части нейросетей, в том числе свойства обучающей выборки и архитектура каждой отдельно взятой сети по-прежнему является «черным ящиком» для исследователей, и параметры, обеспечивающие эффективную работу сети на тех или иных данных, подбираются эмпирически. В дальнейших исследованиях планируется подробно разобрать вопросы эффективности той или иной архитектуры, особенно связанные с методами выбора и обучения тех или иных нейронных сетей. Например, в ходе подготовки данной работы были также рассмотрены архитектуры Seq2Seq и сети, основанные на механизмах внимания. Несмотря на более сложное внутреннее устройство, эти сети не показали каких-либо значимых результатов в предсказании интрон-экзонной структуры.

Отдельным вопросом для исследования являются границы применимости нейросетей, обученных на одних видах, для предсказания интронов в геномах других видов, эволюционно достаточно далеких. Это чрезвычайно важно для аннотации геномов новых модельных организмов, таких как, например, виноградная улитка. С одной стороны, ее геном и транскриптом чрезвычайно важны для задач по исследованию памяти [13]. С другой стороны, из-за обилия повторов и отсутствия близкородственных видов ее аннотация имеющимися алгоритмами крайне затруднена.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Российского научного фонда, грант № 19-74-00141.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания каких-либо исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. G. Yeo, D. Holste, and G. Kreiman, Genome Biol. **5** (10), R74 (2004).
2. E. Furlanis and P. Scheiffele, Annu. Rev. Cell Dev. Biol. **34**, 451 (2018).
3. O. Mauger, F. Lemoine, and P. Scheiffele, Neuron **92** (6), 1266 (2016).
4. G. Biamonti, A. Amato, E. Belloni, et al., Aging Clin. Exp. Res. (2019). DOI: 10.1007/s40520-019-01360-x
5. E. Furlanis, L. Traunmüller, G. Fucile, and P. Scheiffele, Nat. Neurosci. **22** (10), 1709 (2019).
6. Louadi Z. et al. Genes **10** (8), 587 (2019).
7. J. Cheng, T. Y. D. Nguyen, K. J. Cygan, et al., Genome Biol. **20** (1), 48 (2019).
8. F. A. Gers, J. Schmidhuber, and F. Cummins, in *Neural Nets WIRN Vietri-99* (Springer, Lond., 1999), pp. 133–138.
9. J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, arXiv, 1412.3555 (2014).
10. D. P. Kingma and J. Ba, arXiv preprint: arXiv, 1412.6980 (2014).
11. O. Mauger, F. Lemoine, and P. Scheiffele, Neuron **92** (6), 1266 (2016).

12. A. Bankevich, S. Nurk, D. Antipov, et al., J. Comput. Biol. **19** (5), 455 (2012).

13. N. Aseyev, A. K. Vinarskaya, M. Roshchin, et al., Front. Cell. Neurosci. **11**, 348 (2017).

Prediction of the Exon-Intron Structure of a Gene Based on Long Short-Term Memory Neural Network

L.A. Uroshlev, N.V. Bal, and E.A. Chesnokova

Institute of Higher Nervous Activity and Neurophysiology, Russian Academy of Sciences, ul. Butlerova 5a, Moscow, 117485 Russia

This paper suggests several models of long short-term memory neural networks. We trained every model on a full mouse genome to predict the exon-intron structure of a gene. In this work we compare the performance of the neural networks in the test sample and experimental material obtained after screening rat brain cells treated with splicing inhibitors.

Keywords: recurrent neural networks, splicing, LSTM-neural network, GRU-neural network, pladienolide