обнаруженных в молекулярных библиотеках веб-сервера «Pharmit», способны имитировать фармакофорные свойства Fab-фрагмента МКА N6 путем специфических и эффективных взаимодействий с участком белка gp120 ВИЧ-1, критическим для связывания вируса с клеточным рецептором CD4. При этом ключевую роль играют многочисленные ван-дер-ваальсовые контакты лигандов с остатками Phe^{43}–полости gp120, ответственными за взаимодействие ВИЧ-1 с $Phe\text{-}43_{CD4}$, а также водородные связи с остатком $Asp\text{-}368_{gp120}$, образование которых увеличивает аффинность связывания без активации нежелательного аллостерического эффекта [5].

Идентифицированные соединения могут быть использованы в работах по созданию новых противовирусных препаратов – ингибиторов проникновения ВИЧ-1, блокирующих ранние стадии развития ВИЧ-инфекции.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Huang J., Kang B. H., Ishida E. et. al. Identification of a CD4-Binding-Site Antibody to HIV That Evolved Near-Pan Neutralization Breadth // Immunity. 2016. Mol. 45, № 5. P. 1108–1121.
2. Sunseri J., Koes D. R. Identification of a CD4-Binding-Site Antibody to HIV That Evolved Near-Pan Neutralization Breadth // Nucleic Acids Research. 2016. Vol. 44. P. 442–448.
3. Alhossary A., Handoko S. D., Mu Y. et al. Fast, Accurate, and Reliable Molecular Docking With QuickVina 2 // Bioinformatics. 2015. Vol. 31, № 13. P. 2214–2216.
4. Case D. A., Daren T. A., Cheatham T. E. et al. AMBER 11. User Manual. 2010. Univ. California, San Fr.
5. Courter J. R., Madani N., Sodroski J. Structure-based Design, Synthesis and Validation of CD4-mimetic Small Molecule Inhibitors of HIV-1 Entry: Conversion of a Viral Entry Agonist to an Antagonist // Accounts of Chemical Research. 2014. Vol. 47, №. 4. P. 1228–1237.

КЛАССИФИКАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РНК С ПОМОЩЬЮ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

И. В. Климук, В. В. Гринев, В. В. Скакун, Н. Н. Яцков

Кафедра системного анализа и компьютерного моделирования, кафедра генетики, Белорусский государственный университет, Минск, Беларусь

E-mail: ivanklimuk96@gmail.com, {grinev_vv, skakun, yatskou}@bsu.by

Предложена модель для классификации молекул РНК человека. Модель построена на основе многоканальной одномерной сверточной сети и применяет метод бинарного кодирования для геномных последовательностей. Разработанный подход

работает более точно, чем аналогичные методы, а также способен обрабатывать более длинные последовательности.

Ключевые слова: молекулы РНК; биоинформатика; классификация; алгоритм; сверточная нейронная сеть; машинное обучение.

Введение. Транскрипция генов приводит к образованию матричных РНК, а также разнообразных малых и длинных некодирующих РНК [1]. Длинные некодирующие РНК тесно связаны со многими биологическими процессами, например, с многоуровневой регуляцией экспрессии генов. Их структура во многом схожа со структурой матричных РНК, что значительно усложняет задачу точного определения вида молекул. Для решения данной задачи требуется разработка модели классификации кодирующих и некодирующих молекул РНК. В работе [2] представлен алгоритм на основе k-меров, точность которого варьируется от 87% до 99% для молекул РНК разных живых организмов, однако алгоритм является вычислительно затратным и ограничен максимальной длиной анализируемых молекулы РНК. В работе [3] представлены модели на основе векторизации и алгоритма случайного леса. Их преимущество в простоте, интерпретируемости параметров и малых вычислительных затратах. Однако точность при работе с молекулами РНК человека невысокая и варьирует от 90 % до 96 %.

Целью настоящей работы является разработка модели классификации молекул РНК человека на кодирующие и некодирующие, устраняющей недостатки выше упомянутых моделей [2, 3]. Разработана модель на основе одномерной сверточной нейронной сети и подхода бинарного кодирования геномных последовательностей. Проверка работоспособности алгоритмов выполнена на наборе РНК из базы данных NCBI RefSeq. Для увеличения обучающей выборки реализована аугментация данных из класса некодирующих молекул.

Обработка и аугментация данных. Для решения поставленной задачи в качестве обучающей выборки взяты данные из базы NCBI RefSeq. В выборке присутствует 128161 истинная открытая рамка считывания (ОРС) как признак матричных РНК и 4235 некодирующих молекул. После первичной обработки данных и удаления выбросов, размер класса кодирующих молекул сократился до 113071. Так как нейронные сети чаще склонны к переобучению и требуют большого количества данных, для увеличения обучающей выборки и выравнивания соотношения классов в ней проведена аугментация некодирующих молекул РНК. Новые образцы были смоделированы на основе имеющихся ложных ОРС с помощью “отзеркаливания” (т. е. записи в обратном порядке) и дальнейшего выделения случайных подпоследовательностей с использованием

равномерного и нормального распределений [4]. Таким образом, класс некодирующих последовательностей был расширен до 93170, а вся обучающая выборка составила более 200 тысяч примеров. Аугментация проводилась таким образом, чтобы распределение длин последовательностей в обоих классах сохранялось примерно одинаковым.

Для обработки одномерной сверточной нейронной сетью последовательности были преобразованы с помощью бинарного кодирования единицей. Принцип заключается в замене каждого символа алфавита на вектор из 0 и 1, где 1 стоит в соответствии символу на данной позиции (рис. 1). В таком виде каждый символ последовательности представляет собой канал, а вся последовательность – анализируемое «изображение» (по аналогии с двумерными сверточными сетями [5]).

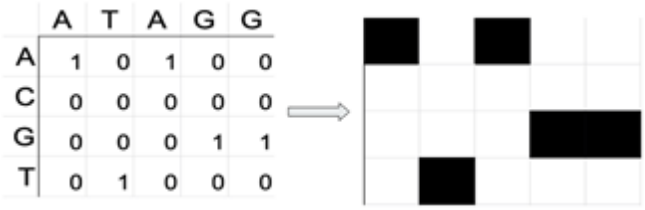

Рис. 1. Пример бинарного кодирования последовательности ATAGG

Архитектура нейронной сети. Ключевой операцией в разработанной модели является одномерная многоканальная свертка [5, 6]. Гиперпараметры, такие как количество слоев и размер сверточных ядер в них, были определены эмпирически, постепенно увеличивая число параметров при выполнении критерия об отсутствии переобучения. Максимальная длина анализируемой последовательности – 6890 нуклеотидов. Архитектура нейронной сети изображена на рис. 2.

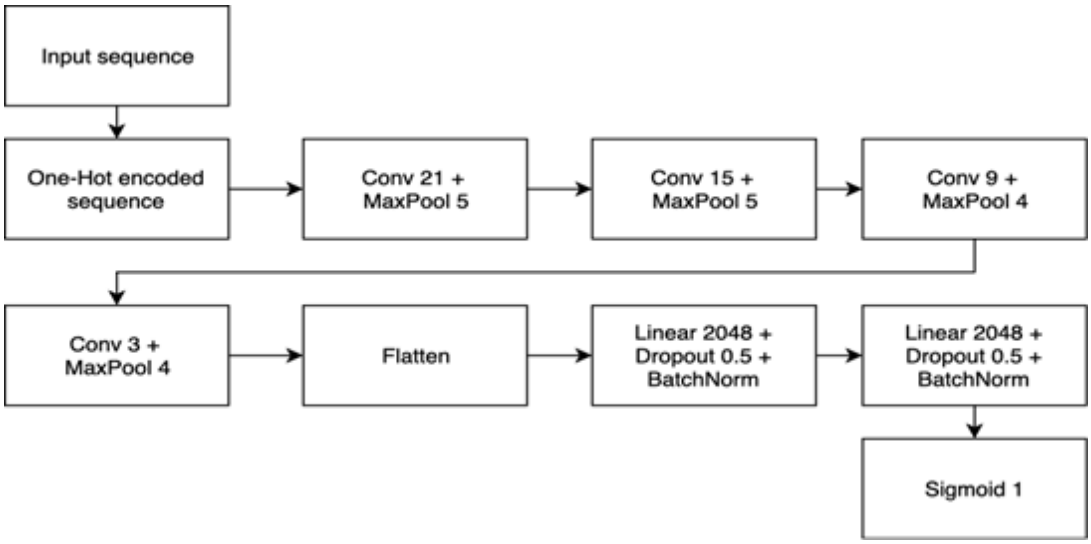

Рис. 2. Архитектура разработанной модели. *Conv* – операция свертки, *MaxPool* – операция макс-пулинга, *Flatten* – операция выпрямления сверточных фильтров, *Linear* – полносвязный слой, *Sigmoid* – функция сигмоиды

Вычислительный эксперимент и результаты. В качестве меры качества классификационной модели рассмотрена точность классификации:

$$accuracy = (TP + TN) / (TP + TN + FP + FN),$$

где *TP* и *TN* – количество верно классифицированных объектов первого и второго класса, соответственно, *FP* и *FN* – количество ошибочно классифицированных объектов первого и второго класса, соответственно.

Для контроля эффекта переобучения и дополнительной оценки точности классификации рассмотрены площадь под *ROC-кривой*, прецизионность (*precision*), полнота (*recall)* и критерий *f1* [7]. Нейронная сеть обучалась в среднем 5-7 эпох оптимизатором *Adam* [8] с начальным шагом обучения 0,01, что занимает порядка 8-10 часов на 24 ядерной машине без видеокарт. Результаты работы модели с различными значениями максимальной длины анализируемой последовательности отображены в таблице. Чем выше длина максимальной анализируемой последовательности *Lmax*, тем выше точность классификации.

Таблица.

Качество работы классификатора от длины максимальной анализируемой последовательности.

Lmax	accuracy	ROC AUC	precision	recall	f1
1880	0,989	0,990	0,987	0,990	0,987
4690	0,995	0,999	0,993	0,996	0,994
6890	0,996	0,999	0,996	0,996	0,996

Выводы. Разработана эффективная модель классификации кодирующих и некодирующих молекул РНК, устраняющая недостатки опубликованных моделей [2, 3]. В работе предложен и проверен метод аугментации данных для увеличения обучающей выборки некодирующих последовательностей. Основным недостатком модели нейронной сети является вычислительная сложность, требуемая на этапе обучения. Разработанная модель нейронной сети составляет основу серверного приложения для предсказания кодирующих молекул РНК с целью определения открытых рамок считывания.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Djebali S., Davis C. A., Merkel A. et al. Landscape Of Transcription In Human Cells // Nature. 2012. Vol. 489. P. 101–108.
2. Wen J., Liu Y., Shi Y. et al. Classification Model For Lncrna And Mrna Based On K-Mers And A Convolutional Neural Network // BMC Bioinformatics. 2019. Vol. 20, №1:469. P. 1–14.

3. Закирова В. Р., Сырокваш Д. А., Гилевский С. В. и др. Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей // Информатика. 2019. Т.12, №2. С. 109–118.
4. Chollet F. Deep learning with Python. New York : Manning Publications Co. 2017. 384 p.
5. Krizhevsky A. et al. Imagenet Classification With Deep Convolutional Neural Networks // NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems. 2012. Vol. 1. P. 1097–1105.
6. Szegedy C., Liu W., Jia Y., Sermanet P. et al. Going deeper with convolutions // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 1–9.
7. Powers D. Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness & Correlation // Journal of Machine Learning Technologies. 2011. Vol. 2. N1. P. 37–63.
8. Kingma D. P., Jimmy L. B. Adam: A Method For Stochastic Optimization // International Conference on Learning Representations (ICLR), 2013. P. 1–15.

ОСОБЕННОСТИ СБОРКИ ГЕНОМА BUCHNERA APHIDICOLA ИЗ ДАННЫХ МЕТАГЕНОМНОГО СЕКВЕНИРОВАНИЯ

Я. П. Кононович, Р. С. Шулинский

Белорусский государственный университет,
биологический факультет, Минск, Беларусь
E-mail: yana.kananovich@gmail.com

Из данных метагеномного секвенировани, полученных из тлей видов Macrosiphum albifrons, Aphis craccivora и Myzus persicae, были собраны три генома Buchnera aphidicola, Приведена методика сборки и оценка полученных нуклеотидных последовательностей.

Ключевые слова: Buchnera aphidicola; метагеномика; сборка генома de novo.

Введение. Расшифровка генома бактерии *Buchnera aphidicola* – облигатного эндосимбионта тли – является важным шагом в понимании молекулярных механизмов адаптации этих насекомых к питанию на различных растениях. Ранее нами была проведена работа по выяснению видового состава микробиома тлей *Aphis pomi*. Тогда изучение вариабельных участков V4–V5 16S рРНК микробиома показало наличие у этого вида тли как минимум 4 видов бактерий-симбионтов [1], что можно ожидать и при исследовании микробиома других видов тлей, коллектированных для настоящего исследования. Это обстоятельство усложняет процесс сборки генома отдельных организмов из данных полногеномного секвенирования. Тем не менее, современные технологии и программное обеспечение позволяют собирать отдельные геномы из данных секвенирования, где в общий пул прочтений входят геномы сразу несколь-