DOI: 10.25205/978-5-4437-1843-9-26

КЛАССИЧЕСКИЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ И НЕЙРОСЕТИ В ЗАДАЧЕ ОПРЕДЕЛЕНИЯ ХОЗЯЕВ РНК-СОДЕРЖАЩИХ ВИРУСОВ НАСЕКОМЫХ, МЛЕКОПИТАЮЩИХ И РАСТЕНИЙ

TRADITIONAL MACHINE LEARNING APPROACHES AND NEURAL NETWORKS IN THE TASK OF HOST IDENTIFICATION FOR RNA VIRUSES INFECTING *INSECTA*, *MAMMALIA*, AND *VIRIDIPLANTAE*

Ф. С. Перелыгин[1,2], А. Н. Лукашев[1,3], Ю. А. Алешина[1,2]

[1] *Институт медицинской паразитологии, тропических и трансмиссивных заболеваний им. Е. И. Марциновского Сеченовского университета, Москва*
[2] *Московский государственный университет им. М. В. Ломоносова*
[3] *НИИ системной биологии и медицины, Москва*

F. S. Perelygin[1,2], A. N. Lukashev[1,3], Y. A. Aleshina[1,2]

[1] *Martsinovsky Institute of Medical Parasitology, Tropical and Vector-Borne Diseases, Sechenov University, Moscow*
[2] *Lomonosov Moscow State University*
[3] *Research Institute for Systems Biology and Medicine, Moscow*

✉ perelygin_f_s@staff.sechenov.ru

Аннотация

В данной работе исследуется проблема определения хозяев вирусов по особенностям состава геномов. Оценивается влияние таксономического состава обучающей выборки, набора признаков, алгоритма и длины нуклеотидной последовательности на качество классификации.

Abstract

The virus-host prediction based on virus genome composition is investigated. The effect of the taxonomic composition of the training sample, the set of features, the algorithm, and the length of the nucleotide sequence on classification quality is evaluated.

Изучение метавирома у различных видов-хозяев привело к открытию множества новых вирусов. Однако поскольку метавиром (особенно фекальный) состоит из генетического материала вирусов разного происхождения (собственные вирусы организма, вирусы из его пищи и т. д.), необходимы методы для точного определения хозяев вирусов по их геномным последовательностям.

Классические алгоритмы машинного обучения (Machine Learning, ML) были применены в задаче предсказания хозяев РНК-содержащих вирусов, инфицирующих три группы организмов: млекопитающих, насекомых и растения. В качестве признаков использовались частоты встречаемости нуклеотидных *k*-меров (последовательностей длиной от 1 до 7).

Как референсные методы на основе выравнивания (tBLASTx) с геномами близкородственных вирусов, так и ML-алгоритмы, обученные на геномах близкородственных вирусов, продемонстрировали высокую точность классификации (минимальная взвешенная F1-мера = 0,88). Однако при идентификации хозяев для вирусов новых родов (не представленных в обучающей выборке) эффективность ML-методов снизилась (наилучшая взвешенная F1-мера = 0,79), но все же оставалась выше, чем у методов, основанных на гомологии (F1-мера = 0,68).

В ходе исследования было показано влияние длины вирусных нуклеотидных последовательностей, используемых для обучения, на качество классификации. В задаче определения хозяев вирусов по коротким фрагментам длиной 400 и 800 нуклеотидов (нт) сравнивались классификаторы, обученные на полных геномах и на коротких фрагментах соответствующей длины. Сравнительный анализ показал, что модели, обученные и протестированные на фрагментах одной длины, демонстрировали более высокое качество классификации по сравнению с полногеномными моделями: прирост медианной взвешенной F1-меры составил 0,04, при этом была улучшена воспроизводимость результатов в 10 повторностях. Разработанный метод позволил достичь F1-меры 0,63 и 0,67 для фрагментов 400 и 800 нуклеотидов соответственно, что на 0,20 превышает показатели традиционного метода tBLASTx.

Применение нейросети из трех сверточных слоев или глубокой сверточной нейросети LegNet [1] не позволило значительно повысить качество классификации в сравнении с референсным классификатором на основе классического ML. ДНК языковая модель на основе блока BERT инструмента BERTax [2] показала лучшее качество классификации (взвешенная F1-мера = 0,77) в сравнении с референсным методом (реализация градиентного бустинга XGBoost, обученная на частотах встречаемости *k*-меров при $k = 4$; взвешенная F1-мера = 0,71). Нейронная сеть, включающая блок BERT модели ViraLM [3], провалилась в задаче предсказания хозяев неизвестных вирусов. При оценке кластеризации фрагментов геномов вирусов и мРНК их хозяев на скрытых слоях блока BERT нейросети ViraLM было показано, что нейросеть хорошо разделяет фрагменты хозяев вирусов от вирусных фрагментов, но практически не выделяет фрагменты вирусов в отдельные группы по хозяевам.

Для того чтобы выяснить, насколько в разных группах хозяев выражены особенности генома, имитируемые РНК-вирусами (например, чтобы избежать распознавания иммунной системой), классификаторы, обученные на вирусных геномах, были протестированы на фрагментах мРНК хозяев вирусов. Модель XGBoost показала хорошее качество классификации при выявлении фрагментов мРНК насекомых (точность = 0,74, F1-мера = 0,65) и млекопитающих (точность = 0,73, F1-мера = 0,65), однако провалилась при определении фрагментов мРНК растений (точность = 0,31, F1-мера = 0,42). Таким образом, можно предположить, что геномы насекомых и млекопитающих содержат более выраженные признаки, к которым приспосабливаются вирусы, в сравнении с растениями.

Литература

1. Penzar D. et al. LegNet: a best-in-class deep learning model for short DNA regulatory regions // Bioinformatics. 2023. Vol. 39, No. 8.

2. Mock F. et al. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks // Proc. Natl. Acad. Sci. 2022. Vol. 119, No. 35.

3. Peng C. et al. ViraLM: empowering virus discovery through the genome foundation model // Bioinformatics. 2024. Vol. 40, No. 12.