

Предсказание хозяев неизвестных РНК-вирусов с использованием классических алгоритмов машинного обучения и нейросетей

Научный руководитель – Алешина Юлия Александровна

Перелыгин Фёдор Сергеевич

E-mail: fedor959@gmail.com

Метавирусные исследования у различных видов-хозяев привели к открытию множества неизвестных вирусов. Так как метавирус содержит генетический материал вирусов из разных источников (собственно вирусы хозяина, вирусы из его пищи и т.д.), необходимы подходы для определения хозяев вирусов по геномной последовательности.

Классические алгоритмы машинного обучения (Machine Learning, ML) были обучены для предсказания хозяев РНК-содержащих вирусов, заражающих 3 группы хозяев – млекопитающих, насекомых и растения. В качестве признаков использовались частоты встречаемости нуклеотидных k-меров (подпоследовательностей длины k) длиной от 1 до 7. Как референсные методы на основе выравнивания с известными геномами вирусов (tBLASTx), так и алгоритмы ML, обученные на выборках геномов близкородственных вирусов, показали высокое качество классификации (минимальная взвешенная F1-мера = 0.88). В случае определения хозяев новых родов вирусов (не использованных при обучении) качество классификации ML снизилось (лучшая взвешенная F1-мера = 0.79), однако было выше, чем у методов, основанных на выравнивании (взвешенная F1-мера = 0.68).

Длина последовательности вирусного генома, используемая для обучения классификатора, влияла на качество предсказания. 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. Оптимальными оказались классификаторы, обученные на фрагментах, длина которых соответствует длине фрагментов, используемых при тестировании: по сравнению с полногеномными моделями качество предсказания повысилось на 0.04, а также улучшилась воспроизводимость в 10 репликах. Предложенный метод позволил определить хозяев неизвестных родов вирусов с взвешенной F1-мерой, равной 0.63 (для фрагментов длины 400 нт) и 0.67 (для фрагментов длины 800 нт), при этом качество классификации оказалось на 0.20 выше, чем у tBLASTx.

Применение свёрточных нейронных сетей в задаче предсказания хозяев новых родов вирусов по фрагментам геномов длины 1500 нт не позволило улучшить качество классификации по сравнению с референсным классификатором на основе классического ML (реализация градиентного бустинга XGBoost). При этом в этой же задаче языковая модель, включающая блок BERT опубликованного инструмента BERTax [1], показала лучшее качество классификации (взвешенная F1-мера = 0.77) по сравнению с XGBoost (взвешенная F1-мера = 0.71).

Работа поддержана грантом РНФ №22-15-00230.

Источники и литература

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