

Предсказание класса длинных некодирующих РНК по расположению в геноме методами глубокого машинного обучения

В.С. Соколов^{1*}, А.Ю. Пронозин², Д.А. Афонников^{1,2}

¹ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

² Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

* v.sokolov2@g.nsu.ru

Цель: Длинные некодирующие РНК (днРНК) – это класс РНК длиной более 200 нуклеотидов, не кодирующих белки. Экспериментальные исследования показывают вовлеченность днРНК в регуляцию экспрессии генов, формирование макромолекулярных комплексов. Известно, что у растений днРНК принимают участие в регуляции устойчивости к различным видам стресса, оказывают влияние на устойчивость к гипоксии, а также принимают участие в развитии плодов, корней и листьев. Существует три класса днРНК по расположению в структуре генома: антисмысловые, межгенные и интронные. От класса днРНК зависят многие ее функции: как она взаимодействует с белок кодирующими генами, как встраивается и каким образом производит регуляцию транскрипции. Например, антисмысловые днРНК могут взаимодействовать с комплементарной молекулой мРНК и препятствовать биосинтезу белка. На сегодняшний день разработано достаточно программ предсказания днРНК. Данные программы применяют различные методы, как самые простые основанные на измерении кодирующего потенциала, так и основанные на машинном обучении (ML). Однако ни одна из существующих программ не предусматривает классификацию днРНК по расположению в геноме. Таким образом, целью работы является разработка и оценка эффективности методов на основе глубокого обучения для предсказания и классификации длинных днРНК.

Материалы и методы: Использованы наборы последовательностей днРНК растений (*Arabidopsis thaliana*, *Zea mays* и др.), загруженных из публичных баз данных. Для классификации применены методы глубокого обучения, такие как предварительно обученная на наборе данных днРНК модель ERNIE BERT с несколькими модификациями, сверточные нейронные сети, а также классические алгоритмы машинного обучения. Для полученных моделей классификации проведена оценка точности для сравнения методов и их модификаций.

Результаты: Для исследуемого набора днРНК получены значения точности предсказания и классификации для каждого алгоритма и модификаций. Модель ERNIE BERT показала следующие результаты: F1-мера для предсказания составила 0.88 и F1-мера для классификации – 0.67. Продемонстрирована сравнительная эффективность различных подходов в задаче классификации днРНК по местоположению в геноме.

Выводы: Использование методов глубокого обучения представляет собой перспективный подход для классификации днРНК по их геномному расположению. Полученные результаты позволяют расширить знания о характеристиках различных типов днРНК и способствуют развитию методов аннотации геномов.

Финансирование: Работа выполнена при поддержке бюджетного проекта № FWNR-2022-0020.