Секция «2.2 Биоинформатика»

Предсказание субклеточной локализации РНК

Научный руководитель – Миронов Андрей Александрович

Тюкаев Артём Алексеевич

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

E-mail: artyomtyukaev@gmail.com

РНК функционируют в различных клеточных компартментах, и их роль часто определяется локализацией [1,3]. Предсказание субклеточной локализации позволяет лучше понять функции транскриптов и общие закономерности их распределения в клетке. Особенно интересна локализация длинных некодирующих РНК (днкРНК), так как они выполняют важные функции в разных компартментах клетки.

Существующие модели локализации РНК имеют ограничения и, как правило, не ориентированы на биологическую интерпретацию выявленных признаков. Целью работы стало создание интерпретируемой модели для выявления последовательностных признаков, определяющих локализацию транскриптов.

Мы проанализировали данные РНК-секвенирования по протоколу APEX-RIP [2], позволяющему различать ядерные и цитоплазматические РНК, выполнили анализ дифференциальной экспрессии и выделили мРНК днкРНК, обогащённые в соответствующих компартментах.

При разбиении данных гомологичные и паралогичные гены полностью относились либо к тренировочной, либо к тестовой выборке, что позволило избежать утечки информации из-за сходства последовательностей.

В качестве базовой модели была обучена логистическая регрессия с L1-регуляризацией на к-мерном составе транскриптов. Наилучшее качество (ROC-AUC = 0,91) достигалось при длине k = 5 (рис. 1a–c). Модели, обученные отдельно на мРНК и днкРНК, демонстрировали сопоставимое качество при тестировании на другом биотипе, что указывает на частично общую природу сигналов локализации.

Затем была обучена сверточная нейросеть из трёх сверточных слоёв. Её качество оказалось несколько выше, чем у логистической регрессии (рис. 2a–d). Для интерпретации модели были рассчитаны нуклеотидные атрибуции, и с помощью алгоритма MoDISco выделены повторяющиеся мотивы, ассоциированные с субклеточной локализацией (рис. 3a–b).

Для проверки их функциональной значимости мы сравнили предсказания для динуклеотидно перемешанных последовательностей и тех же последовательностей с вставленным наиболее распространённым «цитоплазматическим» мотивом. Вставка мотива приводила к увеличению его вклада в предсказание и смещению вероятности в сторону цитоплазматического класса (рис. 3c–d).

Источники и литература

1) Marina Chekulaeva. Mechanistic insights into the basis of widespread RNA localization // Nature Cell Biology volume 26, pages 1037–1046 (2024).
2) Sebastiaan van Heesch, et al. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes // Genome Biology volume 15, Article number: R6 (2014).

3) Pornchai Kaewsapsak, et al. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking // eLife, 6:e29224 (2017).

Иллюстрации

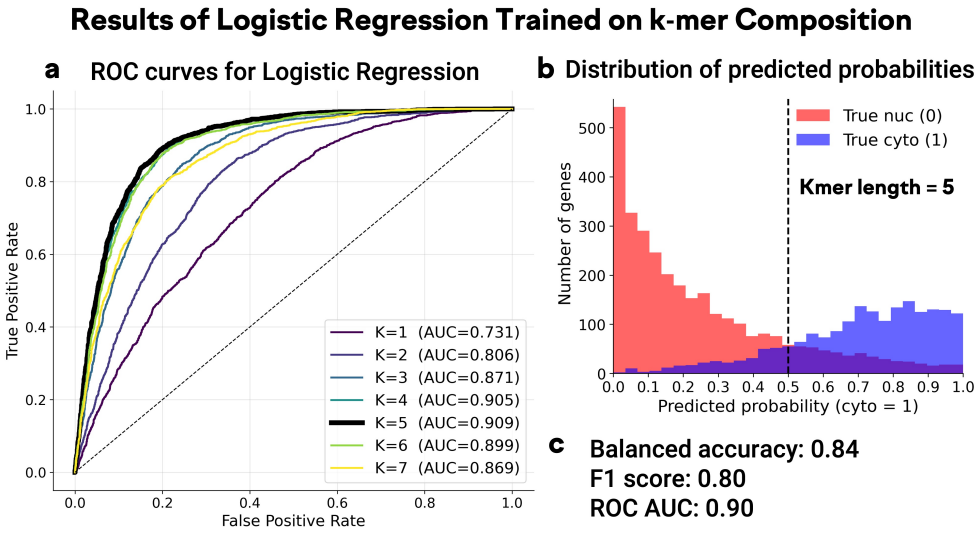

Рис. : Результаты логистической регрессии на k-мерном составе транскриптов. a — ROC-кривые для разных длин k-меров. b — Распределение предсказанных вероятностей (k = 5). c — Основные метрики качества модели (balanced accuracy, F1-score, ROC AUC).

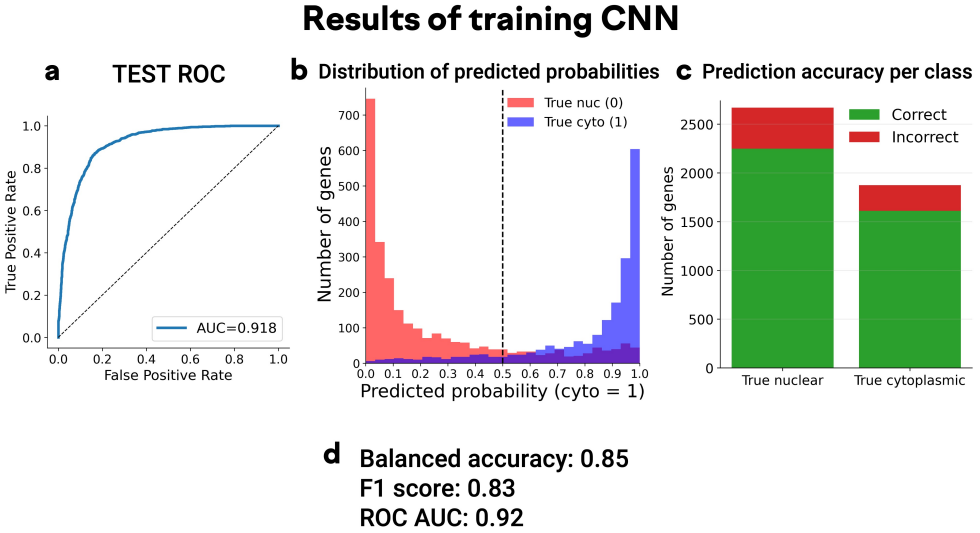

Рис. : Результаты обучения сверточной нейросети (CNN). a — ROC-кривая на тестовой выборке. b — Распределение предсказанных вероятностей. c — Точность предсказаний по классам. d — Основные метрики качества модели (balanced accuracy, F1-score, ROC AUC).

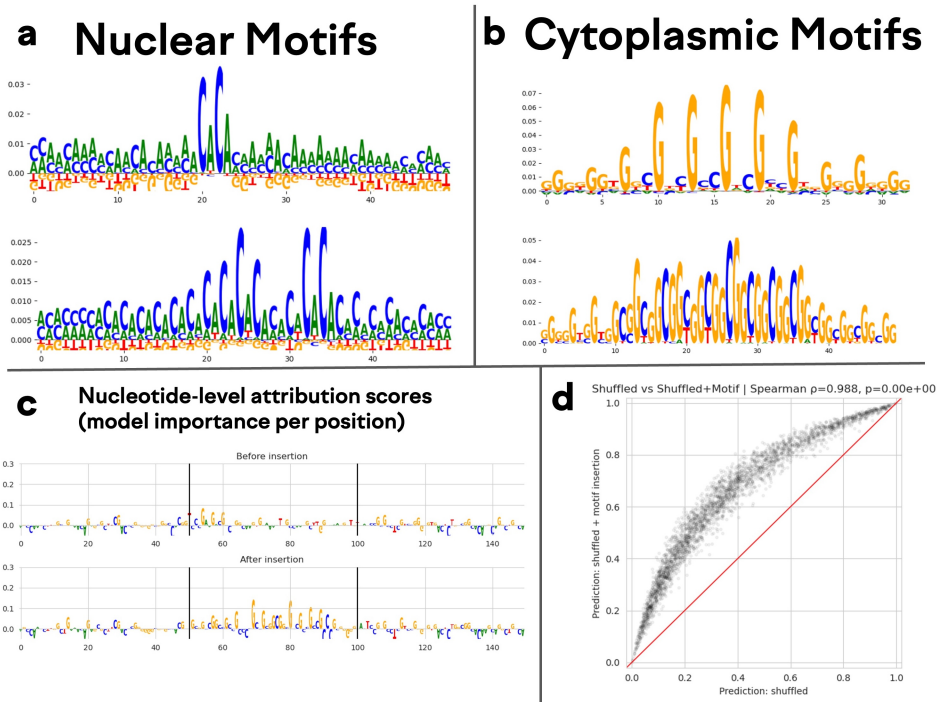

Рис. : Интерпретация CNN-модели и функциональная проверка найденных мотивов. a — Ядерные мотивы, выявленные с помощью MoDISco. b — Цитоплазматические мотивы, выявленные с помощью MoDISco. c — Вклад нуклеотидов до и после вставки мотива. d — Сравнение предсказанных вероятностей для перемешанных последовательностей и последовательностей с вставленным мотивом.