

Импутация взаимодействий белков с низкоэкспрессируемыми РНК методами глубокого обучения

Научный руководитель – Миронов Андрей Александрович

Рыбаков Арсений Константинович

Студент (специалист)

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

E-mail: irisunikova@gmail.com

Существует ряд экспериментальных подходов, направленных на определение РНК-интерактома для заданного белка (CLIP [1], eCLIP [2], fRIP [3]). Каждый из таких протоколов включает 3 стадии: фиксация существующих в клетке РНК-белковых комплексов с помощью формальдегида или ультрафиолета, выделение комплексов, в которых участвует только целевой белок, с помощью афинных к нему антител, и секвенирование РНК, оказавшихся в комплексе с целевым белком. Поскольку количество комплексов с белком для каждой фиксированной РНК пропорционально количеству этой РНК в клетке, IP-эксперименты менее склонны к обнаружению взаимодействий между белками и РНК с низкой экспрессией. Многие из таких РНК являются регуляторными [4], и потеря информации о них может помешать правильному пониманию функционирования белка. Таким образом, является необходимой импутация данных о взаимодействиях белков с низкоэкспрессируемыми РНК, для которой в последнее время набирают популярность методы на основе машинного обучения [5, 6]. В настоящей работе мы представляем нейросетевой подход для дополнения РНК-интерактома белков на основе полученных для них eCLIP-данных.

Результаты экспериментов eCLIP были получены из базы данных ENCODE [7]. Для каждого белка с известными eCLIP-данными обучалась модель для классификации РНК на взаимодействующие и не взаимодействующие с данным белком. Обученная модель далее применялась для разметки низкоэкспрессируемых РНК. Датасет для обучения формировался следующим образом. В качестве позитивных примеров выбирались сайты посадки белка на РНК, определяемые eCLIP. Негативная выборка формировалась из областей РНК, для которых не было экспериментально обнаружено взаимодействий с белком. В качестве классификатора использовалась полносвязная нейронная сеть. Поскольку такая сеть работает со входным вектором фиксированного размера, для каждой РНК рассчитывалось представление на основе вектора частот k-меров. Всего было обучено 120 моделей для клеточной линии K562 и 100 моделей для линии HepG2.

Разработанная нами архитектура демонстрирует качество, превосходящее модели на основе свёрточных нейронных сетей [5] и трансформеров [6]. В дальнейшем планируется построение модели, способной учитывать структурную информацию об РНК на основе SHAPE-экспериментов [8] по определению открытости нуклеотидов в РНК.

Источники и литература

- 1) Ule J. et al. CLIP: a method for identifying protein–RNA interaction sites in living cells // *Methods*. – 2005. – Т. 37. – №. 4. – С. 376-386.
- 2) Van Nostrand, Eric L., et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP) // *Nature methods*. No 13.6. 2016. С. 508-514.

- 3) G Hendrickson, David, et al. Widespread RNA binding by chromatin-associated proteins // Genome biology. No 17. 2016. C. 1-18.
- 4) Morris K. V., Mattick J. S. The rise of regulatory RNA //Nature Reviews Genetics. – 2014. – T. 15. – №. 6. – C. 423-437.
- 5) Sun L. et al. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures //Cell research. – 2021. – T. 31. – №. 5. – C. 495-516.
- 6) Zhu H. et al. Dynamic characterization and interpretation for protein-RNA interactions across diverse cellular conditions using HDRNet //Nature Communications. – 2023. – T. 14. – №. 1. – C. 6824.
- 7) Luo Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal //Nucleic acids research. – 2020. – T. 48. – №. D1. – C. D882-D889.
- 8) Weeks K. M., Mauger D. M. Exploring RNA structural codes with SHAPE chemistry //Accounts of chemical research. – 2011. – T. 44. – №. 12. – C. 1280-1291.