Секция «Биоинформатика»

Предсказание реактивностей нуклеотидов РНК по ее последовательности – Stanford Ribonanza RNA Folding Challenge

Научный руководитель – Пензар Дмитрий Дмитриевич

Вяльцев Валерий Владимирович

Студент (специалист)

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

E-mail: vyaltsevvaleriy7@gmail.com

Точное предсказание структуры РНК может помочь совершить революцию в науке и медицине, облегчив исследователям процесс выявления уникальных РНК мишеней для лекарств и разработки лекарств на основе РНК. Кроме того, полное понимание функционирования организма требует в свою очередь и полного понимания свойств РНК. Таким образом, разработка точной модели для предсказания структуры РНК имеет решающее значение для расширения наших знаний в области биологии и для разработки новых методов лечения заболеваний.

С целью создания модели, эффективно предсказывающей структурные характеристики молекул РНК, был создан международный конкурс Stanford Ribonanza RNA Folding на платформе Kaggle [1], в котором решение нашей команды заняло первое место, значительно опередив остальные команды со всего мира и известные SOTA-решения.

Задачей конкурса стояло предсказание реактивностей нуклеотидов, полученных с помощью методов DMS-MaP и 2A3-MaP [2], для более чем 1.5 млн последовательностей РНК. В основе методов DMS-MaP и 2A3-MaP лежит один принцип: молекулы РНК обрабатывается реагентом, который модифицирует нуклеотиды, и затем при получении кДНК из-за модификации нуклеотиды транскрибируются с ошибками. Реактивность нуклеотида отражает долю мутаций в данной позиции и зависит от структурных особенностей молекулы. В конкурсе организаторы предоставили 800 тысяч последовательностей длины от 170 до 206 п.н. в качестве обучающий данных, на их основе предлагалось обучить модель, чье качество оценивалось с помощью метрики MAE (Mean Absolute Error).

В основе нашего подхода лежит трансформерная encoder-only архитектура (Рис.1A), показавшая себя намного лучше сверточных нейросетей, которые мы также протестировали на данной задаче. Значительно качество сети улучшило добавление BPPM (Base Pair Probability Matrix), вычисленных для каждой последовательности с помощью инструмента EternaFold [3]. В архитектуре сети BPPM прибавляются к значениям attention перед применением softmax, мы также добавили 2D-сверточные слои (Рис.1B), работающие с BPPM, чтобы сеть могла выучить дополнительные полезные закономерности из них.

Организаторы конкурса в тестовый набор данных выделили большое количество последовательностей, превосходящих по длине последовательности из обучающей выборки, с целью отобрать модели, хорошо обобщающие предсказания для более длинных входных данных. В связи с этим в нашей модели мы отказались от абсолютного позиционного кодирования в пользу относительного (Рис.1B), что также позволило значительно улучшить качество предсказания. В качестве финальной модели мы использовали ансамбль из 28 одиночных моделей, который показал лучший MAE в конкурсе.

В ходе дальнейшего исследования, мы смогли еще больше улучшить качество нашей модели, использовав идеи из SqueezeFormer и добавив возможность двусторонней коммуникации между признаками, основанными на BPPM и матрицами внимания.

Источники и литература

1) Rhiju Das, Shujun He, Rui Huang, Jill Townley, Rachael Kretsch, Thomas Karagianes, John Nicol, Grace Nye, Christian Choe, Jonathan Romano, Maggie Demkin, Walter Reade, and Eterna players . (2023). Stanford Ribonanza RNA Folding. Kaggle. https://kaggle.com/competitions/stanford-ribonanza-rna-folding

2) Cheng, C. Y., Kladwang, W., Yesselman, J. D., & Das, R. (2017). RNA structure inference through chemical mapping after accidental or intentional mutations. Proceedings of the National Academy of Sciences of the United States of America, 114(37), 9876–9881. https://doi.org/10.1073/pnas.1619897114

3) Wayment-Steele, H.K., Kladwang, W., Strom, A.I. et al. RNA secondary structure packages evaluated and improved by high-throughput experiments. Nat Methods 19, 1234–1242 (2022). https://doi.org/10.1038/s41592-022-01605-0

Иллюстрации

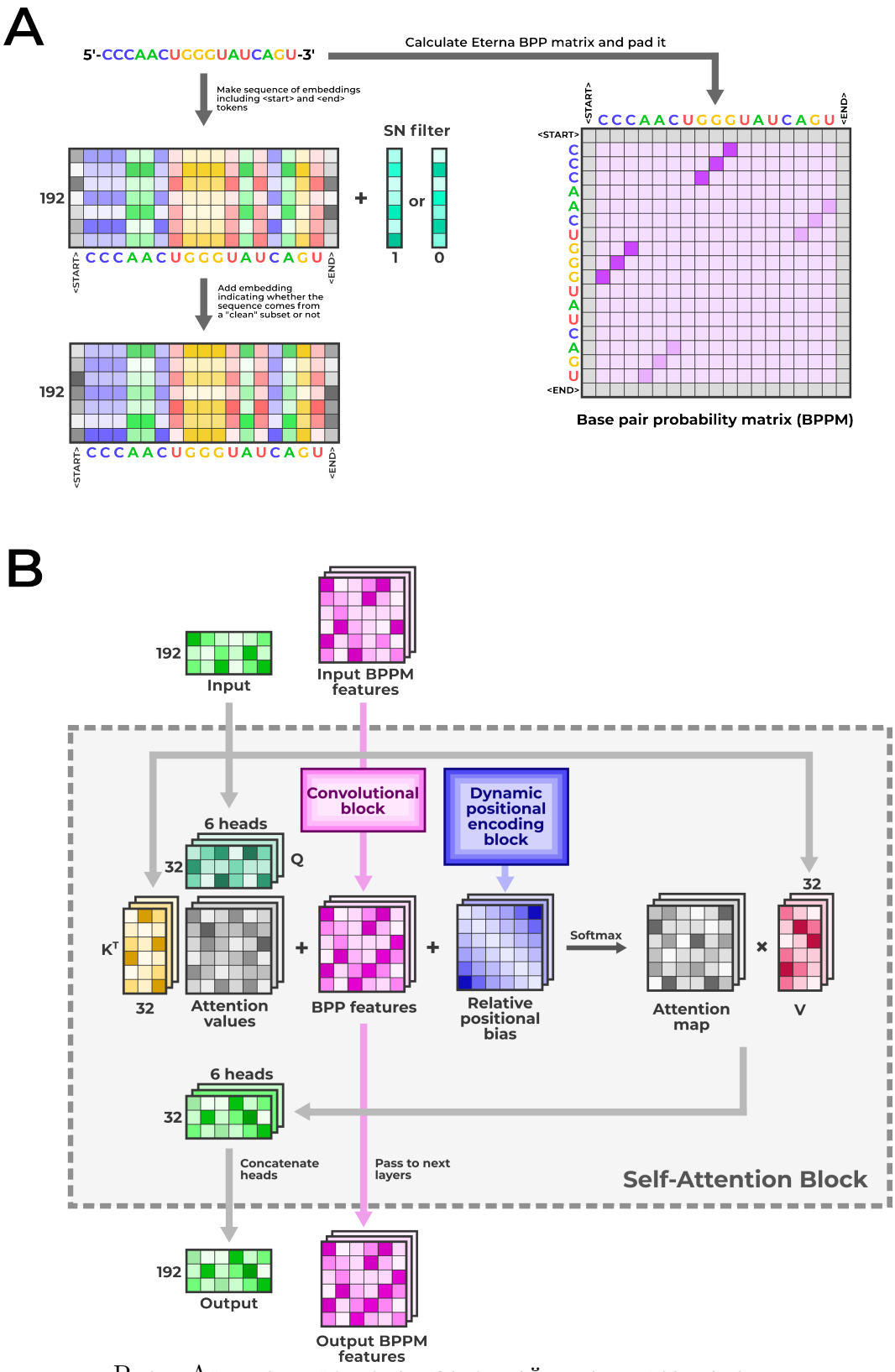

Рис. : Архитектура использованной в конкурсе модели.