Секция «2.2 Биоинформатика»

Метод сопоставления потоков для генерации регуляторных элементов с заданной специфичностью в наборе тканей

Научный руководитель – Зинкевич Арсений Олегович

Рыбаков Арсений Константинович
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: irisunikova@gmail.com

Тканеспецифическая активность генетических элементов в значительной степени определяется сигналами в регуляторных регионах, которые обычно закодированы непосредственно в нуклеотидной последовательности ДНК или РНК. Изучение механизмов работы таких элементов важно для развития технологий генной терапии, персонализированной медицины и, в частности, разработки тканеспецифических векторов для мРНК-терапии. Значительный интерес представляет задача генеративного моделирования последовательностей с заданным профилем активности в известных тканях. Подобные методы при корректном использовании могут иметь ценность для задач генной терапии и тонкой генной инженерии (для направленного редактирования последовательностей) и достижения заданного фенотипического или клинического эффекта.

Существует ряд подходов генеративного моделирования последовательностей ДНК и РНК: Discrete denoising diffusion [1], RNA latent diffusion]2], UTRGAN [3], подходы на основе авторегрессионных языковых моделей [4]. Несмотря на обилие существующих подходов, в существующих работах фактически не рассматриваются подходы к нетривиальному обусловливанию генерируемых объектов на уровень их тканеспецифической активности. Наиболее полным способом обусловливания является генерация последовательностей при условии профиля, состоящего из желаемых уровней активности, которыми она должна обладать в каждом из исследуемых типов клеток или тканей [5]. Учитывая сравнительно небольшое количество данных, доступных для потенциального обучения нейронных сетей, создание модели, способной к эффективной обработке векторных представлений числовых признаков, представляется непростой задачей и требует надёжного генеративного подхода. В настоящий момент одним из наиболее быстроразвивающихся подходов для генеративного моделирования считается метод сопоставления потоков (flow matching) [6], который сочетает в себе основополагающие принципы диффузионных моделей и простоту обучения моделей подобного класса.

Мы показываем, что Dirichlet Flow Matching (DFM) [7] - один из способов адаптации Flow Matching для дискретных данных - демонстрирует наилучшее качество генерации на ряде датасетов из фреймворка MPRA-MNIST и данных об активностях нетранслируемых областей мРНК. Мы также предлагаем набор метрик для оценки качества сгенерированных последовательностей с помощью модели-предсказателя активности по последовательности. Наконец, мы предлагаем подход на основе перемешивания последовательностей с сохранением частот k-меров для заданного k для оценки сложности выученных генеративной моделью закономерностей.

В частности, для последнего набора данных мы показываем, что корреляция Пирсона активностей, запрошенных у DFM, и активностей, предсказанных моделью-предсказателем, достигает 0.82 для клеточной линии MDA-081, однако сохраняется при перемешивании последовательностей с сохранением частот динуклеотидов. Подобные результаты поднимают вопрос о том, как низкая сложность закономерностей, выучиваемых подобными генеративными моделями соотносится с формально высоким качеством.

Источники и литература

1) Avdeyev, Pavel, et al. "Dirichlet diffusion score model for biological sequence generation." International Conference on Machine Learning. PMLR, 2023.
2) Huang, Kaixuan, et al. "Latent diffusion models for controllable rna sequence generation." arXiv preprint arXiv:2409.09828 (2024).
3) Barazandeh, Sina, et al. "UTRGAN: learning to generate 5′ UTR sequences for optimized translation efficiency and gene expression." Bioinformatics Advances 5.1 (2025): vbaf134.
4) Ihtiyar, Musa Nuri, and Arzucan Özgür. "Generative language models on nucleotide sequences of human genes." Scientific Reports 14.1 (2024): 22204.
5) Gosai, Sager J., et al. "Machine-guided design of cell-type-targeting cis-regulatory elements." Nature 634.8036 (2024): 1211-1220.
6) Lipman, Yaron, et al. "Flow matching for generative modeling." arXiv preprint arXiv:2210.02747 (2022).
7) Stark, Hannes, et al. "Dirichlet flow matching with applications to dna sequence design." arXiv preprint arXiv:2402.05841 (2024).