



МЕТОД И АЛГОРИТМЫ АНАЛИЗА И ОБРАБОТКИ ОДНОКЛЕТОЧНЫХ ОМИКСНЫХ ДАННЫХ НА ОСНОВЕ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ.

Мусакаев Руслан

Ташкентский Государственный технический университет

Аспирант (PHD)

Email: marasres08121999@mail.ru

Аннотация В статье рассматриваются современные подходы к анализу и обработке одноклеточных омиксных данных с применением методов глубокого обучения. Особое внимание уделяется архитектурам нейронных сетей, используемым для извлечения скрытых биологических закономерностей, снижению размерности, кластеризации, интеграции многомодальных данных и реконструкции клеточных траекторий. Обсуждаются преимущества нейросетевых методов по сравнению с классическими статистическими подходами, а также существующие ограничения и перспективы развития.

Ключевые слова одноклеточные данные; scRNA-seq; омиксные технологии; глубокое обучение; нейронные сети; автоэнкодеры; вариационные автоэнкодеры; генеративные модели; GAN; графовые нейронные сети; трансформеры; снижение размерности; кластеризация; интеграция данных; клеточная гетерогенность; псевдовремя; импутация данных; биоинформатика; многомодальный анализ; self-supervised обучение.

Annotatsiya Ushbu maqolada chuqur o'rganish (deep learning) usullaridan foydalangan holda bir hujayrali omiks ma'lumotlarini tahlil qilish va qayta ishlashning zamonaviy yondashuvlari ko'rib chiqiladi.

Asosiy e'tibor yashirin biologik qonuniyatlarni aniqlash, o'lchamni kamaytirish, klasterlash, ko'p modalli ma'lumotlarni integratsiya qilish hamda hujayra trayektoriyalarini rekonstruksiya qilish uchun qo'llaniladigan neyron tarmoq arxitekturalariga qaratilgan. Shuningdek, neyron tarmoqlarga asoslangan usullarning klassik statistik yondashuvlarga nisbatan afzalliklari, mavjud cheklolari va rivojlanish istiqbollari muhokama qilinadi.

Kalit so'zlar :*yagona hujayrali ma'lumotlar; scRNA-seq; omiks texnologiyalar; chuqur o'rganish; neyron tarmoqlar; autoenkoderlar; variatsion autoenkoderlar; generativ modellar; GAN; graf neyron tarmoqlari; transformerlar; o'lchamni kamaytirish; klasterlash; ma'lumotlarni integratsiya qilish; hujayra geterogenligi; psevdovaqt; ma'lumotlarni to'ldirish (imputatsiya); bioinformatika; ko'p modalli tahlil; o'z-o'zini nazorat qiluvchi o'rganish.*

Abstract *This paper discusses modern approaches to the analysis and processing of single-cell omics data using deep learning methods. Particular attention is given to neural network architectures used for extracting hidden biological patterns, dimensionality reduction, clustering, multimodal data integration, and reconstruction of cellular trajectories. The advantages of neural network-based methods over classical statistical approaches, as well as existing limitations and future development prospects, are also discussed.*

Keywords *single-cell data; scRNA-seq; omics technologies; deep learning; neural networks; autoencoders; variational autoencoders; generative models; GAN; graph neural networks; transformers; dimensionality reduction; clustering; data integration; cellular*

Введение

Развитие технологий высокопроизводительного секвенирования привело к появлению одноклеточных омиксных данных, включая транскриптомику (scRNA-seq), эпигеномику (scATAC-seq) и протеомику. Эти данные характеризуются высокой размерностью, разреженностью и шумом, что существенно усложняет их анализ.

Классические методы обработки, основанные на линейной алгебре и статистике, часто оказываются недостаточными для выявления сложных нелинейных зависимостей. В связи с этим все большую популярность приобретают методы машинного обучения, в частности нейронные сети, способные моделировать сложные структуры данных.

Особенности одноклеточных омиксных данных

Одноклеточные данные обладают рядом специфических характеристик:

- **Высокая размерность** (десятки тысяч генов на клетку);
- **Разреженность** (большое количество нулевых значений);
- **Шум и технические артефакты** (dropout-эффект);
- **Гетерогенность клеточных популяций**;
- **Нелинейные зависимости между признаками**.

Эти особенности требуют применения методов, способных эффективно извлекать информативные представления.

Архитектуры нейронных сетей для анализа данных

Автоэнкодеры

Автоэнкодеры являются одним из наиболее широко используемых инструментов для снижения размерности. Они состоят из двух частей: энкодера и декодера.

Функционально автоэнкодер реализует отображение:

$$x \rightarrow z \rightarrow x^{\wedge} \quad x \rightarrow z \rightarrow \hat{x} \quad x \rightarrow z \rightarrow x^{\wedge}$$

где xxx — исходные данные, zzz — латентное представление.

В задачах анализа одноклеточных данных применяются:

- **Вариационные автоэнкодеры (VAE)** — для моделирования распределений;

- **Деноизирующие автоэнкодеры** — для устранения шумов.

Глубокие генеративные модели

Генеративные модели позволяют восстанавливать распределения данных и генерировать новые наблюдения. Они применяются для:

- имитации клеточных состояний;
- заполнения пропущенных значений;
- нормализации данных.

Ключевые подходы:

- вариационные автоэнкодеры;
- генеративно-состязательные сети (GAN).

Графовые нейронные сети (GNN)

Одноклеточные данные часто представляются в виде графов, где вершины — клетки, а рёбра отражают сходство.

Графовые нейронные сети позволяют:

- учитывать локальную структуру данных;
- улучшать кластеризацию;

- выявлять клеточные сообщества.

Рекуррентные и трансформерные модели

Для анализа динамики клеточных состояний используются:

- рекуррентные нейронные сети (RNN);
- модели на основе механизма внимания (Transformers).

Они применяются при реконструкции траекторий дифференцировки клеток.

Основные алгоритмические задачи

Снижение размерности

Нейронные сети позволяют получать компактные латентные представления, сохраняющие биологическую значимость данных.

Преимущества: выявление нелинейных структур; устойчивость к шуму; возможность визуализации.

Кластеризация клеток

На основе латентных представлений выполняется кластеризация: с использованием k-means; иерархических методов; встроенных механизмов в нейросетях.

Это позволяет выделять типы и подтипы клеток.

Интеграция многомодальных данных

Современные исследования часто объединяют различные типы омиксных данных.

Нейросетевые методы обеспечивают: согласование различных пространств признаков; устранение батч-эффектов; совместное обучение представлений.

Реконструкция клеточных траекторий

Определение путей дифференцировки клеток является важной задачей.

Нейронные сети позволяют: моделировать псевдовремя; выявлять переходные состояния; прогнозировать будущие состояния клеток.

Импутация данных

Заполнение пропущенных значений осуществляется с помощью: автоэнкодеров; генеративных моделей.

Это повышает качество downstream-анализа.

Преимущества нейросетевых методов

- Способность моделировать сложные нелинейные зависимости;
- Высокая масштабируемость;
- Универсальность применения;
- Возможность обучения на больших объемах данных;
- Интеграция разнородных источников информации.

Ограничения и проблемы

Несмотря на значительные преимущества, нейросетевые методы имеют ряд ограничений:

- Интерпретируемость моделей;
- Высокие вычислительные затраты;
- Зависимость от качества обучающих данных;
- Проблемы воспроизводимости результатов;
- Переобучение при малых выборках.

Перспективы развития

Перспективные направления включают:

- разработку интерпретируемых моделей;
- использование self-supervised обучения;
- интеграцию пространственных омиксных данных;

- развитие гибридных методов (нейросети + статистика);
- применение фундаментальных моделей (foundation models) в биоинформатике.

Заключение

Методы глубокого обучения, в частности нейронные сети, играют ключевую роль в анализе одноклеточных омиксных данных. Они позволяют эффективно решать задачи снижения размерности, кластеризации, интеграции данных и реконструкции динамических процессов. Несмотря на существующие ограничения, дальнейшее развитие нейросетевых подходов открывает новые возможности для понимания клеточной гетерогенности и фундаментальных биологических процессов.

Список литературы

1. Lopez R., Regier J., Cole M.B., Jordan M.I., Yosef N. Deep generative modeling for single-cell transcriptomics // *Nature Methods*. 2018.
2. Eraslan G., Simon L.M., Mircea M., Mueller N.S., Theis F.J. Single-cell RNA-seq denoising using a deep count autoencoder // *Nature Communications*. 2019.
3. Kingma D.P., Welling M. Auto-Encoding Variational Bayes // *arXiv preprint arXiv:1312.6114*. 2013.
4. Goodfellow I. et al. Generative Adversarial Nets // *Advances in Neural Information Processing Systems*. 2014.
5. Kipf T.N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks // *ICLR*. 2017.
6. Vaswani A. et al. Attention Is All You Need // *NeurIPS*. 2017.

7. Stuart T., Satija R. Integrative single-cell analysis // *Nature Reviews Genetics*. 2019.
8. Luecken M.D., Theis F.J. Current best practices in single-cell RNA-seq analysis // *Molecular Systems Biology*. 2019.
9. Lotfollahi M., Naghipourfar M., Luecken M.D., Theis F.J. Mapping single-cell data to reference atlases by transfer learning // *Nature Biotechnology*. 2022.
10. Cao J., Spielmann M., Qiu X. et al. The single-cell transcriptional landscape of mammalian organogenesis // *Nature*. 2019.