УДК 004.89

РАЗРАБОТКА АПТАМЕРОВ НА ОСНОВЕ ЭМБЕДДИНГОВ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Заикина А.А. (ИТМО)
Научный руководитель – кандидат химических наук, Серов Н.С. (ИТМО)

Введение. Аптамеры - синтетические одноцепочечные молекулы ДНК или РНК длиной 20-60 нуклеотидов, способные с высокой специфичностью связываться с белками, пептидами и малыми молекулами [1], различая даже варианты, отличающиеся одной аминокислотой [2]. Их клиническая значимость подтверждена применением аптамерного препарата Macugen [3] и использованием в диагностических анализах формата ELASA [4]. Основным методом получения аптамеров остаётся технология SELEX: итеративный процесс in vitro отбора и амплификации нуклеотидных последовательностей [5], который, несмотря на эффективность, является длительным и трудоёмким. Существующие вычислительные подходы ускоряют разработку аптамеров, однако часто ограничены узкими наборами данных и плохо обобщаются на новые мишени. Развитие методов глубокого обучения, в частности архитектур Transformer, открывает возможности для генерации аптамеров на основе информации о белковой мишени, что актуализирует задачу создания моделей, способных предсказывать аптамерные последовательности по аминокислотной последовательности целевого белка.

Основная часть. Предлагаемый подход основан на использовании архитектуры Transformer, позволяющей учитывать дальние зависимости в биологических последовательностях. Генерация аптамеров формулируется как задача, аналогичная машинному переводу: аминокислотная последовательность белка рассматривается как исходный язык, а нуклеотидная последовательность аптамера — как целевой. Такой подход позволяет учитывать контекст белковой мишени при генерации потенциально специфичных последовательностей.

Для обучения модели был сформирован набор данных из 4693 пар «белок–аптамер», собранных из открытых источников, включая APIPred [6], Apta-Index™ и UTexas [7]. Белковые последовательности кодировались с использованием 1024-мерных эмбеддингов, полученных из предварительно обученной языковой модели белков Evolutionary Scale Modeling, что обеспечивало компактное и информативное представление мишени.

Модель реализована в формате «кодировщик–декодировщик»: на основе эмбеддингов белка формируется латентное представление, по которому авторегрессивно генерируется нуклеотидная последовательность аптамера.

Оценка качества проводилась на уровне статистических характеристик последовательностей. Для сравнения исходных и сгенерированных аптамеров использовались расстояние Левенштейна, энтропия Шеннона, а также распределения длины и содержания GC. Различия между распределениями оценивались с помощью дивергенции Дженсена-Шеннона — симметризированной версии дивергенции Кульбака-Лейблера [8, 9]. Получены умеренные расхождения между исходными и сгенерированными последовательностями: 0,3813 бит для энтропии Шеннона, 0,3239 бит для расстояния Левенштейна, 0,2522 бит для длины и 0,4756 бит для GC состава. Это свидетельствует о том, что модель воспроизводит глобальные структурные характеристики аптамеров и уровень их вариабельности, сохраняя при этом новизну генерируемых последовательностей.

Выводы. Представлена модель глубокого обучения, способная генерировать аптамеры,

специфичные к заданной белковой мишени, на основе эмбеддингов белковых последовательностей. Модель демонстрирует способность воспроизводить ключевые глобальные характеристики природных аптамеров, обеспечивая при этом новизну сгенерированных последовательностей.

Разработанный подход открывает возможность масштабируемого и целенаправленного проектирования аптамеров in silico, что позволяет существенно сократить объём экспериментального скрининга. В практическом контексте это ускоряет выявление перспективных кандидатов для диагностики и терапии, снижая лабораторные затраты и повышая эффективность доклинических исследований.

Список использованных источников:

1. Yang, Lucy F., et al. "Aptamers 101: aptamer discovery and in vitro applications in biosensors and separations." Chemical science 14.19 (2023): 4961-4978.
2. Chen, Liang, et al. "The isolation of an RNA aptamer targeting to p53 protein with single amino acid mutation." Proceedings of the National Academy of Sciences 112.32 (2015): 10002-10007.
3. Lee, Joon-Hwa, et al. "A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF165." Proceedings of the National Academy of Sciences 102.52 (2005): 18902-18907.
4. Toh, Saw Yi, et al. "Aptamers as a replacement for antibodies in enzyme-linked immunosorbent assay." Biosensors and bioelectronics 64 (2015): 392-403.
5. Nasaev, Shamsudin Sh, et al. "Molecular Modeling Methods in the Development of Affine and Specific Protein-Binding Agents." Biochemistry (Moscow) 89.8 (2024): 1451-1473.
6. Fang, Zheng, et al. "APIPred: An XGBoost-Based Method for Predicting Aptamer–Protein Interactions." Journal of chemical information and modeling 64.7 (2023): 2290-2301.
7. Askari, Ali, et al. "UTexas Aptamer Database: the collection and long-term preservation of aptamer sequence information." Nucleic Acids Research 52.D1 (2024): D351-D359.
8. Zielezinski, Andrzej, et al. "Benchmarking of alignment-free sequence comparison methods." Genome biology 20.1 (2019): 144.
9. Lin, Jianhua. "Divergence measures based on the Shannon entropy." IEEE Transactions on Information theory 37.1 (2002): 145-151.