УДК 004.89

МУЛЬТИМОДАЛЬНОЕ МОДЕЛИРОВАНИЕ МИРНК ДУПЛЕКСОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ЭФФЕКТИВНОСТИ НОКДАУНА ГЕНОВ

Шатковский Д.П. (ИТМО), Дружининский С.М. (ИТМО)
Научный руководитель – кандидат химических наук, Серов Н.С. (ИТМО)

Введение. Малые интерферирующие РНК (миРНК) представляют собой двухцепочечные молекулы, участвующие в естественном механизме РНК-интерференции и обеспечивающие направленное подавление экспрессии генов посредством деградации комплементарной матричной РНК [1].

Применение методов классического машинного и глубокого обучения позволяет оптимизировать процесс *in silico* отбора нуклеиновых кислот-кандидатов. Однако большинство существующих моделей ограничено упрощенными представлениями последовательностей и, как правило, не учитывает химические модификации, физико-химические свойства нуклеотидов и межцепочечные взаимодействия [2]. Подобные ограничения могут снижать точность предсказаний и обобщающую способность моделей. В связи с этим актуальной задачей является разработка архитектурного подхода, позволяющего комплексно моделировать дуплекс миРНК как единую физико-химическую систему с учетом структурных особенностей, химических модификаций и характера взаимодействия между цепями.

Основная часть. В настоящем исследовании проведен сравнительный анализ архитектурных подходов к моделированию двухцепочечных молекул миРНК. Для векторного кодирования последовательностей использовалось агрегирование данных, включающих информацию о нуклеотидном составе, химических модификациях, а также эмбеддингах модели RNA-FM и физико-химических дескрипторах, рассчитанных с использованием библиотеки RDKit.

Архитектура модели включает специализированные слои, направленные на экстракцию внутрицепочечных контекстных признаков, моделирование межцепочечного взаимодействия и обеспечение перестановочной инвариантности цепей дуплекса. В качестве энкодеров отдельных цепей сравнивались архитектуры Transformer, BiLSTM, SSM и CNN. Для моделирования взаимодействий смысловой и антисмысловой цепей использовались различные механизмы, основанные на Cross-Attention (Single-Head, Multi-Head), а также их модификации с использованием Bilinear Fusion. На этапе финального объединения признаков анализировались стратегии, основанные на конкатенации, суммировании, усреднении, ранжировании по важности и механизмах внимания.

Сформированные представления использовались для решения задачи регрессии – предсказания эффективности нокдауна с помощью полносвязного слоя. Обучение и валидация проводились на наборе данных, включающем более 3500 пар последовательностей с указанием химических модификаций нуклеотидов, условий проведения эксперимента и наблюдаемой эффективности нокдауна. Качество моделей оценивалось с использованием метрик R^2 и RMSE.

По итогам сравнительного анализа различных конфигураций наилучшие результаты продемонстрировала модель, включающая Transformer в качестве внутрицепочечного энкодера и Cross-Attention с Bilinear Fusion для моделирования межцепочечного взаимодействия ($R^2 \approx 0.73$, RMSE ≈ 15.44 %).

Для сравнения, предложенный нами ранее пайплайн регрессионного метаобучения, решающий ту же downstream задачу с использованием алгоритма LightGBM, демонстрировал $R^2 \approx 0.84$ (SOTA) [3]. Известно, что в задачах с табличными признаками и ограниченным объемом данных ансамблевые методы градиентного бустинга, как правило, демонстрируют более высокую устойчивость по сравнению с нейросетевыми архитектурами. В этом контексте достигнутый результат подтверждает применимость предложенного нейросетевого подхода для решения задачи прогнозирования эффективности миРНК в режиме low-data.

Выводы. Полученные результаты подтверждают целесообразность комплексного моделирования дуплекса миРНК с учётом внутрицепочечных контекстных зависимостей и межцепочечного взаимодействия. Разработанный подход может быть использован для повышения точности *in silico* отбора кандидатов миРНК и оптимизации процесса их экспериментальной валидации.

Список использованных источников:

1. Friedrich M., Aigner A. Therapeutic siRNA: state-of-the-art and future perspectives // BioDrugs. – 2022. – 36. – Pp. 549-571.

2. Martinelli D. D. Machine learning for siRNA efficiency prediction: a systematic review // Health Sciences Review. – 2024. – 11. – 100157.

3. Golovkin I., Shatkovskii D., Serov N. Two-Stage Probability-Enhanced Regression on Property Matrices and LLM Embeddings Enables State-of-the-Art Prediction of Gene Knockdown by Modified siRNAs // Int J Mol Sci. – 2025. – 26(24). – 11791.