УДК 575.112

СОЗДАНИЕ БАЗЫ ДАННЫХ ЭКСПЕРИМЕНТОВ ДЛЯ РАЗРАБОТКИ МЕТОДА СИНТЕЗА СТРУКТУРЫ ГИБРИДИЗАЦИОННЫХ ЗОНДОВ НА ОСНОВЕ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ ИИ

Дравгелис В.А. (ИТМО)

Научные руководители – кандидат технических наук, доцент ФИТиП Сергушичев А.А. (ИТМО); кандидат технических наук, доцент ИПКН Муравьёв С.Б. (ИТМО)

Введение. Метод Fluorescence In Situ Hybridization (FISH) является мощным инструментом для обнаружения и локализации специфических последовательностей ДНК или РНК в клетках и тканях. Этот метод широко применяется в диагностике генетических заболеваний, исследовании рака и хромосомном анализе, благодаря своей высокой специфичности и возможности одновременного обнаружения нескольких мишеней [1]. Однако создание эффективных зондов для FISH сопряжено с рядом сложностей, таких как конформация ДНК, содержание GC и модификации, например метилирование, которые могут влиять на гибридизацию [2]. Целью нашего исследования является разработка метода синтеза структуры гибридизационных зондов на основе генеративных моделей искусственного интеллекта (ИИ), что требует создания базы данных экспериментов для обучения и валидации модели.

Основная часть. Наш проект направлен на создание базы данных экспериментов, которая будет использоваться для разработки модели ИИ, способной генерировать оптимальные последовательности зондов для метода FISH. Наименьшей единицей базы данных является JSON-файл, содержащий информацию об эксперименте и зондах. JSON был выбран благодаря своей гибкости, удобству работы с большими языковыми моделями (LLM) и возможности автоматизированной проверки данных [3]. Каждый JSON-файл разделён на две основные части: данные об эксперименте и информацию о зондах, включая их последовательности и модификации.

Для извлечения данных из научных статей мы используем общедоступные LLM, такие как DeepSeek и ChatPDF, которые позволяют автоматизировать процесс и снизить затраты. Однако эти модели могут генерировать неточные данные, поэтому мы внедрили процесс валидации, включающий проверку соответствия JSON-файлов заданной схеме. Алгоритм проверки контролирует наличие всех необходимых свойств, соответствие типов данных и регулярных выражений для специфических полей, таких как последовательности зондов и их модификации [4].

На данный момент мы разработали структуру JSON, создали запросы для LLM, собрали статьи и извлекли последовательности зондов с их метаданными. Например, успешно извлечены последовательности зондов с модификациями ROX и BHQ2, которые прошли проверку на соответствие нашей схеме.

Выводы. Созданная база данных экспериментов станет основой для разработки генеративной модели ИИ, способной синтезировать оптимальные последовательности зондов для метода FISH. Это позволит учитывать сложные факторы, влияющие на гибридизацию, и повысит точность и эффективность метода, что особенно важно для диагностики и исследований в области генетики и онкологии.

Список использованных источников:

1. Speicher M.R., Carter N.P. The new cytogenetics: blurring the boundaries with molecular biology // Nature Reviews Genetics. – 2005. – Vol. 6. – P. 782–792.

2. Levsky J.M., Singer R.H. Fluorescence in situ hybridization: past, present and future // Journal of Cell Science. – 2003. – Vol. 116. – P. 2833–2838.
3. JSON: JavaScript Object Notation. URL: https://www.json.org/json-en.html (дата обращения: 10.10.2023).
4. Rajpurkar, P., Chen, E., Banerjee, O. et al. AI in health and medicine. // Nat Med. – 2022 – № 28, P. 31–38

Автор _____ Дравгелис В.А.

Научный руководитель_____ Сергушичев А.А.

Научный руководитель_____ Муравьёв С.Б.