УДК 575.112

РАЗРАБОТКА И ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА, ГЕНЕРАЦИИ И ОПТИМИЗАЦИИ ГИБРИДИЗАЦИОННЫХ ЗОНДОВ

Сердюков А.Н. (ИТМО), Иванов А.Б. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Муравьев С.Б.
(ИТМО)

Введение. Гибридационные зонды широко применяются в молекулярной биологии и медицине для детекции специфических последовательностей ДНК. Оптимальный выбор зондов важен для точности и эффективности некоторых молекулярных методов, таких как метод флуоресцентной *in-situ* гибридизации (FISH). Эти методы позволяют проверить наличие заданной подпоследовательности в молекуле ДНК, определить её расположение в молекуле или количественно оценить частоту вхождения. Две ключевые метрики качества – вероятность гибридизации и специфичность – определяют эффективность гибридационных зондов. В простейшем случае в качестве гибридационного зонда может выступать комплементарная к заданной последовательность. Однако, в виду некоторых ограничений, таких как сложная пространственная конформация молекулы ДНК или модификация некоторых её участков, зачастую требуется оптимизация зонда. Ручная оптимизация является трудозатратной и плохо масштабируемой задачей. Автоматизация подбора зондов с использованием методов машинного обучения и генеративных моделей искусственного интеллекта позволяет значительно сократить затраты ресурсов и повысить качество гибридационных методов. Данное исследование направлено на разработку алгоритмов и программных инструментов для анализа, генерации и оптимизации гибридационных зондов.

Основная часть. Главной Целью исследования является создание инструмента, способного в полностью автоматическом режиме генерировать гибридационные зонды по заданной целевой последовательности ДНК с учётом структурных факторов и параметров гибридизации. Для этого были выделены следующие группы задач:

1. Сбор информации и анализ существующих гибридационных экспериментов;
2. Формирование базы данных известных гибридационных зондов и условий экспериментов;
3. Разработка предсказательных моделей для оценки вероятности гибридизации и специфичности зондов;
4. Создание генеративной модели для автоматизированной генерации новых зондов;
5. Валидация полученных зондов в реальных лабораторных экспериментах.

Методы машинного обучения, в частности глубокие нейронные сети, позволяют предсказывать гибридационные свойства зондов, учитывая пространственную структуру ДНК. Генеративные модели, такие как вариационные автокодировщики и трансформеры, могут синтезировать новые последовательности зондов с оптимальными параметрами [1, 2]. Возможно дальнейшее использование данных Hi-C, что позволит учитывать трёхмерную организацию генома.

Для обучения и тестирования моделей создаётся и наполняется специализированная база данных, содержащая последовательности зондов, их

экспериментально подтверждённые характеристики и условия проведения гибридных тестов. Для массового анализа выбранного корпуса статей и прочих научных публикаций применяются решения на основе больших языковых моделей. Для упрощения последующего анализа было принято решение собрать экспериментальные данные в формате JSON [3]. Преимуществом использования больших языковых моделей является возможность составления такого запроса, ответ на который модель представит в структурированном формате, позволяющем последующую автоматизированную обработку. Ввиду того, что большие языковые модели могут сгенерировать ответ, содержащий некорректные данные, например такие гибридные эксперименты, описание которых в публикации не содержится, был также разработан простейший способ проверки на основе регулярных выражений.

Выводы. На данный момент разработан инструментарий для массового анализа научных публикаций по теме гибридных зондов, при помощи больших языковых моделей из статей извлекаются данные о гибридных экспериментах. В качестве структурированного формата был выбран формат JSON. Был составлен список атрибутов гибридного эксперимента, подлежащих извлечению, и разработана JSON-схема. На основе полученной схемы был спроектирован соответствующий запрос к большой языковой модели, ответ на который генерируется в заданном формате по заданной схеме. Результаты извлечения данных об экспериментах были собраны в базу данных. В дальнейшем планируется перейти к разработке модулей для предсказания метрик качества гибридных зондов, а также к созданию и обучению генеративной модели на основе полученных на данном этапе данных. Перспективы дальнейшего исследования включают возможность интеграции дополнительных источников данных, таких как Hi-C, и других параметров молекулярных взаимодействий [4].

Список использованных источников:

1. Killoran, N., Lee, L. J., DeLong, A., Duvenaud, D., Frey, B. J. Generating and designing DNA with deep generative models [Электронный ресурс] // arXiv preprint. 2017. arXiv:1712.06148. URL: <https://arxiv.org/abs/1712.06148> (дата обращения: 10.02.2025).
2. Zhu, W. Analyzing DNA Hybridization via machine learning [Электронный ресурс] // arXiv preprint. 2018. arXiv:1803.11062. URL: <https://arxiv.org/abs/1803.11062> (дата обращения: 10.02.2025).
3. Bray T. The JavaScript Object Notation (JSON) Data Interchange Format // RFC 8259. — 2017. — URL: <https://datatracker.ietf.org/doc/html/rfc8259> (дата обращения: 10.02.2025).
4. Nguyen, E., et al. Sequence modeling and design from molecular to genome scale with Evo // Science. 2024. DOI: 10.1126/science.ado9336.

Сердюков А.Н. (автор)

Подпись

Муравьёв С.Б. (научный руководитель)

Подпись