УДК 004.89

ОТКРЫТИЕ НОВЫХ ДНКЗИМОВ С ИСПОЛЬЗОВАНИЕМ ГЕНЕРАТИВНЫХ МЕТОДОВ И ИЕРАРХИЧЕСКОГО СКРИНИНГА

Головкин И.А.[1], **Дружининский С.М.**[1], **Литуновский Д.Ю.**[1]
Научный руководитель – канд. хим. наук, Серов Н.С.[1]
[1] Университет ИТМО
golovkin2003@list.ru

Работа выполнена в рамках темы НИРСИИ №640100 «Мультимодальное моделирование с использованием графов знаний для прогнозирования свойств и условной генерации сенсорных биополимеров».

Введение

ДНКзимы (деоксирибозимы) представляют собой каталитические молекулы ДНК, способные осуществлять специфическое расщепление РНК и другие химические реакции. Со времени открытия первого ДНКзима в 1994 году данное направление активно развивается, находя применение в молекулярной диагностике, терапии онкологических и воспалительных заболеваний, а также в фундаментальных исследованиях регуляции экспрессии генов [1, 4].

Несмотря на значительный прогресс, ключевой проблемой остается дизайн каталитического ядра ДНКзима. В настоящее время разработка новых последовательностей в значительной степени опирается на методы in vitro-селекции и модификацию известных консенсусных мотивов (например, 8–17 и 10–23). Эффективность таких подходов ограничена высокой контекстной зависимостью активности и необходимостью ресурсозатратного экспериментального скрининга [2].

Современные методы машинного обучения и генеративного моделирования открывают возможности для перехода от эмпирического подбора к целенаправленному конструированию новых каталитических центров. Однако отсутствие масштабных, качественно аннотированных наборов данных и специализированных моделей затрудняет разработку эффективных алгоритмов de novo-дизайна ДНКзимов.

Цель настоящей работы — разработка многоэтапной физически-информированной генеративной платформы для поиска и отбора новых каталитических ядер ДНКзимов с предсказуемыми структурными и энергетическими характеристиками.

Основная часть

Предложенный подход основан на интеграции генеративных моделей и последовательной иерархической фильтрации кандидатов с учетом физико-химических параметров.

На первом этапе формируются обучающие выборки с использованием двух стратегий: термодинамической (отбор по минимальной свободной энергии молекулы) и распределительной (сохранение статистических характеристик реальных активных последовательностей — длины, GC-состава, энтропии, расстояния Левенштейна). Это позволяет создать репрезентативное пространство последовательностей для предварительного обучения моделей.

Для создания кандидатов используются генеративные модели: рекуррентные нейронные сети (LSTM), генеративно-состязательные сети (GAN) и большие языковые модели (LLM). Каждая модель обучается в режиме предварительного обучения на

расширенном наборе данных с последующей тонкой настройкой на выборке экспериментально подтвержденных ДНКзимов.

Для повышения надежности отбора применяется многоступенчатая система скрининга. На этапе первичной фильтрации используется классификационная модель на основе градиентного бустинга, обученная различать активные и неактивные последовательности. Это позволяет существенно снизить долю ложноположительных кандидатов и сократить объем последующей экспериментальной проверки. Далее учитываются термодинамические характеристики (минимальная свободная энергия), структурные признаки, а также соответствие статистическим распределениям реальных ДНКзимов. Дополнительно проводится кластеризация эмбеддингов последовательностей для предсказания вероятного кофактора (двухвалентных катионов), а также пространственное моделирование трехмерной структуры с оценкой способности формировать каталитически компетентную конформацию.

Таким образом, предлагаемый метод сочетает генеративное моделирование с физически обоснованными критериями отбора, что обеспечивает направленный поиск новых каталитических ядер.

Выводы

Определена последовательность действий для генеративного дизайна каталитических ядер ДНКзимов, интегрирующая машинное обучение и физико-химические критерии оценки.

Применение классификационных и регрессионных моделей существенно снижает затраты на экспериментальный скрининг и повышает вероятность выявления функционально активных кандидатов.
Сформированы два набора данных, прошедших все этапы скрининга, для лабораторной проверки.

Практическое использование результатов возможно при разработке новых терапевтических ДНКзимов, молекулярных сенсоров и диагностических платформ. Предлагаемая методология может быть внедрена в лабораторную практику в качестве предварительного этапа in silico-отбора перед экспериментальной валидацией.

Литература

1. Breaker, R. R. In Vitro Selection of Catalytic Polynucleotides / R. R. Breaker // Chemical Reviews. – 1997. – Т. 97. – № 2. – С. 371-390.

2. Caramori, G. Allergen Responses Modified by a GATA3 DNAzyme / G. Caramori, K. F. Chung, P. J. Barnes // The New England Journal of Medicine. – 2015. – Т. 373. – № 12. – С. 1176-1177.

3. The critical role of basement membrane-independent laminin gamma 1 chain during axon regeneration in the CNS / B. Grimpe, S. Dong, C. Doller [и др.] // The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. – 2002. – Т. 22. – № 8. – С. 3144-3160.