

SEQUENCECRAFT: ЦИФРОВОЙ РЕСУРС НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИССЛЕДОВАТЕЛЬСКОГО АНАЛИЗА РНК-РАСЩЕПЛЯЮЩИХ ДЕЗОКСИРИБОЗИМОВ

Дин Е. (ИТМО), Еремеева М.А. (ИТМО).

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

Введение. Дезоксирибозимы или ДНКзимы представляют собой искусственные короткие последовательности ДНК, обладающие многими каталитическими свойствами. В частности, ДНКзимы, способные расщеплять последовательности РНК, имеют огромный потенциал в генной терапии, аналитическом обнаружении маркеров заболеваний и биосенсинге для детекции ионов металлов, бактерий и вирусов [1–3]. Особенно перспективное применение ДНКзимов заключается в их способности подавлять гены посредством специфического связывания и ингибирования целевой РНК [4]. Однако большинство существующих каталитических ядер ДНКзимов имеют низкую эффективность в экспериментах *in vivo*, тогда как SELEX, основанный на скрининге *in vitro*, предлагает длительный и дорогостоящий цикл отбора со средней вероятностью успеха ~ 30%. Более того, SELEX не позволяет проводить прямой отбор химически модифицированных ДНКзимов, которые демонстрируют более высокую активность *in vivo* [5]. Поэтому существует огромная потребность в подходе *in silico* для исследовательского анализа ядер ДНКзимов, чтобы облегчить открытие новых каталитических ядер с превосходной активностью.

Основная часть. В данном исследовании были использованы сведения о более 350 маркированных образцах ДНКзимов с известной каталитической активностью, которая характеризуется наблюдаемой константой скорости. Обработка данных включала в себя приведение к стандартным единицам измерения, дезагрегирование буфера на компоненты и получение характеристик кофактора из библиотеки Python Materials Genomics [6]. С целью найти наиболее показательный способ представления последовательностей ДНК, было проведено сравнение пяти различных подходов к их кодированию: матричное кодирование, PyBioMed [7], Python seqfold, большая языковая модель HyenaDNA и новый подход к представлению данных на основе сверточного автокодировщика, который был обучен на белковых, РНК и ДНК последовательностях. Оценка эффективности представлений проводилась на основе предсказания каталитической активности с помощью модели случайного леса. Среди перечисленных подходов наиболее высокий результат показали скрытые представления сверточного автокодировщика ($Q^2 = 0.259$, $R^2_{\text{test}} = 0.267$). Далее, был проведен процесс скрининга моделей XGBoost, LightGBM, CatBoost и RandomForest для определения оптимального алгоритма. Перед оценкой модели каждый алгоритм прошел процедуру оптимизации гиперпараметров с использованием рандомизированного поиска CV. Затем был отобран проверочный набор данных, содержащий 53 уникальные последовательности, отличный от обучающего и тестового наборов. Качество моделей оценивалось с использованием R^2 и RMSE на тестовых и проверочных наборах, а также метрики Q^2 с помощью 5-кратной стратифицированной перекрестной проверки на обучающем наборе. Далее, лучшая модель LightGBM прошла процесс отбора признаков с помощью рекурсивного исключения, в результате чего были идентифицированы 24 ключевых признака для производительности модели.

Выводы. Финальный алгоритм продемонстрировал точность $Q^2 = 0,93$ и $\text{RMSE} = 0.52$ в количественном предсказании константы наблюдаемой скорости. Анализ признаков показал превосходящую важность таких экспериментальных условий, как концентрация кофактора, pH и температура. Среди них концентрация кофактора оказывает наибольшее влияние на каталитическую активность, что согласуется с экспериментальными знаниями, так как этот

параметр является центральным для процесса гидролиза. Более того, модель смогла выделить важность таких признаков кофактора, как заряд и сродство к электрону, влияющих на нуклеофильную атаку на фосфодиэфирную связь.

Для популяризации и практического внедрения полученного алгоритма была разработана и развернута платформа SequenceCraft с открытым исходным кодом, позволяющая ученым-экспериментаторам выполнять исследовательский анализ ДНКзимов *in silico*. На платформе предоставлена возможность неограниченного использования инструментов для предсказания каталитической активности и предварительного анализа кандидата, а также продемонстрирована открытая база данных, которая применялась для обучения модели.

Список использованных источников:

1. DNazymes as activity-based sensors for metal ions: recent applications, demonstrated advantages, current challenges, and future directions / Lake, R. J., Yang, Z., Zhang, J. [et al.]. – DOI 10.1021/acs.accounts.9b00419 // *Analytical Chemistry*. — 2019. — Vol. 52, № 12. – p. 3275-3286. – URL: <https://pubs.acs.org/doi/10.1021/acs.accounts.9b00419>.
2. A DNzyme-based colorimetric paper sensor for *Helicobacter pylori* / Ali, M. M., Wolfe, M., Tram, K. [et al.]. – DOI 10.1002/ange.201901873 // *Angewandte Chemie*. — 2019. — Vol. 131, № 29. – p. 10012-10016. – URL: <https://onlinelibrary.wiley.com/doi/10.1002/ange.201901873>.
3. DNzyme walker for homogeneous detection of enterovirus EV71 and CVB3 / Du, M., Zheng, J., Tian, S. [et al.]. – DOI 10.1021/acs.analchem.1c00335 // *Analytical Chemistry*. — 2021. — Vol. 93, № 13. – p. 5606-5611. – URL: <https://pubs.acs.org/doi/10.1021/acs.analchem.1c00335>.
4. On gene silencing by the X10-23 DNzyme / Taylor, A. I., Holliger, P. – DOI 10.1038/s41557-021-00645-x // *Nature Chemistry*. — 2022. — Vol. 14, № 8. – p. 855-858. – URL: <https://www.nature.com/articles/s41557-022-00990-5>.
5. SELEX: Critical factors and optimization strategies for successful aptamer selection / Kohlberger, M., Gadermaier, G. – DOI 10.1002/bab.2244 // *Biotechnology and applied biochemistry*. — 2022. — Vol. 69, № 5. – p. 1771-1792. – URL: <https://iubmb.onlinelibrary.wiley.com/doi/10.1002/bab.2244>.
6. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis / Ong, S. P., Richards, W. D., Jain, A. [et al.]. – DOI 10.1016/j.commatsci.2012.10.028 // *Computational Materials Science*. — 2013. — Vol. 68. – p. 314-319. – URL: <https://doi.org/10.1016/j.commatsci.2012.10.028>.
7. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions / Dong, J., Yao, Z. J., Zhang, L. [et al.]. – DOI 10.1186/s13321-018-0270-2 // *Journal of cheminformatics*. — 2018. — Vol. 10. – p. 1-11. – URL: <https://link.springer.com/article/10.1186/s13321-018-0270-2>.