

УДК 004.89

ИНТЕРПРЕТИРУЕМЫЕ ДЕСКРИПТОРЫ ПЕПТИДОВ С УЧЁТОМ ФИЗИКО-ХИМИЧЕСКИХ СВОЙСТВ АМИНОКИСЛОТ И МЕТАЛЛИЧЕСКИХ КОФАКТОРОВ

Нам Е.В. (Университет ИТМО), Серов Н.С. (Университет ИТМО)

Научный руководитель – профессор, доктор химических наук, Виноградов В.В.
(Университет ИТМО)

Введение. В последние десятилетия неуклонно растет количество разрабатываемых белковых препаратов: объем рынка рекомбинантных белков в 2023 году составил 2.49 млрд долларов США, при этом прогнозируется дальнейший рост спроса на данный класс терапевтических средств. Это обусловлено их высокой селективностью и эффективностью, терапевтические белки обладают значительным потенциалом в лечении различных заболеваний, включая рак, иммунные нарушения и редкие генетические заболевания [1]. Однако существующие методы разработки белковых препаратов требуют больших временных и финансовых ресурсов, что делает актуальным использование инновационных подходов, таких как машинное обучение и искусственный интеллект [2]. Подобные подходы уже показали свою эффективность в ряде областей, например, в органической химии, где методы машинного обучения позволяют прогнозировать продукты органических химических реакций, их активность и многое другое [3-4]. Это возможно в первую очередь благодаря существованию молекулярных дескрипторов, позволяющих эффективно передавать информацию о химических соединениях для обучения моделей. На данный момент существуют предсказательные и генеративные модели, применяющиеся в области исследований новых белковых препаратов [2]. Однако отсутствие дескрипторов белков, сочетающих описание их строения и химических свойств, делает невозможной разработку высокоэффективных моделей машинного и глубокого обучения.

Основная часть. Целью данной работы является разработка дескрипторов белков, несущих информацию об их структурных и химических свойствах. Объектом исследования являются дескрипторы, характеризующие фолдинг – процесс организации аминокислотных цепей в третичную структуру белка. Без учета данного явления невозможно создать точные модели машинного обучения, так как от него во многом и зависят конечные химические свойства белков. Фолдинг, в свою очередь, зависит от многих факторов, одним из которых является взаимодействие цепи аминокислот с ионами двухвалентных металлов, стабилизирующих формирующуюся структуру белка [5]. Для создания новых дескрипторов нами была использована база данных с результатами квантово-химических расчетов на основе теории функционала плотности для различных конформеров протеиногенных аминокислот [6]. Разработанные дескрипторы в сочетании с молекулярными дескрипторами аминокислот будут использованы в качестве входных данных для создания скрытых представлений с помощью сжатия автоэнкодерной архитектурой. Автоэнкодер тренировался на > 5.000.000 белках и 4.500.000 нуклеиновых кислотах. Полученные скрытые представления будут использованы для предсказания физико-химических свойств белковых структур с целью изучения влияния характеризующих фолдинг дескрипторов на точность моделей машинного и глубокого обучения.

Выводы. Таким образом, создание дескрипторов, характеризующих взаимодействие ферментов с ионами металлов, позволит лучше охарактеризовать процессы фолдинга белков в моделях машинного и глубокого обучения, значительно повысив их точность. Упомянутая выше автоэнкодерная архитектура уже реализована и обучается до среднеквадратичной ошибки восстановления нормализованных в диапазоне $[-1;1]$ значений, равной 0.005. Важно отметить что работа не ограничивается учетом взаимодействия ионов металлов с белками в прогнозирующих моделях машинного обучения, и общей целью проекта является учет белковых структур во всей их сложности, что приведет к появлению точных моделей, прогнозирующих физико-химические свойства и биологические активности белковых структур и генерирующих новые структуры с заранее заданными свойствами.

Список использованных источников:

1. Wang, L., Wang, N., Zhang, W. *et al.* Therapeutic peptides: current applications and future directions. *Sig Transduct Target Ther* **7**, 48 (2022).
2. J. Emonts, J.F. Buyel. An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling, *Comput. Struct. Biotechnol. J*, **2023**, V. 21, P. 3234-3247.
3. Jorner, K., Tomberg, A., Bauer, C. *et al.* Organic reactivity from mechanism to machine learning. *Nat Rev Chem* **5**, 240–255 (2021).
4. Byungju Lee, Jaekyun Yoo and Kisuk Kang. Predicting the chemical reactivity of organic materials using a machine-learning approach. *Chem. Sci.*, 2020, **11**, 7813-7822.
5. Héctor López-Laguna, et. al. Divalent cations: a molecular glue for protein materials. *Trends in Biochemical Sciences*. **2020**, V. 45, Iss. 11, P. 992 - 1003.
6. Hu, X., Lenz-Himmer, MO. & Baldauf, C. Better force fields start with better data: A data set of cation dipeptide interactions. *Sci Data*. **2022**. V. 9, Iss. 327.