УДК 004.85

ИССЛЕДОВАНИЕ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ ТРЕТИЧНОЙ СТРУКТУРЫ БЕЛКА

Артемьев А.Д. (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент Кугаевских А.В.
(Университет ИТМО)

Введение. Проблема предсказания структуры белка занимает важное место в молекулярной биологии. Существующие методы глубокого обучения, основанные на анализе аминокислотных последовательностей, сталкиваются с трудностями при работе с орфанными и синтетическими белками из-за отсутствия аналогов в базах данных. Методы, использующие большие языковые модели для кодирования начальной последовательности белка, обещают увеличение точности и скорости работы, что требует детального анализа их эффективности на различных типах белков [1, 3].

Основная часть. В исследовании были рассмотрены три современные модели глубокого обучения: ESMFold [1], AlphaFold2 в виде ColabFold [2], и RGN2 [3], с акцентом на их способность к предсказанию структуры белка без привлечения информации о гомологах. Были выбраны две группы белков для сравнительного анализа: орфанные белки (orphan), не имеющие близких гомологов и характеризующиеся минимальной глубиной MSA, и синтетические белки (de novo), отмеченные в базе данных PDB как 'synthetic construct'. В качестве объектов исследования выбраны орфанные белки 6A3A_D, 6F0F_B, 7AL0_A и синтетические белки 6WRW_C, 6XNS_B, 6XH5_B, которые представляют разную длину последовательностей и структурные особенности. Оценка моделей проводилась по метрикам RMSD и TM-score, что позволило выполнить количественный и визуальный анализ предсказаний моделей.

Выводы. Результаты исследования подтверждают, что ESMFold демонстрирует наилучшую производительность при предсказании структур орфанных и синтетических белков, опережая другие модели по точности и скорости. ColabFold также показал хорошие результаты, но его производительность оказалась несколько ниже по сравнению с ESMFold, особенно в случаях отсутствия гомологов. Модель RGN2 не смогла показать ожидаемую эффективность, особенно при анализе белков с альфа-спиралями. Эти выводы подчеркивают важность выбора соответствующей модели для анализа конкретных типов белков и могут служить отправной точкой для будущих исследований в области генеративного дизайна белков и детального анализа их структур, открывая новые возможности для научного сообщества в понимании и моделировании белковых структур.

Список использованных источников:

1. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. d. S., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. doi: https://doi.org/10.1101/2022.07.20.500902
2. Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589. doi: https://doi.org/10.1038/s41586-021-03819-2
3. Chowdhury, R., Bouatta, N., Biswas, S., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. Nature Biotechnology, 40, 1617–1623. doi: https://doi.org/10.1038/s41587-022-01432-w