

Отчёт о проделанной работе с использованием оборудования ИВЦ НГУ

Тема: Предсказание ЕС-номера фермента на основе методов искусственного интеллекта с учётом оперонной структуры геномов прокариот.

Аннотация

Разработан и оценён метод верификации предсказаний классификационных номеров ферментов (ЕС) с помощью оперонного контекста. Поверх предсказаний двух моделей глубокого обучения — ProteInfer и CLEAN — применяется оперонный фильтр: предсказанный ЕС-номер принимается, если он функционально согласуется с генами-соседями по оперону (общий метаболический путь KEGG либо совпадение ЕС-номера). На основе данных ODB, NCBI RefSeq, UniProt и KEGG построен оценочный набор из 51890 белков (30977 оперонов, 2095 организмов); стратификация по гомологии к обучающей выборке выполнена с помощью MMseqs2. Показано, что оперонный фильтр поверх ProteInfer (порог $\tau = 0,90$) повышает F1 macro до $\approx 0,887$ при покрытии 0,716, причём наибольший прирост достигается в страте низкой гомологии.

Состав коллектива

- Карманов Сергей Юрьевич — магистрант, факультет естественных наук (ФЕН) НГУ; лаборатория искусственного интеллекта и больших геномных данных, ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН».
- Деменков Павел Сергеевич — научный руководитель, кандидат технических наук, научный сотрудник, ИЦиГ СО РАН.

Информация по гранту

Работа выполнена без грантовой поддержки.

Научное содержание работы

Постановка задачи. Модели глубокого обучения (ProteInfer, CLEAN) предсказывают ЕС-номера ферментов по аминокислотной последовательности, но дают уверенные ошибки, особенно на белках с низкой гомологией к обучающей выборке. Цель работы — проверить, может ли функциональный контекст оперона служить независимым критерием верификации таких предсказаний и повышать их надёжность.

Описание работы. Построен полный конвейер подготовки данных и оценки: консервативные опероны (ODB) сопоставлялись с геномными аннотациями NCBI RefSeq, последовательности и ЕС-аннотации брались из UniProt, метаболические пути — из KEGG. На объединённой обучающей выборке моделей с помощью MMseqs2 проводился попарный поиск гомологии, по которому оценочные белки разбивались на страты (HIGH > 70 %, MEDIUM 40–70 %, LOW < 40 %, NONE). Предсказания ProteInfer и CLEAN оценивались до и после применения оперонного фильтра по уровням ЕС-иерархии и стратам гомологии; порог уверенности τ выбирался по максимуму F1 macro после фильтра. Использованное программное обеспечение: Python (pandas, numpy, pyarrow, scikit-learn, matplotlib), MMseqs2, модели ProteInfer и CLEAN.

Полученные результаты

Построен оценочный набор из 51890 белков, 30 77 оперонов и 2095 организмов. Оперонный фильтр поверх ProteInfer при $\tau = 0,90$ повышает F1 macro до $\approx 0,887$ при покрытии 0,716 и устойчиво превосходит модель CLEAN; наибольший прирост качества фильтр даёт в страте низкой гомологии (LOW), где модели ошибаются чаще. Подтверждена биологическая основа метода: у подавляющего большинства оперонов гены-соседи участвуют в общих метаболических путях KEGG. Результаты вошли в выпускную квалификационную работу магистра, защищённую 11.06.2026.

Эффект от использования кластера ИВЦ НГУ

На ресурсах ИВЦ НГУ выполнены наиболее вычислительно ёмкие этапы работы: попарный поиск гомологии MMseqs2 по обучающей выборке из сотен тысяч последовательностей для стратификации оценочного набора, а также прогон и оценка моделей ProteInfer и CLEAN. Использовались дисковое хранилище /mnt/storage, рабочая область /mnt/scratch и планировщик задач PBS Pro.

Публикации

1. Выпускная квалификационная работа магистра «Предсказание ЕС-номера фермента на основе методов искусственного интеллекта с учётом оперонной структуры геномов прокариот» (защищена 11.06.2026), НГУ.