УДК 004.89

РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ДЛЯ РАНЖИРОВАНИЯ ПОТЕНЦИАЛЬНЫХ БИОМОЛЕКУЛЯРНЫХ ВЗАИМОДЕЙСТВИЙ НА ОСНОВЕ ГЕТЕРОГЕННОГО ГРАФА ЗНАНИЙ

Богданов П. И.[1], Еремеева М. А.[1]
Научный руководитель – канд. хим. наук, доцент Серов Н. С.[1]
[1]Университет ИТМО
pascal1@bk.ru

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

Введение

Разработка биосенсорных систем требует подбора распознающих элементов разной химической природы способных селективно связываться с целевыми аналитами. Экспериментальные методы поиска взаимодействий трудоемки и ограничены масштабируемостью [1]. В условиях роста объема данных актуальной задачей становится выявление и приоритизация взаимодействий с учетом контекстуальной связности. Граф знаний обеспечивает масштабируемую структуру для интеграции разнородных сущностей и отношений, позволяя выявлять латентные зависимости и обосновывать новые связи, не наблюдавшихся в существующих базах данных.

Основная часть

В рамках работы построен гетерогенный граф знаний в Neo4j, интегрирующий данные из открытых источников. Граф включает в себя ~1.4M сущностей (пептиды, белки, ДНК, РНК, малые молекулы) и порядка ~6M связей двух типов:

- interacts with – эмпирически подтвержденные взаимодействия;
- has similarity – вычисленное сходство структур или последовательностей.

С использованием построенного графа решена задача прогнозирования отсутствующих связей типа interacts with:

1) Обучение в трансдуктивной постановке: множество узлов фиксировано, а часть ребер скрывалась для тестирования при обязательном присутствии всех сущностей в обучающей выборке.
2) Модели KGE обучены для задачи прогнозирования связи. Скрытые части известных связей использовались для валидации модели в трансдуктивном режиме. Оценка модели проводилась на двух связях (MRR = 0,67, Hits@10 = 0,88), а эмбеддинги применялись для ранжирования ненаблюдаемых пар сущностей с целью приоритизации потенциальных взаимодействий.

Дополнительно выполнена связь-специфичная оценка для связи interacts with на прикладных поднаборах (антибиотики/сигнальные белки/белки, ассоциированные с онкологическими процессами). Во всех случаях модель демонстрирует устойчивое извлечение релевантных взаимодействий в верхней части ранжированного списка (NDCG@10=0,92, NDCG@10=0,73, NDCG@10=0,80 соответственно).

Выводы

Разработан масштабируемый гетерогенный граф знаний, сочетающий в себе информацию о разных типах сущностей для поиска потенциально новых распознающих элементов и модель ранжирования на основе KGE. Показано, что в условиях разреженного графа знаний предложенный подход позволяет выявлять отсутствующие

связи и приоритизировать пары распознающий элемент-аналит, сокращая экспериментальный поиск.

Литература

1. Rettie S.A., Juergens D., и др. Accurate de novo design of high-affinity protein-binding macrocycles using deep learning // Nature Chemical Biology. 2025. P. 1–9.
2. Lerer A., Wu L., Shen J. и др. PyTorch-BigGraph: A large-scale graph embedding system // Proceedings of the 2nd Conference on Systems and Machine Learning. — 2019. — P. 1–16.