Построение бинарной семантической сети межлекарственных взаимодействий на основе поисковых запросов

Степанов Григорий Александрович

Лаборатория молекулярного моделирования и направленного синтеза № 44

Актуальность. Оценка возможности возникновения межлекарственных взаимодействий (далее DDI) при совместном применении нескольких лекарственных препаратов является одной из ключевых задач клинической фармакологии. Ныне существующие базы данных DDI основаны на инструкциях по применению лекарств, которые зачастую не обновляются долгое время и не содержат всей доступной сейчас информации. Большая часть DDI связана с белками семейства цитохромов, которые участвуют в метаболизме чужеродных веществ.

Подход к оценке DDI, предложенный в данном исследовании, выгодно отличается простотой реализации и исполнения алгоритма, а также возможностью построения модели на основе минимального набора данных и базовых знаний в предметной области. В его основе лежит предположения о том, что упоминания факта или свойства напрямую зависит от «практической важности» данного факта или свойства, и что вероятность совместного упоминания двух объектов зависит от «взаимосвязи» данных объектов.

На основе предположений можно ожидать, что количество ответов на поисковые запросы, содержащие названия лекарства и белка, будет тем больше, чем шире изучено их взаимодействие.

Цель. На основе количества ответов на запросы разной структуры выявить DDI (и доступность данных о DDI для конкретных лекарств).

Этапы работы. Для подтверждения возможности делать выводы исходя из количества ответов был проведен поиск синонимов названий лекарств. Изначально были взяты 4 источника с разной степенью курирования (от автоматически создаваемых баз данных до курированных специалистами NCATS NIH). После чего были сделаны поисковые запросы (для реализации потоковых запросов был применен метод подключения через sock5-port), включающие в себя:

- Один из синонимов для каждого лекарства,
- Два синонима одного лекарства,
- Названия двух разных лекарств.

С помощью алгоритма машинного обучения (неглубокое решающее дерево) была достигнута точность 93% в определении, является ли конкретное слово синонимом лекарства. В дальнейшем точность была повышена более, чем на 3% при добавлении большего количества рандомизированных пар (названий двух разных лекарств).

Для дальнейшей работы понадобилась база данных, включающая в себя все известные DDI для подтвержденных FDA лекарственных средств. В связи с этим была собрана самая большая база DDI, включающая в себя информацию из KEGG, DrugBank, DrugInteraction-Medicine, ChEMBL и NCATS.

Для построения предсказательной модели взаимодействия лекарства с цитохромом были сделаны запросы, включающие в себя названия лекарства и белка автономно, совместно, а также с упоминанием ключевых слов (например, мидазолам, который широко применяется для изучения взаимодействий с P450 CYP3A4).

На основе собранных данных была построена модель машинного обучения SVM, которая на основании двух запросов была способна предсказать белок-лекарственное взаимодействие с точностью 70% (Рис. 1).

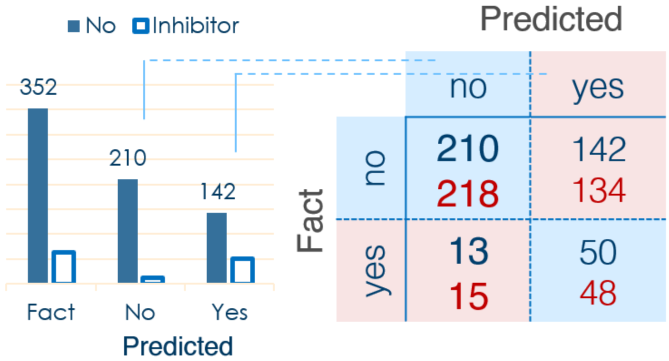

Рис. 1. Результат предсказания взаимодействия на основе модели SVM (числа с синим выделением) и на основе логистической регрессии (красное выделение)

Следующей стадией является построение предсказательной модели на основе метода Random Forest, который позволит улучшить результаты. Так же начата работа над алгоритмом подтверждения клинической значимости каждого взаимодействия на основе данных ClinicalTrials.gov, что позволит улучшить качество входных данных для дальнейшего построения предсказательных моделей

Степанов Григорий Александрович (подпись)

Новиков Федор Николаевич, к.х.н. (подпись)

Свитанько Игорь Валентинович, зав. лаб., д.х.н. (подпись)