Р.Г. Ефремов: Как учитывали (и, вероятно, корректировали?) «межлабораторные ошибки» при подготовке выборки активных соединений (Слайд 5)?

Л.А. Столбов: В данном вопросе ориентировались на опыт представленный в публикациях, в частности Fourches,D., et al. (2016) Trust, but verify II: A practical guide to chemogenomics Data Curation, J Chem Inf Model 56, 1243.

Как отмечено в докладе процесс предобработки данных сопряжен с удалением дубликатов и, несмотря на то, что эта информация теряется в конечной обучающей выборке, она может быть использована для оценки вариабельности данных в различных исследованиях.

В указанных выборках в качестве дубликатов часто выступали хорошо изученные соединения, одобренные FDA в качестве ингибиторов ВИЧ. Для экспериментальных значений дубликатов рассчитывался межквартильный размах - IQR (для каждой структуры, имеющий дублированные записи активности), в качестве факторов принимали экспериментальный протокол. Для каждого исследования смотрели попадают ли активности входящих в исследование дубликатов в IQR; если нет, искали возможную объективную причину в литературе (например, исследование активности реакционной смеси сразу после реакции, без предварительной очистки). Данные такого исследования исключались из обучающей выборки, за исключением случаев, когда все оставшиеся структуры не имели схожих в остальной выборке (при оценке по сходству по MNA и QNA дескрипторам, упомянутым в презентации)

Коррекция в данном случае не проводилась.

Р.Г. Ефремов: Проводили ли анализ полученных выборок соединений, например, по следующим параметрам: представленность различных классов химических соединений и экспериментальных протоколов определения активности, степень вырожденности и пр.? Сравнение результатов тестирования на внешней выборке (слайд 9) и на всех данных (слайд 10), по-видимому, говорят о проблемах с созданными наборами данных.

Л.А. Столбов:

Классы химических соединений:

Ввиду наличия нескольких функциональных групп в практически каждом соединении, исследованном на анти-ВИЧ активность, классификация соединений в обучающей выборке по химическим классам не проводилась. Тем не менее, для создания сбалансированных выборок было произведено разбиение на группы по сходству (также с использованием MNA и QNA дескрипторов)

Экспериментальные протоколы:

Измерения активностей были представлены 3 основными группами - данные получены с помощью масс-спектрометрического анализа, флуориметрического, другое/не указан. Для различных мишеней для представленных физико-химических методов в некоторых случаях группу разбивали в зависимости от экспериментального протокола (например, с присутствием ионов Mg для ингибиторов интегразы или без них).

Так в некоторых случаях объединить данные разных протоколов не удалось (3' и Strand transfer протоколы для интегразы).

Для протеазы наоборот, модели построенные на двух разных выборках, разделенных по физико-химическому методу показывали высокий коэффициент детерминации при взаимном прогнозировании, для дубликатов между двумя выборками также наблюдалась корреляция ($R > 0.9$). При этом возможная линейная коррекция из одних данных в другие.

Тестирование на внешней выборке выявило определенные сложности. Для этого проводилась оценка новых дескрипторов среди соединений в данных Integrity (для которых наблюдался низкий коэффициент детерминации). В объединенных выборках (слайд 10) после валидации были получены приемлемые коэффициенты детерминации 0.75, 0.77, 0.65 соответственно для моделей ингибиторов интегразы, протеазы и обратной транскриптазы. В случае интегразы и протеазы результаты валидации были допустимыми уже на данных ChEMBL. С обратной транскриптазой же ситуация оказалась особенная: для данного фермента существуют ингибиторы с разным механизмом действия (нуклеотидные ингибиторы полимеразы, ненуклеотидные ингибиторы полимеразы и ненуклеотидные ингибиторы РНАзы Н), что повлияло на предсказательную

способность модели в худшую сторону. Наилучшим вариантом в перспективе будет их разделение, однако на данный момент это сопряжено с рядом сложностей (например, отсутствие в большинстве публикаций данных о конкретной субъединице обратной транскриптазы к которой измерена активность).

Р.Г. Ефремов: Есть ли возможность создания на основании полученных выборок лиганд-специфичных QSAR-моделей (желательно с использованием экспериментальных данных, полученных в сходных условиях)? Потенциально это может повысить надежность предсказания.
Л.А. Столбов: Если я правильно понимаю вопрос, то да, при отборе соединений по сходству, субграфу или иным способом из каждой полученной обучающей выборки можно построить более специфичную модель, подходящую для одного "химического класса" с более высокой надежностью предсказания. Однако стоит принять во внимание тот факт, что область применимости модели существенно сократится, что может противоречить плану виртуального скрининга.

Р.Г. Ефремов: Чем отличаются QSAR-модели, результаты применения которых представлены на слайдах 9 и 10? Можно качественно охарактеризовать основные различия?
Л.А. Столбов: Модели на слайде 10 построены по обобщенным данным и используют наибольшее число дескрипторов и, таким образом, имеют большее число степеней свободы по сравнению с моделями на слайде 9. Качественным же различием является то, что для моделей, которые построены по обобщенным данным существуют соединения, которые находятся в их области применимости и при этом не попадающие в область применимости моделей на неполных данных, но не наоборот.