УДК 519.23+615.2+616.98

ВЗАИМОСВЯЗИ «СТРУКТУРА-АКТИВНОСТЬ» ДЛЯ ИНГИБИТОРОВ ОБРАТНОЙ ТРАНСКРИПТАЗЫ ВИЧ-1: КАК ПОВЫСИТЬ ТОЧНОСТЬ И ПРЕДСКАЗАТЕЛЬНУЮ СПОСОБНОСТЬ ПОЛУЧАЕМЫХ МОДЕЛЕЙ?[1]

О.А.Тарасова, Д.А.Филимонов, В.В.Поройков
Научно-исследовательский институт биомедицинской химии имени В.Н.Ореховича, Москва, Россия

STRUCTURE-ACTIVITY RELATIONSHIPS OF HIV-1 REVERSE TRANSCRIPTASE INHIBITORS: HOW TO INCREASE THE ACCURACY AND PREDICTABILITY OF MODELS?

O.A.Tarasova, D.A.Filimonov, V.V.Poroikov
Institute of Biomedical Chemistry, Moscow, Russia

Целью настоящего исследования является оценка влияния вариабельности данных на качество моделей структурного системного анализа (ССА) и разработка подходов, позволяющих повысить точность и предсказательную способность этих моделей. Материалы и методы. Моделирование взаимосвязей «структура-активность» для ингибиторов обратной транскриптазы ВИЧ-1 производилось с использованием баз данных биологически активных соединений. Компьютерный анализ и моделирование взаимосвязей между структурой и биологической активностью химических соединений позволяет предсказать активность для не исследованных экспериментально веществ, в том числе для тех, которые только планируется синтезировать, а базы данных биологически-активных соединений и научные публикации являются существенно важным источником для создания обучающих выборок при построении моделей ССА. Существует значительный разброс количественных значений активности (IC_{50} и K_i), полученных для одних и тех же соединений, в особенности, если их измерение было проведено в различных лабораториях, что является причиной значительной вариабельности значений в базах данных. Результаты. Исследована применимость свободно и коммерчески доступных баз данных для получения точных и предсказательных моделей ССА с антиретровирусной активностью (ингибирование обратной транскриптазы ВИЧ-1). Определено, что точность моделей ССА зависит от способа построения обучающей выборки, а также выявлены определенные ограничения баз данных для создания обучающих выборок, содержащих соединения, испытанные в максимально схожих условиях эксперимента. Таким образом, необходима разработка метода отбора низкомолекулярных соединений с требуемой биологической активностью, протестированных в схожих биологических условиях (по данным, содержащимся в научных публикациях), для последующего построения моделей ССА. Такой метод позволит получать выборки низкомолекулярных соединений с меньшей вариабельностью в количественных данных об их биологической активности для последующего построения наиболее высокоточных моделей ССА. В свою очередь данные модели ССА могут быть в дальнейшем применены для дизайна новых препаратов с антиретровирусной активностью.

Ключевые слова: ВИЧ-1, базы данных биологически активных веществ, компьютерное моделирование, взаимосвязь между структурой и активностью, вариабельность данных.

Study objective was to assess the effect of data inconsistency on the quality of structural systemic analysis (SSA) models of reverse transcriptase inhibitors and to develop approaches to increasing the accuracy and predictive power of such models. Materials and methods: Structure-activity relationships in HIV-1 reverse transcriptase inhibitors were modeled using biologically active compound databases. Computer-assisted analysis and modeling of relationships between the structures and biological activities of chemical compounds allows predicting the activity of substances that were not studied experimentally, including even those only intended to be synthesized. Databases of biologically active compounds are extremely valuable sources for the selection of samples used to train SSA models. The quantitative characteristics of activity (IC_{50} and Ki) of a particular compound may be highly variable, especially if they have been determined at different laboratories, and this may cause

[1] Доложено на мероприятии Школы молодых ученых: «ВИЧ-обусловленные иммуносупрессии и их последствия» 13–15 апреля 2016 года.

considerable discrepancies in databases. Results: The adequacy of freely accessed and commercial databases to developing accurate and predictive SSA models for substances having an antiretroviral activity (HIV-1 reverse transcriptase inhibition) was assessed. The accuracy of SSA models was found to depend on the procedures of construction of a training sample. Certain limitation were found in databases used to construct training samples when sample entries have been testes under almost identical conditions. Therefore, it is expedient to develop a method for selecting low molecular weight compounds featuring low variability of their quantitative characteristics for being used in training of SSA models. Such method would increase the accuracy of SSA models employed in the design of novel antiretroviral substances.

Key words: HIV-1, databases of biologically active compounds, computer-assisted structure-activity relationships analysis, data inconsistency.

Введение. В настоящее время миллионы людей заражены вирусом иммунодефицита человека первого типа (ВИЧ-1), который является причиной возникновения синдрома приобретенного иммунодефицита человека (СПИД). Несмотря на то, что существуют современные схемы лечения, потребность в новых высокоэффективных препаратах против ВИЧ/СПИД сохраняется, и поиск новых более эффективных и безопасных антиретровирусных препаратов для терапии ВИЧ/СПИД остается актуальным [1], однако применение с этой целью экспериментальных методов сопряжено со значительными финансовыми и временными затратами. Компьютерный анализ и моделирование взаимосвязей между структурой и биологической активностью (ССА) химических соединений позволяет предсказать активность для не исследованных экспериментально веществ, в том числе для тех, которые только планируется синтезировать, а также позволяет проводить мультикритериальный отбор (учет биологической активности, побочных эффектов, токсичности и т. д.) на этапе разработки перспективных соединений.

Для построения моделей ССА обычно используют обучающие выборки, содержащие информацию о структуре химических соединений, представленной в двумерной или трехмерной форме, а также известные для этих соединений количественные показатели биологической активности, характеризующие связывание с конкретной фармакологической мишенью (например, константа ингибирования K_i или полуингибирующая концентрация IC_{50}).

Базы данных (БД) низкомолекулярных биологически активных соединений и научные публикации являются важным источником информации для построения моделей ССА. Такие БД содержат информацию, собранную из огромного количества различных публикаций. Показано, что существует значительный разброс количественных значений IC_{50} и K_i, полученных для одних и тех же соединений, в особенности, если их измерение было проведено в различных лабораториях, что является причиной значительной вариабельности значений в БД [2]. Таким образом, использование информации из БД без ее предварительной фильтрации не позволяет получить достаточно точные и обладающие предсказательной способностью модели ССА [2, 3].

Целью настоящего исследования является оценка влияния вариабельности данных на качество моделей ССА и разработка подходов, позволяющих повысить точность и предсказательную способность этих моделей.

Материалы и методы. При моделировании взаимосвязей «структура-активность» для ингибиторов обратной транскриптазы (ОТ) ВИЧ-1 с использованием баз данных биологически активных соединений были использованы выборки ингибиторов ОТ ВИЧ-1, извлеченные из коммерчески доступной БД Thomson Reuters Integrity и свободно доступной БД ChEMBL 19. Далее произведено разделение полученных выборок в соответствии со следующими критериями:

— отбор всех соединений, протестированных в отношении конкретного количественного показателя биологической активности химического соединения (а именно, IC_{50});

— отбор соединений, протестированных в схожих экспериментальных условиях (единый материал и метод исследования, согласно информации из соответствующей БД);

— отбор соединений, данные о биологическом тестировании которых извлечены из одной публикации. Модели ССА были построены в компьютерной программе GUSAR [4, 5].

Точность моделей была оценена методом скользящего контроля с исключением 30% соединений. Оценка предсказательной способности моделей

производилась методом пятикратной кросс-валидации [2].

БД Integrity. Точность моделей ССА была низкой, если для обучения использовали выборки, сформированные в соответствии с первым критерием отбора соединений (R^2=0,54, R^2_{LMO}=0,28, R^2_k=0,24, где R^2 — коэффициент детерминации между экспериментально измеренными величинами активности (IC_{50}) и спрогнозированными величинами активности; R^2_{LMO} — коэффициент детерминации при скользящем контроле с исключением 30% соединений; R^2_k — коэффициент детерминации, рассчитанный в результате пятикратной кросс-валидации).

Удаление соединений, для которых в БД не содержалось информации о конкретном материале и методе исследования, позволило добиться улучшения качества моделей (R^2=0,99, R^2_{LMO}=0,60, R^2_k=0,50). Таким образом, для ингибиторов ОТ исключение из обучающей выборки соединений, для которых нет информации о материале и методе исследования, приводит к повышению качества моделей ССА по сравнению с использованием данных из БД Integrity без дополнительной фильтрации. Разделение на выборки в соответствии с определенным материалом и методом исследования, если в качестве материала были использованы клеточные линии, инокулированные ВИЧ-1, также приводит к повышению качества моделей (для лучшей модели: R^2=0,85, R^2_{LMO}=0,76, R^2_k=0,64; материал экспериментального исследования: мононуклеарные клетки крови человека, метод исследования — реакция «антиген-антитело»). Схожие результаты были получены для случая разделения обучающих выборок в соответствии с методиками, в которых материалом исследования был выделенный и очищенный фермент — обратная транскриптаза ВИЧ-1 (R^2=0,99, R^2_{LMO}=0,60, R^2_k=0,58, см. также [2]). Формирование выборок согласно третьему критерию не было возможно для информации из БД Integrity, поскольку количество структур в обучающих выборках в этом случае не превышало пяти, что недостаточно для построения предсказательных моделей ССА.

БД ChEMBL 19. В целом качество моделей, построенных с применением обучающих выборок из БД ChEMBL 19, не отличается от такового для данных из БД Integrity при использовании первого и второго критерив формирования обучающих выборок (до разделения на материалы и методы исследования: R^2=0,56, R^2_{LMO}=0,54, R^2_k=0,44). Этот факт можно объяснить недостаточной детализированностью при аннотировании данных в соответствии с определенным материалом и методом исследования в БД ChEMBL 19. Однако аннотирование данных в соответствии с конкретной публикацией в БД ChEMBL 19 позволило сформировать выборки согласно третьему критерию, и точность моделей, построенных с использованием этих обучающих выборок, была существенно выше, чем точность моделей, построенных с использованием выборок, сформированных в соответствии с первым и вторым критериями (для лучшей модели: R^2=0,64, R^2_{LMO}=0,64, R^2_k=0,62).

Формирование обучающих выборок с меньшей вариабельностью в количественных данных об их биологической активности. Основываясь на выше изложенных результатах анализа для ингибиторов ОТ ВИЧ-1, был разработан протокол отбора низкомолекулярных соединений с требуемой биологической активностью, протестированных в схожих биологических условиях, для последующего построения моделей ССА (рисунок).

Использованы инструменты анализа текста, реализованные в среде разработки KNIME [6], с целью извлечения ключевых слов, характеризующих условия биологического эксперимента из текстов статей, идентифицируемых по регистрационным номерам PMID в БД PubMed. Далее произведена оценка сходства условий эксперимента, например, путем расчета доли совпадения ключевых слов среди усредненного числа ключевых слов, выбранных из двух рассматриваемых публикаций. На основе сопоставления долей совпадающих ключевых слов осуществлялся отбор однородных публикаций, в которых содержатся данные о биологически активных соединениях, протестированных в схожих условиях эксперимента. Предварительное тестирование разработанного протокола позволило идентифицировать две пары публикаций (PMID: 9240358 и 11472208; 7538590 и 16107158 — см. публикации [7–10]), в которых условия эксперимента схожи по применяемому материалу на основе анализа 37 статей, выбранных из БД PubMed в результате запросов по ключевым словам.

Заключение. На основе представленных результатов рекомендовано два возможных пути увеличения согласованности данных, которые можно использовать для построения моделей ССА. Во-первых, при подготовке научных публикаций необходимо приводить детализированные формализованные протоколы описания методик тестиро-

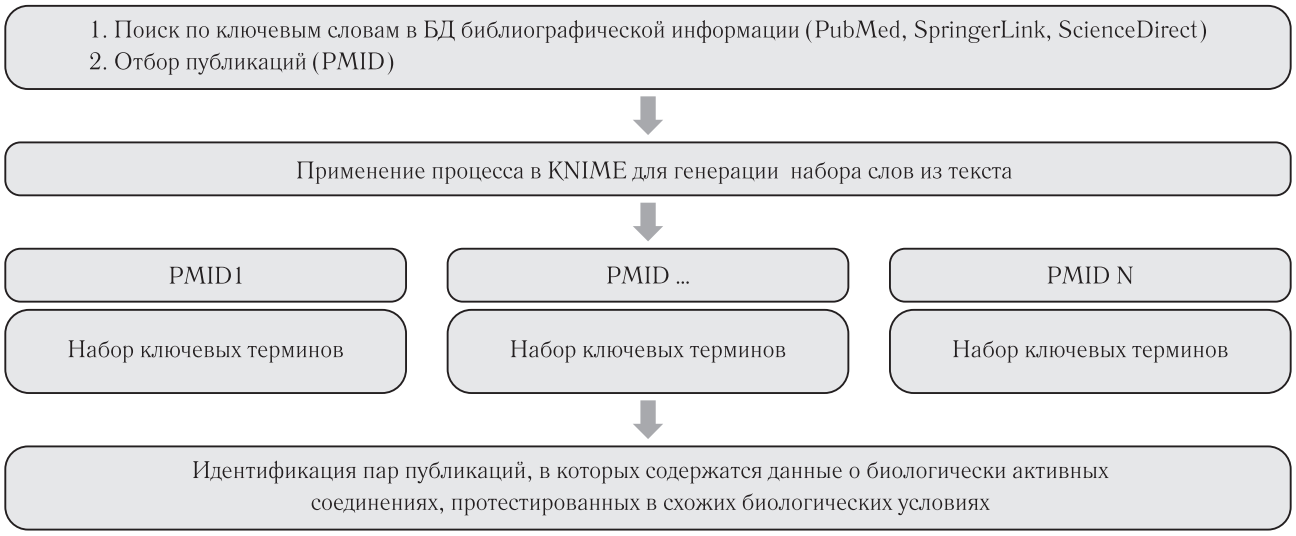

Рисунок. Протокол отбора публикаций, содержащих данные о биологически активных соединениях, протестированных в схожих условиях.

вания биологически активных соединений [11], которые могут быть использованы для сопоставления условий эксперимента. Во-вторых, для увеличения практической применимости моделей ССА они должны быть построены с использованием обучающих выборок, содержащих информацию о соединениях, которые протестированы в определенных максимально близких условиях эксперимента, что позволит получить достаточно точные и обладающие высокой предсказательной способностью модели ССА.

Благодарности:

Работа выполнена при поддержке гранта РФФИ № 16-34-60187.

ЛИТЕРАТУРА

1. *Guasch L., Zakharov A., Tarasova O., Poroikov V.V, Liao C., Nicklaus M.C.* Novel HIV-1 integrase inhibitor development by virtual screening based on QSAR models // Curr. Top. Med. Chem.— 2015.— Vol. 16, № 4.— P. 441–448.
2. *Kramer C., Kalliokoski T., Gedeck P., Vulpetti A.* The experimental uncertainty of heterogeneous public K(i) data // J. Med. Chem.— 2012.— Vol. 55, № 11.— P. 5165–5173.
3. *Tarasova O., Urusova A., Filimonov D., Nicklaus M.C., Zakharov A.V., Poroikov V.V.* QSAR Modeling Using Large-Scale Databases: Case Study for HIV-1 Reverse Transcriptase Inhibitors // J. Chem. Inf. Model.— 2015.— Vol. 55, № 7.— P. 1388–1399.
4. *Filimonov D., Zakharov A., Lagunin A., Poroikov V.V.* QNA based «Star Track» QSAR approach // SAR and QSAR Environ. Res.— 2009.— Vol. 20, № 7–8.— P. 679–709.
5. *Zakharov A., Peach M., Sitzmann M., Nicklaus M.C.* A new approach to radial basis function approximation and its application to QSAR // J. Chem. Inf. Model.— 2014.— Vol. 54, № 3.— P. 713–719.
6. *Berthold M., Cebron N., Dill F., Gabriel T., Kötter T., Meinl T., Ohl P., Sieb C., Thiel K., Wieswedel K.* The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg, 2007.
7. *Kelly T., Proudfoot J., McNeil D., Patel U., David E., Hargrave K., Grob P., Cardozo M., Agarwal A., Adams J.* Novel non-nucleoside inhibitors of human immunodeficiency virus type 1 reverse transcriptase. 6.2-Indol-3-yl- and 2-azaindol-3-yl-dipyridodiazepinones // J. Med. Chem.— 1997.— Vol. 40, № 15.— P. 2430–2433.
8. *Mai A., Sbardella G., Artico M., Massa S., Novellino E., Greco G., Lavecchia A.* Structure-based design, synthesis, and biological evaluation of conformationally restricted novel 2-alkylthio-6-[1-(2,6-difluorophenyl)alkyl]-3,4-dihydro-5-alkylpyrimidin-4(3H)-ones as non-nucleoside inhibitors of HIV-1 reverse transcriptase // J. Med. Chem.— 2001.— Vol. 44, № 16.— P. 2544–2554.
9. *Wyatt P., Bethell R., Cammack N., Charon D., Dodic N., Dumaitre B., Evans D., Green D., Hopewell P., Humber D., Lamont R., Orr D., Plested S., Ryan M., Sollis S., Storer R., Weingarten G.* Benzophenone derivatives: a novel series of potent and selective inhibitors of human immunodeficiency virus type 1 reverse transcriptase // J. Med. Chem.— 1995.— Vol. 38, № 10.— P. 1657–1665.
10. *O'Meara J., Yoakim C., Bonneau P., Bos M., Cordingley M., Deziel R., Doyon L.* Novel 8-substituted dipyridodiazepinone inhibitors with a broad-spectrum of activity against HIV-1 strains resistant to non-nucleoside reverse transcriptase inhibitors // J. Med. Chem.— 2005.— Vol. 48, № 17.— P. 5580–5588.

11. *Orchard S., Al-Lazikani B., Bryant S., Clark D., Calder E.* Minimum information about a bioactive entity (MIABE) // Nat. Rev. Drug. Discov.— 2011.— Vol. 10, № 9.— P. 661–669.

References

1. Guasch L., Zakharov A., Tarasova O., Poroikov V.V, Liao C., Nicklaus M.C. Novel HIV-1 integrase inhibitor development by virtual screening based on QSAR models, *Curr. Top. Med. Chem.*, 2015, vol. 16, No 4, pp. 441–448.
2. Kramer C., Kalliokoski T., Gedeck P., Vulpetti A. The experimental uncertainty of heterogeneous public K(i) data, *J. Med. Chem.*, 2012, vol. 55, No. 11, pp. 5165–5173.
3. Tarasova O., Urusova A., Filimonov D., Nicklaus M.C., Zakharov A.V., Poroikov V.V. QSAR Modeling Using Large-Scale Databases: Case Study for HIV-1 Reverse Transcriptase Inhibitors, *J. Chem. Inf. Model.*, 2015, vol. 55, No. 7, pp. 1388–1399.
4. Filimonov D., Zakharov A., Lagunin A., Poroikov V.V. QNA based «Star Track» QSAR approach, *SAR and QSAR Environ. Res.*, 2009, vol. 20, No. 7–8, pp. 679–709.
5. Zakharov A., Peach M., Sitzmann M., Nicklaus M.C. A new approach to radial basis function approximation and its application to QSAR, *J. Chem. Inf. Model.*, 2014, vol. 54, No. 3, pp. 713–719.
6. Berthold M., Cebron N., Dill F., Dill F., Gabriel T., Kötter T., Meinl T., Ohl P., Sieb C., Thiel K., Wieswedel K., In: *Studies in Classification, Data Analysis, and Knowledge Organization,* Springer: Heidelberg, 2007.
7. Kelly T., Proudfoot J., McNeil D., Patel U., David E., Hargrave K., Grob P., Cardozo M., Agarwal A., Adams J. Novel non-nucleoside inhibitors of human immunodeficiency virus type 1 reverse transcriptase. 6.2-Indol-3-yl- and 2-azaindol-3-yl-dipyridodiazepinones, *J. Med. Chem.*, 1997, vol. 40, No. 15, pp. 2430–2433.
8. Mai A., Sbardella G., Artico M., Massa S., Novellino E., Greco G., Lavecchia A. Structure-based design, synthesis, and biological evaluation of conformationally restricted novel 2-alkylthio-6-[1-(2,6-difluorophenyl)alkyl]-3,4-dihydro-5-alkylpyrimidin-4(3H)-ones as non-nucleoside inhibitors of HIV-1 reverse transcriptase, *J. Med. Chem.*, 2001, vol. 44, No. 16, pp. 2544–2554.
9. Wyatt P., Bethell R., Cammack N., Charon D., Dodic N., Dumaitre B., Evans D., Green D., Hopewell P., Humber D., Lamont R., Orr D., Plested S., Ryan M., Sollis S., Storer R., Weingarten G. Benzophenone derivatives: a novel series of potent and selective inhibitors of human immunodeficiency virus type 1 reverse transcriptase, *J. Med. Chem.*, 1995, vol. 38, No. 10, pp. 1657–1665.
10. O'Meara J., Yoakim C., Bonneau P., Bos M., Cordingley M., Deziel R., Doyon L. Novel 8-substituted dipyridodiazepinone inhibitors with a broad-spectrum of activity against HIV-1 strains resistant to non-nucleoside reverse transcriptase inhibitors, *J. Med. Chem.*, 2005, vol. 48, No. 17, pp. 5580–5588.
11. Orchard S., Al-Lazikani B., Bryant S., Clark D., Calder E. Minimum information about a bioactive entity (MIABE), *Nat. Rev. Drug. Discov.*, 2011, vol. 10, No. 9, pp. 661–669.

Статья поступила *10.05.2016 г.*

Контактная информация: *Тарасова Ольга Александровна, e-mail: olga.a.tarasova@gmail.com*

Коллектив авторов:

Тарасова Ольга Александровна — к.б.н., н.с. лаборатории структурно-функционального конструирования лекарств Научно-исследовательского института биомедицинской химии имени В.Н.Ореховича, 119121, Москва, ул. Погодинская, 10, к. 8, +7 499 255-30-29, e-mail: olga.a.tarasova@gmail.com;

Филимонов Дмитрий Алексеевич — к.ф.-м.н., в.н.с. лаборатории структурно-функционального конструирования лекарств Научно-исследовательского института биомедицинской химии имени В.Н.Ореховича, 119121, Москва, ул. Погодинская, 10, к. 8, +7 499 255-30-29, e-mail: dmitry.filimonov@ibmc.msk.ru;

Пройков Владимир Васильевич — д.б.н., к.ф.-м.н., профессор, руководитель отдела биоинформатики, зав. лабораторией структурно-функционального конструирования лекарств Научно-исследовательского института биомедицинской химии имени В.Н.Ореховича, 119121, Москва, ул. Погодинская, 10, к. 8, +7 499 255-30-29, e-mail: vladimir.poroikov@ibmc.msk.ru.