Биомедицинская химия, 2019 том 65, вып. 2, с. 73-79.

БИОИНФОРМАТИКА И ХЕМОИНФОРМАТИКА

ПОИСК НОВЫХ АНТИРЕТРОВИРУСНЫХ СОЕДИНЕНИЙ В ХИМИЧЕСКОМ ПРОСТРАНСТВЕ “БОЛЬШИХ ДАННЫХ” БИБЛИОТЕКИ SAVI

П.И. Савосина[1]*, Л.А. Столбов[1], Д.С. Дружиловский[1], Д.А. Филимонов[1], М. Никлаус[2], В.В. Поройков[1]

[1]Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича, 119121, Москва, Погодинская ул., 10, стр. 8; *эл. почта: polinkasavosina@gmail.com

[2]Лаборатория химической биологии, Центр исследования онкологии, Национальный институт онкологии, Национальные институты здравоохранения, Фредерик, Мэриленд 21702, США (Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland, 21702 USA)

Несмотря на значительные успехи в применении высокоактивной антиретровирусной терапии, разработка новых препаратов для лечения ВИЧ-инфекции остается актуальной, поскольку существующие лекарственные средства не обеспечивают полного излечения от ВИЧ-инфекции, вызывают серьезные побочные эффекты и приводят к возникновению резистентности. В 2015 году консорциумом американских и европейских ученых и специалистов начат проект по созданию библиотеки SAVI (Synthetically Accessible Virtual Inventory), в рамках которого было сгенерировано *in silico* свыше 283 млн. структур новых легко-синтезируемых органических молекул с целью поиска среди них более безопасных и эффективных фармакологических веществ. Мы разработали алгоритм сравнения больших химических баз данных (БД) на основе представления структурных формул в формате SMILES и оценили возможности выявления в БД SAVI новых антиретровирусных соединений. Проанализировав пересечение SAVI с 97 млн. структур БД PubChem, мы обнаружили, что лишь малая часть SAVI (~0,015%) представлена в PubChem, что свидетельствует о значительной новизне этой виртуальной библиотеки. Однако, среди этих структур было выявлено 632 соединения, протестированных на анти-ВИЧ активность, из которых 41 обладали искомой активностью. Таким образом, наши исследования впервые продемонстрировали, что SAVI является перспективным источником для поиска новых анти-ВИЧ соединений.

Ключевые слова: “большие данные”; SAVI; PubChem; новые лекарственно-подобные соединения; антиретровирусная активность; прогноз PASS

DOI: 10.18097/PBMC20196502073

ВВЕДЕНИЕ

ВИЧ-инфекция остаётся одним из наиболее опасных заболеваний, поскольку число инфицированных ежегодно увеличивается на более чем 1,8 млн. человек, а смертность от причин, связанных с ВИЧ/СПИД, составляет почти 1 млн. случаев в год [1]. В 1987 году был введен в медицинскую практику Азидотимидин – первый антиретровирусный препарат для терапии ВИЧ-инфекции. Последующие клинические исследования показали, что использование монотерапии не обеспечивает устойчивой элиминации вируса. Применение комбинированной терапии замедлило прогрессирование заболевания и привело к уменьшению оппортунистических инфекций [2]. В настоящее время разрешены к медицинскому применению 36 антиретровирусных препаратов, относящихся к шести классам: нуклеозидные и ненуклеозидные ингибиторы обратной транскриптазы, ингибиторы протеазы, ингибиторы интегразы, ингибиторы слияния, антагонисты хемокиновых рецепторов [3, 4]. Современный подход к терапии ВИЧ-инфекции называют высокоактивной антиретровирусной терапией (ВААРТ), включающей сочетанное применение 2-4 препаратов из различных фармакологических классов [2].

Несмотря на значительные достижения в разработке и применении ВААРТ, остаются не решёнными важные проблемы. Одной из них является возникновение серьёзных побочных эффектов, таких как остеопороз, кардиомиопатии, нефролитиаз, нарушения свертываемости крови и др. [5], которые требуют немедленного терапевтического вмешательства. Снижению или полному отсутствию эффективности лечения способствует развитие лекарственной резистентности вируса ко всем классам анти-ВИЧ препаратов [6]. Таким образом, применение существующих лекарственных средств не обеспечивает полного излечения ВИЧ-инфицированных пациентов и сопряжено с возникновением указанных выше проблем, что свидетельствует об актуальности поиска и создания новых антиретровирусных препаратов.

Большинство работ по поиску новых анти-ВИЧ соединений базируется на изучении ограниченного химического пространства уже синтезированных лекарственно-подобных органических веществ, представленного такими базами данных (БД), как ChemNavigator [7], ZINC [8] и др. Использование для поиска новых фармакологических веществ существенно более объёмных БД, в которых представлены виртуально сгенерированные структуры молекул (например, GDB-17 [9]), затруднено отсутствием информации о методах синтеза этих соединений.

В настоящее время самой большой свободно доступной БД уже синтезированных органических соединений является PubChem, которая содержит информацию о структуре и свойствах более чем

97 млн. молекул [10]. Однако, теоретический размер химического пространства лекарственно-подобных соединений составляет 10^{40} молекул [11], что свидетельствует о возможностях существенного увеличения количества и разнообразия органических соединений, среди которых могут быть найдены новые фармакологические вещества.

Одной из первых попыток генерации *in silico* виртуальной библиотеки органических соединений, содержащей наряду со структурными формулами продуктов синтетических реакций также информацию об исходных реагентах и реакциях синтеза, является проект SAVI (Synthetically Accessible Virtual Inventory), начатый консорциумом американских и европейских ученых и специалистов в 2015 году [12]. Целью этого проекта было создание очень большой базы данных, содержащей информацию о 10^8-10^9 потенциально легко-синтезируемых соединений, которая может быть использована для поиска новых лекарственно-подобных веществ среди ещё не исследованных и не представленных в других БД молекул. Современная версия свободно доступной химической библиотеки SAVI содержит информацию о более чем 283 млн. органических соединений, полученных *in silico* с использованием набора правил для 14 химических реакций (трансформаций) и более 377 тыс. детально аннотированных исходных реагентов (building blocks) [13].

Данные о химических соединениях библиотеки SAVI доступны на веб-сайте NCI/NIH в формате SDF [12]. Каждое вещество охарактеризовано 62-мя дескрипторами, содержащими информацию об используемых исходных реагентах (идентификаторы в каталоге Sigma-Aldrich, стоимость и т.д.), о предполагаемой реакции (условия, защита, предполагаемый выход, оценка стоимости синтеза и др.), а также расчетные оценки свойств пригодности продуктов реакции для разработки новых лекарственных препаратов (соответствие “правилу трёх”, “правилу пяти Липински”, коэффициент распределения *н*-октанол/вода, доля sp3-гибридизированных атомов углеродов, общая полярная площадь поверхности TPSA, прогноз генотоксичности, и др.). Таким образом, число различных записей в химической библиотеке SAVI составляет более 10^{10}, что позволяет отнести её к категории “больших данных” (Big Data) [14].

Для компьютерной обработки подобных больших библиотек химических данных в настоящее время не существует готовых решений. Несмотря на разнообразие предложенных алгоритмов, разработанные программные комплексы требуют для обработки таких данных слишком больших временных затрат. Многие широко используемые в хемоинформатике подходы требуют ручной обработки исследуемых соединений, что невозможно при анализе больших данных.

Таким образом, одной из основных проблем анализа больших химических данных является отсутствие компьютерных программ, обеспечивающих высокую скорость обработки информации. Трудности с масштабированием программного обеспечения связаны, прежде всего, с высокой сложностью алгоритмов [15]. Применение параллельных вычислений и распределенного хранения данных является решением проблемы масштабируемости и необходимо для обработки больших данных при изучении ещё неисследованного химического пространства.

Целью нашей работы стала разработка подхода к обработке больших химических данных *in silico* и его применение для анализа 283 млн. молекул библиотеки SAVI, чтобы оценить перспективы её использования для поиска новых антиретровирусных веществ.

МЕТОДИКА

Программный комплекс ChemAxon Instant JChem

Для выполнения нашей работы был необходим программный комплекс, который позволял бы создавать, обновлять и анализировать информацию о структурных формулах и других характеристиках молекул, расположенную на удалённых SQL-серверах. Для решения этой задачи мы выбрали пакет программ Instant JChem (IJC, ChemAxon), с помощью которого было реализовано централизованное управление массивом данных, представленных в библиотеке SAVI:

- JChem Manager обеспечивает возможность создавать структуру таблицы на основе реляционной СУБД, подключаться к удалённой таблице, загружать и экспортировать структурные и дополнительные данные;
- JChem Base обеспечивает возможность подключиться к уже существующей таблице, содержащей информацию о структуре и различных свойствах молекул, просматривать и фильтровать данные, выполнять SQL-запросы для поиска и обработки информации;
- Marvin Sketch – приложение для ввода в компьютер, редактирования и просмотра структурных формул молекул и информации о химических реакциях.

Таким образом, программный комплекс IJC позволяет создавать на удалённых SQL-серверах таблицы, импортировать в них различную информацию о химических соединениях из файлов различных форматов (MOL, SDF, XLS и др.), получать необходимые данные с помощью SQL-запросов и экспортировать их, просматривать и редактировать структуры веществ непосредственно в таблице из единого места доступа посредством подключения к удалённо расположенной БД на основе различных СУБД (MySQL, PostgreSQL, Oracle и др.).

Алгоритм сравнения двух больших библиотек химических данных

Для оценки SAVI как источника новых соединений, потенциально обладающих антиретровирусной активностью, мы решили сопоставить её со свободно доступной базой данных PubChem, содержащей информацию о структуре и свойствах около 97 млн. органических веществ.

В настоящее время в хемоинформатике применяется несколько форматов представления структурных формул: MOL, SDF, SMILES и InChI. Файл в формате MOL содержит информацию о матрице связности (Connection Table), характеризующей атомы и связи между ними. SDF (Structure-Data File) содержит информацию о структурных формулах и других характеристиках для одного или более веществ (набор записей в формате MOL, характеризующих отдельные молекулы, и связанных с каждой молекулой данных) [16]. SMILES (Simplified Molecular Input Line Entry System) является линейным форматом представления информации о структурных формулах молекул в виде строки символов ASCII, где каждый атом описывается соответствующим ему символом периодической системы элементов. Запись SMILES формируется в соответствии с определёнными правилами [17]. Международный химический идентификатор IUPAC InChI также является линейным форматом представления данных, согласно которому структура молекулы описывается “слоями”, разделёнными символом “/”. Обязательные слои представлены формулой соединения в соответствии с правилами Хилла, данными о связанности атомов и количестве атомов водорода. Далее при необходимости могут быть указаны другие слои: заряд, стереохимия двойных связей и т.д. [18]. Для сравнения двух упомянутых выше библиотек химических данных нами был выбран способ представления структурных данных в формате SMILES, поскольку такая информация занимает значительно меньше памяти, чем структура молекулы в форматах MOL или SDF, и более точно отражает структуру вещества, в отличие от формата INCHI, который не всегда корректно отображает структуру молекулы [18].

Для проведения процедуры сравнения были использованы коды SMILES, представленные в таблицах MySQL для библиотеки SAVI и в загруженных с сайта (https://pubchem.ncbi.nlm.nih.gov) SDF файлах. На выходе формировался файл, в котором для обозначения совпавших структур использовались уникальные идентификаторы каждой библиотеки: E_NAME для SAVI, CID для PubChem, а также был записан соответствующий код SMILES. Используя полученные при сравнении CID соединений, были отобраны все записи о биологическом тестировании данных молекул из БД PubChem. С помощью текстового фильтра, применяемого к названию и мишеням биологического тестирования, были выбраны только методы, нацеленные на исследование анти-ВИЧ активности. Далее были извлечены результаты биологического тестирования для каждого отобранного соединения из данных, представленных в свободном доступе на PubChem.

Общая схема обработки данных библиотеки SAVI, представлена на рисунке 1.

Прогноз антиретровирусной активности с использованием PASS

Прогнозирование анти-ВИЧ активности для соединений, представленных в библиотеке SAVI, осуществляли с использованием компьютерной программы PASS (Prediction of Activity Spectra for Substance). PASS позволяет оценивать вероятности наличия (Pa) и отсутствия (Pi) конкретной биологической активности на основе анализа взаимосвязей структура-активность для веществ обучающей выборки с установленной экспериментально активностью. Структура исследуемых соединений представлена в виде

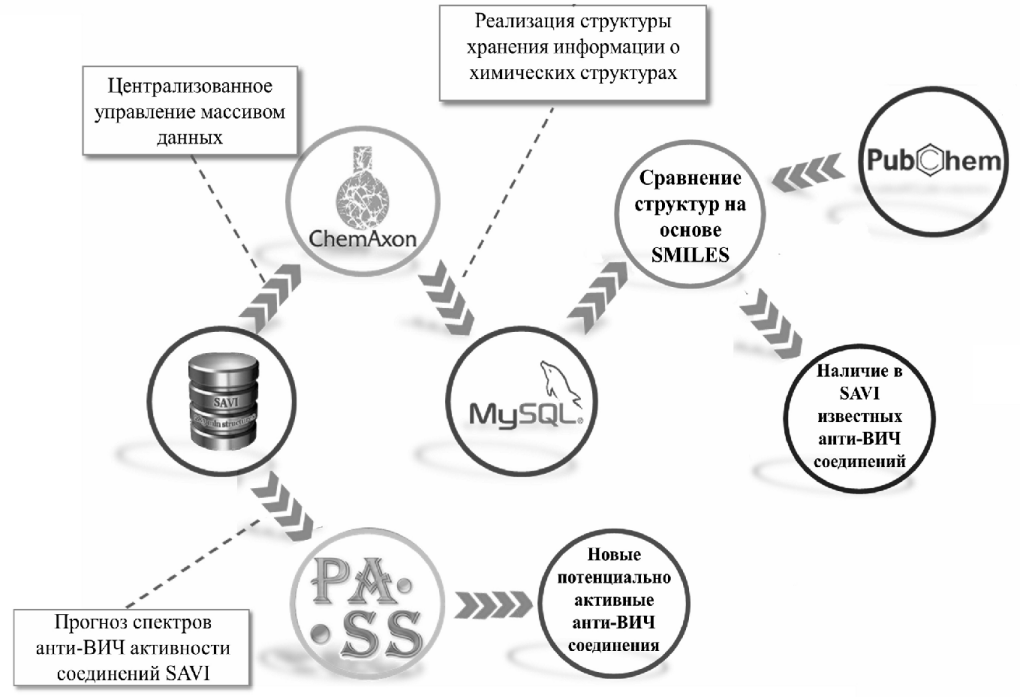

Рисунок 1. Схема обработки информации, представленной в SAVI, для выявления уже изученных анти-ВИЧ соединений с целью подтверждения их наличия в исследуемом химическом пространстве.

дескрипторов многоуровневых атомных окрестностей (Multilevel Neighborhoods of Atoms – MNA). Алгоритм установления взаимосвязей "структура-активность" на основе анализа имеющейся информации для соединений обучающей выборки и прогнозирования активности для новых (не включённых в обучающую выборку) веществ основан на Байесовских оценках. Более подробное описание используемого в PASS подхода приведено в публикациях [19, 20].

Прогноз может быть получен для полностью определённых, однокомпонентных, незаряженных структур, содержащих не менее трёх атомов углерода, имеющих молекулярную массу не более 1250 а.е.м. Результат прогноза представляется как список видов активности с соответствующими вероятностями Pa и Pi.

База знаний (SAR Base) программы PASS формируется в процессе обучения программы на основе анализа информации, содержащейся в обучающей выборке. В рамках настоящей работы нами были построены две специализированные SAR Base. Первая, SAR I, была построена на основе информации о 4289 ингибиторах протеазы, 2500 ингибиторах обратной транскриптазы и 918 ингибиторах интегразы ВИЧ-1, содержащихся в стандартной версии PASS. Вторая, SAR II, была сформирована путём добавления к первой базе 389996 соединений с известными 3868 видами молекулярных механизмов действия. Для структур, отобранных при пороге Pa>0,3 с использованием одной SAR Base, выполняли прогноз с использованием другой SAR Base. Таким образом, соединения, идентифицированные с использованием SAR II, были подвергнуты прогнозу, основанному на SAR I, и наоборот. Средняя точность прогноза (IAP) для обеих SAR Base, которая является вероятностью того, что для случайно взятой пары активного и неактивного соединений прогноз будет правильным (значение Pa для активного соединения будет выше, чем для неактивного), составила 99%. Точность прогноза для ингибиторов интегразы, протеазы и обратной транскриптазы составила 99%, 99% и 98%, соответственно.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Данные, полученные при сравнении SAVI и PubChem

Пересечение SAVI со структурами БД PubChem составило 43267 химических структур (~0,015% библиотеки SAVI), среди которых 632 были протестированы против ВИЧ (рис. 2). Из них по данным биологического тестирования антиретровирусная активность была найдена у 41 вещества. Восемь соединений ингибировали интегразу ВИЧ-1, 15 – обратную транскриптазу; 13 соединений блокировали взаимодействие вирусного белка Vif с фермент-каталитическим полипептидом типа 3G, корректирующего мРНК апопротеина B (APOBEC3G); 8 веществ были идентифицированы как молекулы, связывающиеся с трансактивируемым регуляторным элементом в области длинных концевых повторов РНК вируса (TAR); для 3 соединений активность против ВИЧ была доказана в экспериментах

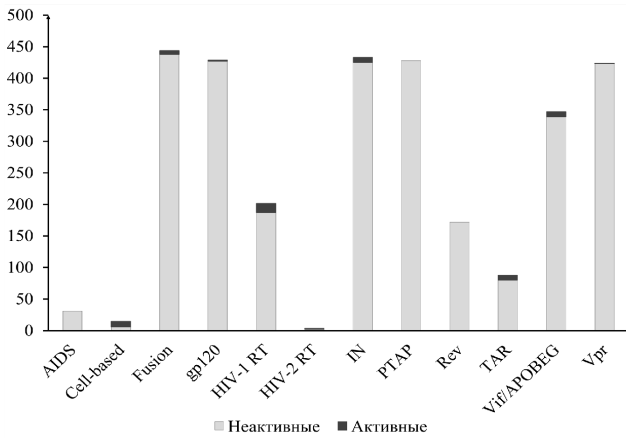

Рисунок 2. Распределение изученных на анти-ВИЧ активность соединений, общих для SAVI и PubChem.

на клеточных культурах; две молекулы проявляли активность в отношении вирусного белка gp120; одно соединение – в отношении vpr. Также было установлено действие 12 молекул как ингибиторов слияния вируса с клеткой. Активность трёх веществ была подтверждена как в экспериментах на клеточных культурах, так и против конкретной мишени. Для 11 соединений было установлено действие на несколько мишеней. Однако, среди соединений, общих для обеих БД, не было выявлено ни одного вещества, протестированного в отношении протеазы ВИЧ-1.

Поскольку 283 млн. соединений библиотеки SAVI были сгенерированы с использованием ограниченного списка из 14 трансформаций для 377 тыс. "строительных блоков", в составе этой библиотеки могут быть выявлены аналоги найденных нами 41-го соединения, для которых установлено наличие анти-ВИЧ активности. Поиск таких аналогов предполагается осуществить в ходе дальнейших исследований.

На рисунке 3 представлены примеры трёх выявленных нами в ходе проведенного анализа соединений, обладающих антиретровирусной активностью: a – ингибитор интегразы; b – ингибитор слияния вируса с клеткой; c – ингибитор обратной транскриптазы.

Соединение, представленное на рисунке 3a, проявляет противовирусную активность в отношении ряда резистентных штаммов ВИЧ-1 путём ингибирования каталитической функции, независимой от клеточного фактора транскрипции LEDGF/p75 (Lens Epithelium-Derived Growth Factor), благодаря связыванию с аллостерическим центром фермента и нарушению белок-белкового взаимодействия IN-LEDGF/p75 [21-23]. Взаимодействие IN-LEDGF/p75 является перспективным молекулярным механизмом, воздействие на который может использоваться для антиретровирусной терапии, и на данный момент известно сравнительно мало соединений с подобным механизмом действия. Соединение, представленное на рисунке 3b, ингибирует образование капсидного белка p24 и нарушает формирование gp41 6-HB структуры при слиянии мембраны вируса с мембраной

клетки-хозяина, тем самым подавляя репликацию ВИЧ-1 в клеточных линиях [24, 25]. Отмечается, что данное вещество обладает более низкой цитотоксичностью по сравнению с другими аналогами, что свидетельствует о перспективности его дальнейшего исследования [26]. Соединение, представленное на рисунке 3c, является аналогом первого допущенного к клиническому применению нуклеозидного ингибитора обратной транскриптазы (Азидотимидин) и проявляет активность в низких микромолярных концентрациях как против ВИЧ-1 дикого типа, так и против штаммов, резистентных к ненуклеозидным ингибиторам обратной транскриптазы [27].

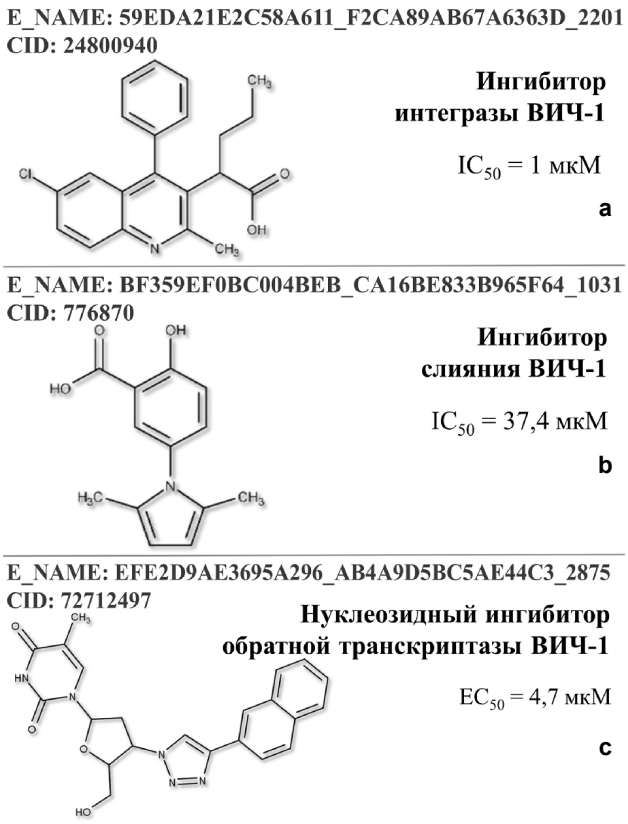

Рисунок 3. Примеры анти-ВИЧ соединений, выявленных при пересечении SAVI и PubChem. Для каждого соединения представлены соответствующие идентификаторы: CID – идентификатор PubChem, E_NAME – идентификатор SAVI.

Прогноз спектров анти-ВИЧ активности для соединений из библиотеки SAVI

Для выявления новых, ранее неисследованных соединений, потенциально активных в отношении ВИЧ-1, мы выполнили прогноз трёх видов антиретровирусной активности (ингибиторы обратной транскриптазы, протеазы и интегразы) для молекул из библиотеки SAVI, не содержащихся в PubChem. С вероятностью Pa>0,3 выявлено 535334 потенциальных ингибиторов обратной транскриптазы, 89838 – ингибиторов интегразы, и 36605 – ингибиторов протеазы ВИЧ-1 (всего 661777 соединений).

На рисунке 4 представлены распределения количества прогнозируемых соединений в зависимости от величины вероятности наличия активности Pa для двух использованных в рамках настоящего исследования SAR Base (SAR II и SAR I). Как видно из данных на рисунке 4, при пороге Pa>0,7 для синтеза и биологического тестирования может быть отобрано 10 ингибиторов интегразы, 63 ингибиторов протеазы и 32029 ингибиторов обратной транскриптазы ВИЧ-1.

ЗАКЛЮЧЕНИЕ

Поиск новых антиретровирусных соединений для терапии ВИЧ-инфекции и профилактики сопутствующих заболеваний остаётся актуальной задачей биомедицины. Наличие серьезных побочных эффектов и возникновение резистентности к используемым в настоящее время лекарственным препаратам обуславливают необходимость расширения химического пространства для поиска новых веществ, обладающих анти-ВИЧ активностью.

Одной из первых виртуальных баз данных легко синтезируемых органических соединений является библиотека SAVI, современная версия которой содержит информацию о более чем 283 млн. молекул (более 10^{10} различных данных). Для хранения и обработки такого рода “больших химических данных”

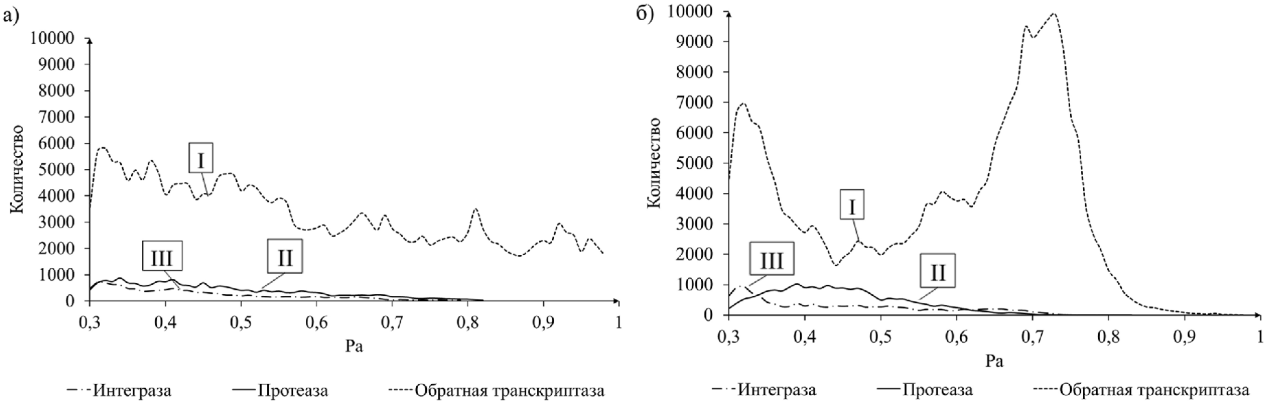

Рисунок 4. Распределение потенциальных антиретровирусных соединений из библиотеки SAVI в зависимости от величины вероятности наличия активности Pa: а – для обучающей выборки SAR II; б – для обучающей выборки SAR I. I – ингибиторы обратной транскриптазы; II – ингибиторы интегразы; III – ингибиторы протеазы ВИЧ-1.

нами развёрнута информационно-вычислительная инфраструктура, обеспечивающая централизованное управление массивом данных SAVI, расположенном на удалённых SQL-серверах.

Нами разработан алгоритм сравнения больших химических БД на основе представления структурных формул в формате SMILES, с использованием которого мы сопоставили базы данных SAVI и PubChem. Показано, что лишь около 0,015% соединений SAVI также включено в БД PubChem, что свидетельствует о существенной новизне изучаемых нами данных. Вместе с тем, проведённый нами анализ позволил выявить 632 соединения, ранее протестированных на анти-ВИЧ активность, среди которых 41 было активным.

Прогноз антиретровирусной активности для 99,99% новых (не входящих в пересечение с PubChem) соединений SAVI позволил идентифицировать потенциальные ингибиторы обратной транскриптазы, протеазы и интегразы ВИЧ-1. Показано, что в зависимости от пороговых значений вероятности Pa имеются значительные возможности для отбора конкретных молекул с целью их синтеза и биологического тестирования. Полученные результаты позволяют рассматривать SAVI как перспективный источник новых антиретровирусных соединений.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена в рамках гранта РФФИ-NIH № 17-54-30015-НИЗ_а.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит каких-либо исследований с участием людей или с использованием животных в качестве объектов.

ЛИТЕРАТУРА

1. HIV/AIDS. Data and statistics. Retrieved January, 2019, from https://www.who.int/hiv/data/en/
2. *Pau A.K., George J.M.* (2014) Infect. Dis. Clin. North. Am., **28**(3), 371-402.
3. Antiretroviral drugs used in the treatment of HIV infection. Retrieved January, 2019, from https://www.fda.gov/forpatients/illness/hivaids/treatment/ucm118915.htm
4. *Lu D.Y., Wu H.Y., Yarla N.S., Xu B., Ding J., Lu T.R.* (2018) Infect. Disord. Drug Targets, **18**(1), 15.
5. *Montessori V., Press N., Harris M., Akagi L., Montaner J.S.* (2004) CMAJ, **170**(2), 229-238.
6. *Iyidogan P., Anderson K.S.* (2014) Viruses, **6**(10), 4095-4139.
7. ChemNavigator. Retrieved January, 2019, from https://www.chemnavigator.com/
8. ZINC. Retrieved January, 2019, from http://zinc.docking.org/
9. GBD17. Retrieved January, 2019, from http://gdb.unibe.ch/
10. PubChem. Retrieved January, 2019, from https://pubchem.ncbi.nlm.nih.gov/
11. *Peach M.L., Nicklaus M.C.* (2018) in: Applied Chemoinformatics: Achievements and Future Opportunities (Engel T., Gasteiger J., eds.) Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, p. 391.
12. SAVI. Retrieved January, 2019, from https://cactus.nci.nih.gov/download/savi_download/
13. *Pevzner Y.U., Ihlenfeldt W.D., Nicklaus M.C.* (2017) Synthetically accessible virtual inventory (SAVI). Abstracts of the 253rd American Chemical Society National Meeting (San-Francisco, CA: CINF), 141.
14. *Tetko I.V., Engkvist O., Koch U., Reymond J.L., Chen H.* (2016) Mol. Inform., **35**(11-12), 615-621.
15. *Yang A., Troup M., Ho J.W.K.* (2017) Comput. Struct. Biotechnol. J., **15**, 379-386.
16. *Dalby A., Nourse J.G., Hounshell W.D., Gushurst A.K.I., Grier D.L., Leland B.A., Laufer J.* (1992) J. Chem. Inform. Comput. Sci., **32**(3), 244-255.
17. OpenSMILES. Retrieved January 2019, from http://opensmiles.org/
18. *Heller S., McNaught A., Stein S., Tchekhovskoi D., Pletnev I.* (2013) J. Cheminform., **5**(1), 7.
19. *Filimonov D.A., Druzhilovskiy D.S., Lagunin A.A., Gloriozova T.A., Rudik A.V., Dmitriev A.V., Pogodin P.V., Poroikov V.V.* (2018) Biomedical Chemistry: Research and Methods, **1**(1), e00004. DOI: 10.18097/bmcrm00004
20. *Filimonov D.A., Lagunin A.A., Gloriozova T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V.* (2014) Chem. Heterocycl. Compd., **50**, 444-457.
21. *Christ F., Voet A., Marchand A., Nicolet S., Desimmie B.A., Marchand D., Bardiot D., Van der Veken N.J., Van Remoortel B., Strelkov S.V., De Maeyer M., Chaltin P., Debyser Z.* (2010) Nat. Chem. Biol., **6**(6), 442-448.
22. *Di Santo R.* (2014) J. Med. Chem., **57**(3), 539-566.
23. *Hu G., Li X., Zhang X., Li Y., Ma L., Yang L.M., Liu G., Li W., Huang J., Shen X., Hu L., Zheng Y.T., Tang Y.* (2012) J. Med. Chem., **55**(22), 10108-10117.
24. *Liu K., Lu H., Hou L., Qi Z., Teixeira C., Barbault F., Fan B.T., Liu S., Jiang S., Xie L.* (2008) J. Med. Chem., **51**(24), 7843-7854.
25. *He X.Y., Lu L., Qiu J., Zou P., Yu F., Jiang X.K., Li L., Jiang S., Liu S., Xie L.* (2013) Bioorg. Med. Chem., **21**(23), 7539-7548.
26. *Patel R.V., Park S.W.* (2015) Bioorg. Med. Chem., **21**(23), 5247-5263.
27. *Sirivolu V.R., Vernekar S.K., Ilina T., Myshakina N.S., Parniak M.A., Wang Z.* (2013) J. Med. Chem., **56**(21), 8765-8780.

Поступила в редакцию: 22. 02. 2019.
После доработки: 01. 03. 2019.
Принята к печати: 04. 03. 2019.

DISCOVERING NEW ANTIRETROVIRAL COMPOUNDS IN "BIG DATA" CHEMICAL SPACE OF THE SAVI LIBRARY

P.I. Savosina[1]*, L.A. Stolbov[1], D.S. Druzhilovskiy[1], D.A. Filimonov[1], M.C. Nicklaus[2], V.V. Poroikov[1]

[1]Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: polinkasavosina@gmail.com
[2]Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland 21702, United States

Despite significant advances in the application of highly active antiretroviral therapy, the development of new drugs for the treatment of HIV infection remains an important task because the existing drugs do not provide a complete cure, cause serious side effects and lead to the emergence of resistance. In 2015, a consortium of American and European scientists and specialists launched a project to create the SAVI (Synthetically Accessible Virtual Inventory) library. Its 2016 version of over 283 million structures of new easily synthesizable organic molecules, each annotated with a proposed synthetic route, were generated *in silico* for the purpose of searching for safer and more potent pharmacological substances. We have developed an algorithm for comparing large chemical databases (DB) based on the representation of structural formulas in SMILES codes, and evaluated the possibility of detecting new antiretroviral compounds in the SAVI database. After analyzing the intersection of SAVI with 97 million structures of the PubChem database, we found that only a small part of the SAVI (~0.015%) is represented in PubChem, which indicates a significant novelty of this virtual library. However, among those structures, 632 compounds tested for anti-HIV activity were detected, 41 of which had the desired activity. Thus, our studies for the first time demonstrated that SAVI is a promising source for the search for new anti-HIV compounds.

Key words: "Big Data"; SAVI; PubChem; new drug-like compounds; antiretroviral activity; PASS prediction