ОБЗОР

КОМПЬЮТЕРНОЕ ПРОГНОЗИРОВАНИЕ СПЕКТРОВ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ ХИМИЧЕСКИХ СОЕДИНЕНИЙ: ВОЗМОЖНОСТИ И ОГРАНИЧЕНИЯ

Д.А. Филимонов[1], Д.С. Дружиловский[1], А.А. Лагунин[1,2], Т.А. Глориозова[1], А.В. Рудик[1], А.В. Дмитриев[1], П.В. Погодин[1], В.В. Поройков[1]*

[1]Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121 Москва, ул. Погодинская, 10 стр. 8; *эл. почта: vladimir.poroikov@ibmc.msk.ru
[2]Российский национальный исследовательский медицинский университет имени Н.И. Пирогова Минздрава России,
117997, Москва, ул. Островитянова, д. 1.

Важной характеристикой химических соединений является их биологическая активность, поскольку её наличие может стать основой для использования вещества в терапевтических целях, либо, напротив, ограничить возможности его практического применения вследствие проявления побочных и токсических эффектов. Компьютерная оценка спектра биологической активности дает возможность определить наиболее перспективные направления для тестирования фармакологического действия конкретных веществ и отсеять потенциально опасные молекулы на ранних стадиях исследований. Свыше 25 лет нами осуществляется разработка и совершенствование компьютерной программы PASS (Prediction of Activity Spectra for Substances), предназначенной для прогнозирования спектра биологической активности вещества по структурной формуле его молекул. Прогноз осуществляется на основе анализа зависимостей “структура-активность” для соединений обучающей выборки, в настоящее время содержащей информацию о структурах и известных видах биологической активности более чем для миллиона молекул. Описание структуры молекул органического соединения реализовано в PASS посредством дескрипторов атомных окрестностей (Multilevel Neighborhoods of Atoms), прогнозирование активности для новых соединений выполняется алгоритмом на основе “наивного Байесовского подхода” и зависимостей “структура-активность”, выявляемых при анализе обучающей выборки. Нами созданы и совершенствуются как локальные версии программы PASS, так и свободно доступные в Интернет веб-ресурсы на основе PASS (http://way2drug.com): прогноз нескольких тысяч видов биологической активности (фармакологические эффекты, молекулярные механизмы действия, специфическая токсичность и побочное действие, метаболизм, а также влияние на нежелательные мишени, молекулярный транспорт, генную экспрессию), прогноз цитотоксичности для опухолевых и неопухолевых клеточных линий, прогноз канцерогенности, прогноз индуцированных органическими соединениями изменений профилей экспрессии генов, прогноз взаимодействия с ферментами метаболизма лекарств, в том числе прогноз сайтов метаболизма, а также прогноз принадлежности к субстратам и/или метаболитам этих ферментов. Веб-ресурс Way2Drug используют свыше 19 тысяч исследователей более чем из 100 стран мира, что позволило им осуществить свыше 600 тысяч прогнозов и опубликовать около 500 работ с описанием полученных результатов. Анализ опубликованных работ показывает, что в некоторых случаях приводимая авторами этих публикаций интерпретация результатов прогноза требует корректировки. В рамках настоящей работы мы представим теоретическое обоснование и рассмотрим на конкретных примерах возможности и ограничения компьютерного прогнозирования спектров биологической активности.

Ключевые слова: анализ зависимостей “структура-активность”; спектр биологической активности; компьютерное прогнозирование; PASS; точность; предсказательная способность; веб-ресурс Way2Drug

DOI: 10.18097/BMCRM00004

ВВЕДЕНИЕ

Наличие у химического соединения биологической активности дает возможность использовать его в качестве субстанции лекарственного препарата для терапии определенной патологии [1]. С другой стороны, биологическая активность может стать причиной проявления веществом побочных и токсических эффектов, что ограничивает возможности его практического применения [2].

В настоящее время около 80 млн различных химических соединений доступно для тестирования в виде уже синтезированных образцов [3]; в регистрационной системе Chemical Abstracts Service содержится информация о 138 млн органических и неорганических веществ, описанных в литературе с начала XIX столетия [4]; *in silico* сгенерированы сотни миллионов структурных формул органических молекул вместе с исходными реагентами и реакциями синтеза [5] и свыше 166 млрд структурных формул, полностью покрывающих химическое пространство, включающее до 17 атомов C, N, O, S и галогенов [6].

Число известных молекулярных мишеней лекарственных препаратов в организме человека составляет несколько тысяч [7, 8], а число фармакотерапевтических эффектов, возникающих при взаимодействии фармакологических веществ с этими мишенями – несколько сотен [9].

Экспериментальное тестирование взаимодействия многих миллионов химических соединений с тысячами молекулярных мишеней невозможно как с экономической, так и с практической точки зрения [10]. Таким образом, возникает необходимость предварительного отбора молекул, с наибольшей вероятностью взаимодействующих с целевыми молекулярными мишенями и проявляющих, благодаря такому взаимодействию, необходимое фармакотерапевтическое действие. С этой целью сегодня широко применяют компьютерные методы дизайна лекарств, основанные как на структуре

макромолекулы-мишени, так и на структуре лигандов [11, 12].

Использование методов, основанных на структуре мишени, требует наличия информации о пространственной структуре макромолекулы-мишени, сопряжено с необходимостью проведения ресурсоемких вычислений и связано с рядом других ограничений [13]. В Интернете доступны некоторые веб-ресурсы, позволяющие предсказывать профили биологической активности на основе молекулярного докинга и анализа ассоциативных взаимосвязей по аффинности к различным молекулярным мишеням лекарственных соединений из обучающей выборки (например, [14]).

Условием применения методов, основанных на структуре лигандов, является наличие «обучающих примеров» в виде массива информации о структуре молекул набора соединений и их взаимодействии с целевой мишенью или проявлении ими заданной биологической активности в некотором стандартизированном биологическом тесте. Часто это условие невыполнимо, особенно для новых фармакологических мишеней, представляющих особый интерес. В этих случаях применяют методы оценки биологической активности соединений на основе анализа структурного сходства их молекул, что не всегда гарантирует получение надежных оценок [15, 16]. Тем не менее, поиск «по сходству» является встроенной процедурой в некоторых базах данных коммерчески доступных образцов органических соединений (например, [3]), что помогает пользователю найти «хоть что-нибудь похожее» на структурную формулу, использованную в качестве запроса.

Появление свободно-доступных через Интернет баз данных, содержащих информацию о структуре и биологической активности химических соединений [17-19], создало необходимые предпосылки для развития методов дизайна лекарств, основанных на структуре лигандов и позволяющих прогнозировать профили биологической активности для новых веществ [20-25]. Для прогноза биологической активности используется либо анализ структурного сходства (например, SwissTargetPrediction), либо методы машинного обучения (например, ChemProt).

Программа PASS (Prediction of Activity Spectra for Substances) была создана намного раньше, чем упомянутые выше разработки, что было отмечено в недавней публикации Андреаса Бендера с сотр.: «*One of the earliest and most widely used examples of data-mining target elucidation is the continuously curated and expanded Prediction of Activity Spectra for Substances (PASS) software, which was assimilated from the bioactivites of more than 270,000 compound-ligand pairs*» [26].

В 1990 году была опубликована первая работа, в которой упоминалось о реализации нами прогнозирования спектров биологической активности [27]. В 1993 году были продемонстрированы преимущества программы PASS в сравнении с предсказаниями специалистов-экспертов для независимой выборки веществ [28]. Двумя годами позже было опубликовано детальное описание используемого в PASS подхода, включая некоторые примеры применения [29]. В 1996 году представлено первое описание PASS на английском языке [30]. В 1999 году был реализован первый в мире свободно-доступный веб-ресурс, позволявший пользователям осуществлять прогнозирование спектров биологической активности через Интернет [31, 32]. А в 2003 году результаты прогноза спектров биологической активности с использованием PASS для 250 тысяч соединений из базы данных Национального института рака США (Open NCI database) были представлены на веб сайте NCI/NIH [33].

Более детальную информацию об истории развития программы PASS можно найти в публикациях [34-37].

Ниже мы рассмотрим реализованный в настоящее время в PASS метод анализа зависимостей «структура-активность» и прогноза активности для новых веществ.

МАТЕРИАЛЫ И МЕТОДЫ

Биологическая активность органического соединения представляет собой результат его взаимодействия с биологическим объектом. Она зависит от характеристик соединения (структуры его молекулы), биологического объекта (вид, пол, возраст, и др.), способа воздействия (путь введения, доза) и особенностей условий эксперимента. В PASS биологическая активность описывается **качественно** (“активное” или “неактивное”); при количественных данных соединение признается «активным», если полуэффективная концентрация меньше 10 мкМ.

Спектр биологической активности органического соединения – это множество различных видов биологической активности, которые отражают результат его взаимодействия с различными биологическими объектами. Он отражает “внутренние”, присущие данному соединению свойства, зависящие только от строения его молекулы. Вводя это обобщающее понятие, мы обеспечиваем возможность объединения больших массивов данных из различных источников, поскольку информация из конкретной публикации не охватывает всех аспектов биологического действия описываемого в нем органического соединения. При этом мы следуем принципу «презумпции невиновности»: в PASS принимается, что соединение не обладает теми видами биологической активности, которые не указаны в его спектре. Хотя нельзя исключить ситуации, когда информация о какой-либо активности органического соединения не была найдена в доступных источниках, либо оно обладает некоторой биологической активностью, но на эту активность соединение еще не испытывалось. Это приближение не оказывает существенного влияния на результаты анализа зависимости «структура–активность» и выполняемого на этой основе прогноза благодаря статистической устойчивости используемого в PASS метода расчёта [38]. Прогнозируемый PASS спектр биологической активности органического соединения

включает в себя фармакологические эффекты, молекулярные механизмы действия, специфическую токсичность и побочное действие, метаболизм, а также влияние на нежелательные мишени, молекулярный транспорт, генную экспрессию (рис. 1).

Необходимо подчеркнуть, что для прогноза с помощью PASS может быть использован любой способ объективной классификации органических соединений. Если соответствующие классы действительно определяются особенностями структуры молекул, то прогноз принадлежности к этим классам может быть вполне успешным. Например, интервал значений некоторой количественной величины можно рассматривать в PASS как «активность»: если значение величины принадлежит этому интервалу, то соединение «активно», и «неактивно» в иных случаях. Поэтому ясно, что применимость PASS гораздо шире прогноза только спектра биологической активности.

Описание структуры органического соединения основано на его структурной формуле, так как это единственная доступная информация о соединении на ранних стадиях его исследования (соединение может только планироваться к синтезу). Для описания структуры химического соединения нами были предложены специальные дескрипторы, которые мы назвали дескрипторами **MNA** (Multilevel Neighborhoods of Atoms – многоуровневые атомные окрестности) [39]. Эти дескрипторы были разработаны с учетом нашего практического опыта по решению задач о поиске зависимостей «структура–свойство» для гетерогенных выборок органических соединений, обладающих широким набором видов биологической активности [34, 35].

Структурная формула, записываемая в соответствии с номенклатурными правилами в химии, отражает атомный состав и взаимное расположение атомов в молекуле. Дескрипторы MNA основаны на таком представлении структурной формулы, в котором, согласно валентностям и зарядам атомов, явно указаны все атомы водорода и не учитываются типы связей: природа не знает, что такое «стёртые водороды», а кратность связей во многих случаях на самом деле должна быть дробной – например, в ароматическом кольце или в группе $–NO_2$, – можно лишь утверждать, имеется ли между данными двумя атомами достаточно устойчивая химическая связь или нет. В таком виде структурная формула становится однозначной даже формально – она не зависит, например, от альтернативных способов изображения ароматических систем.

На основе описанного представления структурной формулы дескрипторы MNA для каждого атома молекулы строятся рекурсивно следующим образом:

дескриптор MNA 0-го уровня – метка $\boldsymbol{A}$ самого атома; дескриптор MNA любого следующего уровня – условное обозначение структурного фрагмента $\boldsymbol{A}(\boldsymbol{D}_1\boldsymbol{D}_2...\boldsymbol{D}_i...)$, где $\boldsymbol{D}_i$ – дескриптор MNA предыдущего уровня для i-го непосредственного соседа данного атома с меткой $\boldsymbol{A}$.

Дескрипторы соседей $\boldsymbol{D}_1\boldsymbol{D}_2...\boldsymbol{D}_i...$ записываются в каком-нибудь однозначном порядке, например, лексикографическом.

Эта итерационная процедура может быть продолжена до любого уровня. Важно подчеркнуть, что метки атомов могут не только соответствовать общепринятым символам химических элементов, но и включать любую дополнительную информацию, например, о принадлежности атома к цепи или к какой-либо циклической системе, или что он является сайтом метаболизма.

На рисунке 2 представлена структура молекулы противоэпилептического препарата Топирамат и пример дескрипторов MNA для атома серы. Структура молекулы в PASS представлена как бесповторное

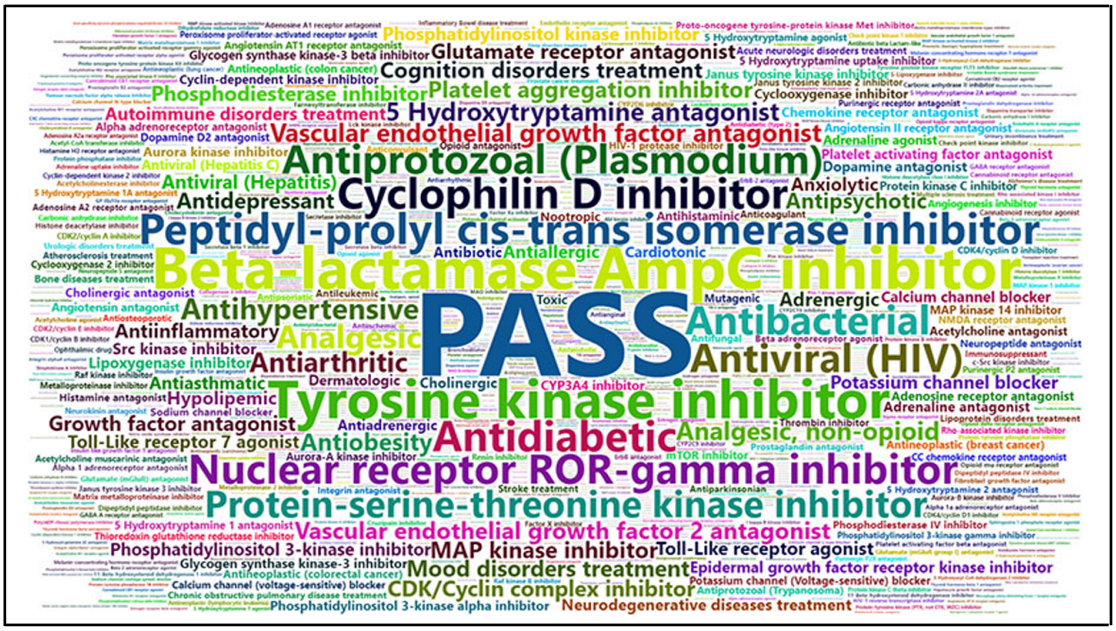
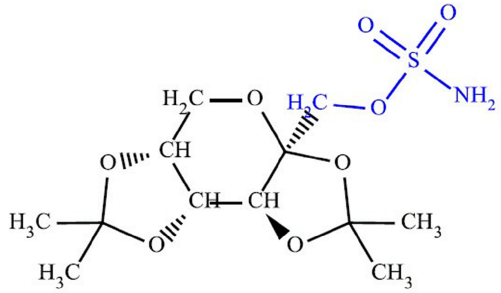


Рисунок 1. Облачное представление множества видов биологической активности, прогнозируемых PASS (версия 2017). Размер шрифта пропорционален количеству соединений с соответствующей активностью в обучающей выборке.

множество дескрипторов MNA 1-го и 2-го уровней. В дескрипторах 2-го уровня используется индикатор «–» для обозначения атомов, которые не входят ни в какие циклы. Пример множества дескрипторов MNA, описывающих в PASS соединение Топирамат, приведён на рисунке 3.

Оценить влияние стереоизомерии на проявление биологической активности PASS не позволяет, поскольку для многих видов активности в настоящее время невозможно создать репрезентативную обучающую выборку, учитывающую особенности пространственной структуры включенных в неё веществ. Оценка влияния на активность особенностей пространственной структуры изучаемых веществ может быть получена с применением методов молекулярного моделирования [40].



MNA/0: S
MNA/1: S(NOOO)
MNA/2: S(N(HHS)O(CS)O(S)O(S))
. . .

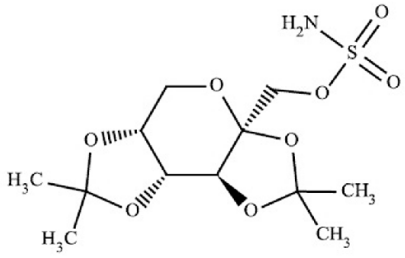
Рисунок 2. Дескрипторы MNA для атома серы молекулы соединения Топирамат (Topiramate, PubChem CID: 5284627). Фрагмент, соответствующий входящим в пример MNA/2 атомам, выделен синим цветом.

Важной особенностью дескрипторов MNA является их открытость – дескрипторы порождаются на основе самой структурной формулы, а не на основе какого-либо заранее составленного списка структурных фрагментов. Другая их особенность заключается в сохранении целостности фрагментов структуры в том смысле, что для каждого дескриптора MNA можно, при некотором навыке, изобразить соответствующий ему фрагмент.

В программе PASS органические соединения считаются эквивалентными, если их молекулярные структуры описываются одинаковым набором дескрипторов. Так как дескрипторы MNA не отражают стереохимические особенности молекулы, структуры, которые отличаются только стереохимически, формально считаются эквивалентными.

В PASS используются файлы, содержащие данные о химической структуре в форматах MOL или SDF [41]. Многие молекулярные редакторы и системы управления базами данных позволяют экспортировать данные в этих форматах. Дескрипторы MNA (как для прогнозирования спектра активности соединения, так и для добавления соединения в базу знаний с информацией о связи «структура–активность» SAR Base) генерируются только в случае, если структура соединения удовлетворяет следующим критериям:

- каждый атом в молекуле должен быть обозначен соответствующим символом из периодической таблицы элементов; символы неопределенного атома A, Q, * или метки R группы не допускаются;
- каждая связь в молекуле должна быть ковалентной простой, двойной или тройной связью;
- в структуре соединения должно присутствовать не менее трёх атомов углерода;



MNA/1	MNA/2
	C(C(CCO-H)C(CCO-H)O(CC)-H(C))
	C(C(CCO-H)C(COO-C)O(CC)-H(C))
	C(C(CCO-H)C(CO-H-H)O(CC)-H(C))
	C(C(CCO-H)O(CC)O(CC)-C(C-H-H-O))
	C(C(CCO-H)O(CC)-H(C)-H(C))
	C(O(CC)O(CC)-C(C-H-H-H)-C(C-H-H-H))
	O(C(CCO-H)C(OO-C-C))
	O(C(COO-C)C(CO-H-H))
	O(C(COO-C)C(OO-C-C))
	-H(C(CCO-H))
	-H(C(CO-H-H))
	-H(-C(C-H-H-H))
	-H(-C(C-H-H-O))
	-H(-N(-H-H-S))
	-C(C(COO-C)-H(-C)-H(-C)-O(-C-S))
	-C(C(OO-C-C)-H(-C)-H(-C)-H(-C))
	-N(-H(-N)-H(-N)-S(-N-O-O-O))
	-O(-C(C-H-H-O)-S(-N-O-O-O))
	-O(-S(-N-O-O-O))
	-S(-N(-H-H-S)-O(-C-S)-O(-S)-O(-S))
HC	
HN	
CHHHC	
CHHCO	
CHCCO	
CCCOO	
NHHS	
OCC	
OCS	
OS	
SNOOO	

Рисунок 3. Структурная формула соединения Топирамат и множество дескрипторов MNA, описывающих в PASS структуру его молекул.

- структура соединения должна быть однокомпонентной; её части, состоящие из одного атома, такие как Cl, Cl^-, OH^-, Na^+ и т.д. (атомы водорода не принимаются во внимание) исключаются из процесса генерации MNA дескрипторов;
- сумма зарядов атомов молекулы должна быть равна нулю;
- абсолютная молекулярная масса соединения должна быть меньше 1250.

Если структура не удовлетворяет этим критериям или имеются какие-либо другие ошибки во входных данных, то генерируется соответствующее сообщение об ошибке.

Эти требования *де факто* являются стандартом, который в настоящее время общепринят и широко используется при подготовке обучающих выборок для построения зависимостей «структура-свойство» [42-44].

Обучающая выборка PASS состоит из тщательно отобранных записей о структуре и биологической активности органических соединений. Файл SAR Base (расширение «.SAR») создаётся в ходе обучения на основе обучающей выборки. **SAR Base** включает в себя словарь названий видов биологической активности (8054 терминов в версии 2017) и словарь дескрипторов MNA (106816 в версии 2017), данные, представленные в виде набора дескрипторов MNA соединений из обучающей выборки со спектрами их биологической активности (1025468 записей в версии 2017) и извлечённые в результате обучения знания о зависимостях «структура–активность» (7604 «правила» в версии 2017).

К сожалению, пользуясь только публично доступными источниками, невозможно составить достаточно большую коллекцию биологически активных соединений, для которых были бы известны результаты тестирования на все виды биологической активности. По этой причине некоторые виды биологической активности в SAR Base PASS (версия 2017) представлены более чем 100000 органических соединений (136451 – противоопухолевые), а другие – только несколькими (405 видов активности по 3 соединения, 384 – по 4, 389 – по 5, и т.д.).

В разных источниках информации биологические активности органических соединений описаны неодинаковыми терминами. Поэтому термины, описывающие спектры активности соединений в обучающей выборке, приводятся к единому «стандартизованному» виду.

Алгоритм прогноза PASS удобнее всего описывать на основе классического байесовского подхода, который можно сформулировать следующим образом [45-47]. Для химического соединения C по структуре его молекул, записанной в виде множества $\{D_1,...,D_m\}$ из m дескрипторов MNA, оценим вероятность $P(A|C)$ того, что соединение C принадлежит к классу A.

Согласно формуле Байеса:

$$P(A|C) = \frac{P(C|A) \cdot P(A)}{P(C)} \quad (1)$$

где $P(C|A)$ условная вероятность структуры C при условии, что химическое соединение принадлежит к классу A; $P(A)$ априорная вероятность принадлежности химического соединения к классу A; $P(C)$ – априорная вероятность структуры C.

Если допустить, что дескрипторы $D_1,...,D_m$ независимы в совокупности, то можно, согласно «наивному Байесовскому подходу», записать $P(C|A)$ как произведение условных вероятностей для отдельных дескрипторов:

$$P(C|A) \cong P(D_1, \ldots, D_m|A) = \prod_{i=1}^{m} P(D_i|A) \quad (2)$$

Это выражение приближённое, поскольку дескрипторы MNA заведомо являются зависимыми в силу способа их построения. Но из-за отсутствия приемлемых альтернатив нам остаётся лишь не забывать о приближённости получаемых формул.

В результате простых алгебраических преобразований получаем следующее выражение для логарифма отношения правдоподобия условной вероятности $P(A|C)$ для класса A и $P(B|C)$ для класса B в виде:

$$ln\left[\frac{P(A|C)}{P(B|C)}\right] \cong ln\left[\frac{P(A)}{P(B)}\right] + \sum_{i=1}^{m}\left\{ln\left[\frac{P(A|D_i)}{P(B|D_i)}\right] - ln\left[\frac{P(A)}{P(B)}\right]\right\} \quad (3)$$

Смысл полученного выражения вполне прозрачен: логарифм апостериорного отношения правдоподобия есть сумма логарифма априорного отношения правдоподобия и суммы вкладов отдельных дескрипторов. При этом, если принадлежность к классам A и B не зависит от данного дескриптора, то $P(A|D_i)$=P(A), $P(B|D_i)=P(B)$ и этот дескриптор не влияет на результат – его вклад в сумму нулевой. Это и есть результат классического байесовского подхода в приближении «наивный Байес» (Naïve Bayes) [45-47].

Формулу (3) можно записать в более традиционном виде. Для этого перенумеруем дескрипторы MNA по словарю и введем «координаты» соединений $x_i(C)$ так, что $x_i(C)$=1, если дескриптор D_i входит в описание соединения C, и $x_i(C)$=0 во всех остальных случаях. Тогда формулу (3) для логарифма отношения правдоподобия принадлежности соединения C к классам A и B можно записать в следующем виде:

$$ln\left[\frac{P(A|C)}{P(B|C)}\right] \cong a_0 + \sum_i a_i x_i(C) \quad (4a)$$

$$a_0 = ln\left[\frac{P(A)}{P(B)}\right] \quad (4b)$$

$$a_i = ln\left[\frac{P(A|D_i)}{P(B|D_i)}\right] - ln\left[\frac{P(A)}{P(B)}\right] \quad (4c)$$

Формула (4) имеет вид обычной линейной регрессии, однако, коэффициенты регрессии a_i

в ней вычисляются не в результате минимизации какого-либо критерия согласия, а непосредственно на основе описанного байесовского подхода.

В настоящее время в PASS класс B – это все соединения, не принадлежащие к классу A, $P(B)$=1-$P(A)$, $P(B|D_i)$=1-$P(A|D_i)$, и получаем из (3):

$$ln\left[\frac{P(A|C)}{1-P(A|C)}\right] \cong ln\left[\frac{P(A)}{1-P(A)}\right] + \sum_{i=1}^{m}\left\{ln\left[\frac{P(A|D_i)}{1-P(A|D_i)}\right] - ln\left[\frac{P(A)}{1-P(A)}\right]\right\} \quad \textbf{(5)}$$

Алгоритм PASS использует частотные оценки вероятностей $P(A_k)$ и $P(A_k|D_i)$ для класса A_k соединений, содержащих активность A_k в спектре активности:

$$\boldsymbol{P(A_k) = \frac{N_k}{N}} \quad \textbf{(6a)}$$

$$\boldsymbol{P(A_k|D_i) = \frac{N_{ik}}{N_i}} \quad \textbf{(6b)}$$

где целые числа N ... N_{ik} это: N – общее количество соединений в SAR Base; N_i – количество соединений, содержащих дескриптор D_i в описании структуры молекул; N_k – количество соединений, содержащих активность A_k в спектре активности; N_{ik} – количество соединений, содержащих и дескриптор D_i в описании структуры молекул, и активность A_k в спектре активности.

Помимо уже отмеченной приближенности подхода, формулы (3), (4) и (5) при использовании частотных оценок вероятностей (6) имеют существенный, хорошо известный недостаток – оценки (6b) могут принимать нулевое значение. Один из наиболее популярных методов преодоления этого недостатка – «коррекция по Лапласу» [46, 47], согласно которому оценки (6) заменяются на следующие:

$$\boldsymbol{P(A_k) = \frac{N_k + \alpha_k}{N + \beta}} \quad \textbf{(7a)}$$

$$\boldsymbol{P(A_k|D_i) = \frac{N_{ik} + \alpha_{ik}}{N_i + \beta_i}} \quad \textbf{(7b)}$$

с положительными числами α и β. В «каноническом» случае должно быть α≡1 и β≡2. По результатам наших исследований использование значений $\alpha_k = \alpha_{ik} = P(A_k)$ и β≡1 (при этом оценки (7a) тождественны оценкам (6a)) дало гораздо лучшие результаты, однако, самый лучший результат даёт арксинусное преобразование Фишера:

$$\boldsymbol{ln\left[\frac{P(A|D_i)}{1-P(A|D_i)}\right] \rightarrow ArcSin\left[2P(A|D_i) - 1\right]} \quad \textbf{(8)}$$

Точность прогноза также повысилась после замены суммы вкладов дескрипторов их средним значением, что, видимо, компенсирует допущение о независимости дескрипторов.

Для рассматриваемой задачи классификации логарифм априорного отношения правдоподобия (4b) не несёт информации о конкретном прогнозируемом органическом соединении и может быть опущен.

Описанный выше подход поясняет, почему алгоритм прогноза PASS основан на следующей специальной B статистике: по структуре молекул химического соединения, записанной в виде множества $\{D_1,..., D_m\}$ из m дескрипторов MNA, для каждой активности A_k подсчитываются величины B_k:

$$\boldsymbol{B_k = (S_k - S_{0k}) / (1 - S_k \cdot S_{0k})} \quad \textbf{(9a)}$$

$$\boldsymbol{S_k = Sin\left[\sum_{i=1}^{m} ArcSin\left(2P(A_k|D_i) - 1\right)/m\right]} \quad \textbf{(9b)}$$

$$\boldsymbol{S_{0k} = 2P(A_k) - 1} \quad \textbf{(9c)}$$

При этом для активности A_k, если для всех дескрипторов $P(A_k|D_i)$=1, то B_k=1; если для всех дескрипторов $P(A_k|D_i)$=0, то B_k=-1; если связи между дескрипторами и активностью A_k нет и $P(A_k|D_i) \approx P(A_k)$, то $B_k \approx 0$.

Результат прогноза биологической активности представляется в PASS в виде вероятностей *Pa* «быть активным» («to be active») и *Pi* «быть неактивным» («to be inactive»). Зависимости, необходимые для получения вероятностей *Pa* и и *Pi* по значениям *B* статистики, и оценки точности прогноза PASS являются конечным результатом **процедуры обучения**, которая состоит в следующем. По данным SAR Base, сформированной на основе обучающей выборки, для каждой активности A_k для каждого из N_k активных и для каждого из N-N_k неактивных соединений вычисляются значения *B* статистики. Вычисления проводятся в режиме скользящего контроля с исключением по одному, то есть после «исключения» этого соединения из SAR Base, для чего достаточно не включать его в суммы. По полученным выборкам *B* статистики строятся гладкие полиномиальные оценки функций *Pa* и *Pi* как описано в [38]. На рисунке 4 представлены *Pa*(*B*) и *Pi*(*B*) для активности «Антигипертензивное».

Из примера на рисунке 4 видно, что значения *Pi* монотонно убывают при возрастании значений *Pa* и сумма *Pa* и *Pi* меньше или равна 1. Вероятности *Pa* и *Pi* являются также, по построению, оценками вероятности ошибок прогноза 1-го и 2-го рода, соответственно, а 1-*Pa* и 1-*Pi* – оценками чувствительности и специфичности. Вероятности *Pa* и *Pi* можно рассматривать и как меры принадлежности прогнозируемого соединения к нечётким множествам «активных» и «неактивных» органических соединений. Все эти интерпретации вероятностей *Pa* и *Pi* эквивалентны и полезны для анализа результатов прогноза. На их основе можно сконструировать самые разные критерии анализа результатов прогноза, соответствующие решению конкретных практических задач.

Точность прогноза каждой активности оценивается в PASS как вероятность того, что для произвольной пары новых активного и неактивного соединений значение *Pa* для активного соединения будет выше, чем значение *Pa* для неактивного соединения, и называется инвариантной точностью прогноза (IAP) [35, 38]. Она эквивалентна критерию площади

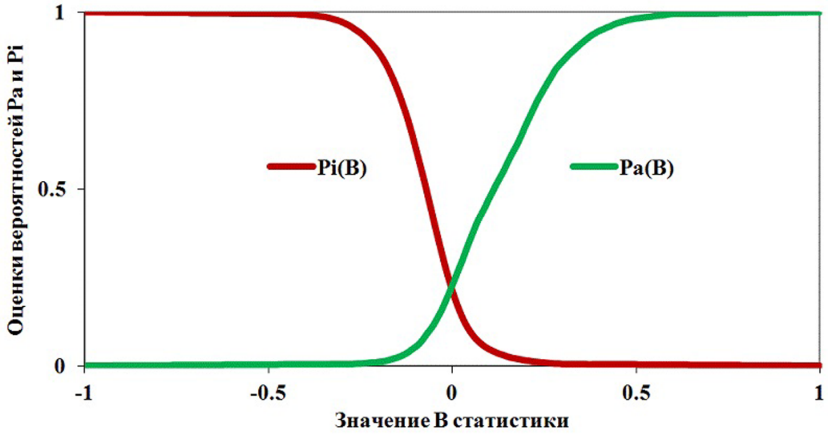


Рисунок 4. Зависимости *Pa*(*B*) и *Pi*(*B*) для активности «Антигипертензивное» на основе данных, представленных в SAR Base PASS версии 2017.

под кривой оперативной характеристики (AUC ROC). На рисунке 5 на примере активности «противоопухолевое» показаны зависимости между *Pa* и *Pi*, чувствительностью, специфичностью, точностью и сбалансированной точностью.

Площадь под кривой зависимости 1-*Pa* от *Pi* (кривой чувствительности), показанной на рисунке 5 чёрным цветом, и есть AUC ROC, совпадающая с IAP [35]. Точка пересечения всех кривых соответствует равенству *Pa* и *Pi*, и, соответственно, равенству вероятностей ошибок первого и второго рода, равенству чувствительности и специфичности, и, примерно, максимуму сбалансированной точности. Значение *Pa* = *Pi* в этой точке равно минимаксной оценке точности прогноза при полном отсутствии

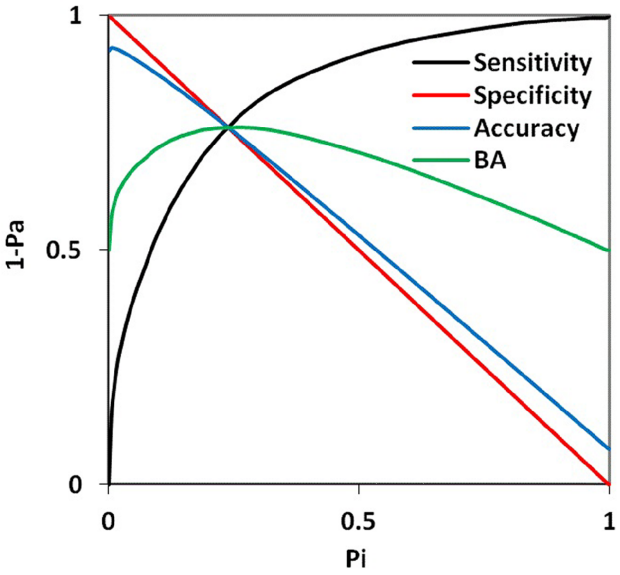


Рисунок 5. Пример зависимостей между чувствительностью («Sensitivity» ≡ 1-*Pa*, чёрная кривая), специфичностью («Specificity» ≡ 1-*Pi*, красная линия), точностью (конкордансом) («Accuracy» (Concordance), синяя кривая), и сбалансированной точностью («BA», Balanced Accuracy, BA = (Sensitivity+Specificity)/2) как функций порога по вероятности ошибок второго рода (*Pi*) для активности «противоопухолевое».

априорной информации как о платежной матрице, так и вероятности встречаемости активности в какой-либо выборке.

Оценка влияния неполноты данных на качество прогноза. Поскольку обучающая выборка не может содержать полной информации о биологической активности включенных в нее соединений (ни одно из химических соединений не исследовано на все возможные виды биологической активности), мы провели специальное исследование [38] с целью оценки влияния неполноты информации в обучающей выборке на качество прогноза. Использовали выборку, содержащую около 19000 веществ из базы данных MDDR (так называемых “Principal Compounds”, для которых в MDDR были приведены экспериментальные данные о биологической активности). 120 различных видов активности было представлено в этой выборке не менее чем 100 соединениями. В ходе компьютерных экспериментов всю выборку 50 раз случайным образом делили на две равные подвыборки, одна из которых использовалась в качестве обучающей, а другая – в качестве тестовой, и наоборот (всего, таким образом, было выполнено по 100 экспериментов). Чтобы смоделировать неполноту информации, из обучающих выборок случайным образом исключали 20, 40, 60, 80% информации о структуре или биологической активности. В ходе обучения рассчитывали средние значения ошибки прогноза. Было показано, что исключение до 60% информации позволяет получать разумные оценки биологической активности для веществ тестовых выборок, то есть алгоритм программы PASS обладает робастностью (статистической устойчивостью) по отношению к неполноте данных в обучающей выборке. В данной работе также продемонстрировано, что оценка точности по скользящему контролю с исключением по одному даже более жёсткая, чем по перекрёстному контролю.

Нами было проведено несколько компьютерных экспериментов по сравнению предсказательной способности PASS с другими, свободно-доступными через Интернет, веб-ресурсами.

В 2008 году отличных от PASS веб-ресурсов в Интернете, прогнозирующих спектры биологической активности веществ, не было обнаружено; поэтому для сравнения качества прогноза мы сопоставили между собой результаты оценки некоторых других характеристик с применением различных методов [48]. Наилучшее согласие между результатами прогноза было получено для величины коэффициента распределения «*н*-октанол–вода» logP (для семи методов коэффициенты корреляции R варьировали от 0.65 до 0.90); менее согласованные между собой результаты были получены для прогноза растворимости в воде (R = 0.01–0.73 для четырёх методов) и параметра «drug-likeness» (R = 0.19–0.73 для трёх методов). Качество прогноза PASS было оценено на основе анализа независимых от авторов программы публикаций: было найдено 15 работ, в которых результаты прогноза были подтверждены в эксперименте для веществ, принадлежащих к различным химическим классам, и разнообразных видов биологической активности [48].

В 2016 году для веществ из тестовой выборки лекарственных препаратов, разрешенных к медицинскому применению в 2014 г., были проведены оценки качества прогноза с использованием четырёх веб-ресурсов [49]. Оказалось, что значения чувствительности S для четырёх рассмотренных методов убывают в следующем порядке: PASS > SuperPred > DRAR-CPI > SwissTargetPrediction (S = 0.95; 0.53; 0.41; 0.37). На основе полученных результатов мы пришли к выводу о преимуществе реализованных в PASS дескрипторов MNA и алгоритма классификации, по сравнению с используемыми в SuperPred и SwissTargetPrediction методах оценки по сходству или поиску ассоциаций на основе молекулярного докинга в DRAR-CPI [49].

В 2017 году было сопоставлено качество прогноза исходных и репозиционированных фармакотерапевтических эффектов с использованием шести доступных в Интернете веб-ресурсов (ChemProt, PASS, SEA, SuperPred, SwissTargetPrediction, TargetHunter) с использованием двух тестовых выборок: 50 репозиционированных лекарств и 12 препаратов, недавно запатентованных по новому назначению [50]. Для первой выборки значения чувствительности варьировали от 0.64 (TarPred) до 1.00 (PASS) для исходных показаний, и от 0.64 (TarPred) до 0.98 (PASS) для репозиционированных показаний. Для второй выборки – от 0.08 (SuperPred) до 1.00 (PASS) для исходных показаний, и от 0.00 (SuperPred) до 1.00 (PASS) для репозиционированных показаний. Таким образом был сделан вывод о «самодостаточности» прогноза PASS и отсутствии необходимости консенсусных прогнозов на основе комбинирования результатов PASS и каких-либо других веб ресурсов [50].

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ ПРОГНОЗА PASS

Пользователь PASS получает результат прогноза спектра биологической активности в виде упорядоченного списка оценок вероятностей *Pa* и *Pi* принадлежности прогнозируемого соединения к классам «активных» и всех прочих соединений, и названий соответствующих активностей. Упорядочение выполняется по убыванию разности *Pa*-*Pi*, так что более вероятные виды активности находятся в верхней части спрогнозированного спектра. Спрогнозированный спектр активности может анализироваться любым желаемым образом, но по умолчанию в него включаются активности, для которых *Pa*>*Pi*.

Необходимо помнить, что вероятность *Pa* отражает, прежде всего, сходство структуры молекул данного органического соединения со структурами молекул, наиболее типичных в соответствующем подмножестве «активных» соединений в обучающей выборке. Поэтому никакой прямой корреляции вычисляемых величин *Pa* с количественными характеристиками активности, как правило, нет. Действительно активное соединение, но имеющее нетипичную для обучающей выборки структуру молекул, может иметь согласно прогнозу низкое значение *Pa*, даже, возможно, *Pa*<*Pi*, поскольку значения величин *Pa* для активных и *Pi* для неактивных соединений из обучающей выборки (подсчитанные с их исключением!) распределены строго равномерно, что следует из способа построения функций *Pa*(*B*) и *Pi*(*B*) [38].

Необходимо также помнить о том, что основное для алгоритма PASS выражение (9b) можно записать в аналогичном (4) виде:

$$S_k = Sin\left[\sum_i a_i x_i(C) / \sum_i x_i(C)\right] \qquad \textbf{(10a)}$$

$$a_i = ArcSin\left(2P\left(A_k|D_i\right) - 1\right) \qquad \textbf{(10b)}$$

откуда все соединения, удовлетворяющие условию $\Sigma_i(a_i\text{-}b)x_i(C)\approx 0$, будут иметь соответствующие $S_k=Sin(b)$ одинаковые значения *Pa* и *Pi*, хотя их структуры могут быть и совсем непохожими друг на друга. Из всего написанного выше следует и интерпретация результатов прогноза.

Если, например, величина *Pa* равна 0.9, то для 90% соединений из обучающей выборки, проявляющих эту активность, значение *B* статистики меньше, чем для исследуемого соединения, и только для 10% – больше. И, соответственно, если мы отклоним предположение о том, что это соединение обладает активностью, то, в среднем, мы с вероятностью 0.9 совершим ошибку. Если же величина *Pa* меньше 0.5, то, следовательно, более половины активных соединений из обучающей выборки имеют значение *B* статистики больше, чем для данного соединения, и если мы отклоним предположение о том, что оно обладает активностью, то совершим ошибку с вероятностью менее 0.5.

Другой важный аспект интерпретации результатов прогноза связан с новизной анализируемого соединения по сравнению с соединениями в обучающей выборке. Результатам прогноза 0.3<*Pa*<0.7 примерно соответствуют наиболее

вероятные структуры активных соединений в SAR Base (наибольший наклон на рисунке 4), хотя в силу (10a), оно и может сильно отличаться от всех соединений с данной активностью в SAR Base, но, вероятнее всего, оно «такое же». Если же Pa>0.7, то шансы обнаружить активность в эксперименте довольно высоки, и соединение, скорее всего, сочетает в себе наиболее важные особенности активных соединений, имеет очень мало общего с остальными соединениями в SAR Base (левый нижний угол на рисунке 5) и даже может оказаться родоначальником нового химического класса для рассматриваемого вида биологической активности.

Ещё один важный аспект интерпретации результатов прогноза состоит в предположении, что характеристики выборки соединений, для которых выполнен прогноз, подобны характеристикам соединений в SAR Base, для которых построены оценки Pa и Pi, что необходимо для применимости имеющихся оценок Pa и Pi для анализа прогноза. Необходимо помнить, что в обычных выборках очень мало активных соединений. Например, в SAR Base PASS (версия 2017) для половины из 5050 прогнозируемых активностей имеется менее 30 активных соединений (в среднем 473). Таким образом, для 2525 видов активности априорная вероятность $P(A)$ менее 0.00003, и в среднем по всем активностям $P(A)$=0.00046. Даже для наиболее «популярных» видов активности она мала – для «Beta-lactamase AmpC inhibitor» (см. рис. 1) $P(A)$=0.02. Эти оценки означают, что в чисто случайной выборке из тысячи соединений априорная вероятность найти хотя бы одно соединение с заданной активностью менее 1/2. Поскольку в SAR Base PASS включаются только те соединения, для которых известна хотя бы одна найденная экспериментально активность, то в общем случае активные соединения встречаются ещё реже.

Если в исследуемой выборке N_1 соединений с желаемой активностью и N_0 неактивных соединений, то по результатам прогноза PASS при заданном пороге Pa условно активными будет признано $N_1(1\text{-}Pa)+N_0Pi$ соединений, среди которых действительно активных всего $N_1(1\text{-}Pa)$ соединений. В силу редкой встречаемости активных соединений даже при высокой точности прогноза вполне возможно, что $N_1(1\text{-}Pa)<<N_0Pi$ – ложноположительных прогнозов гораздо больше, чем истинных.

Совершенно аналогично в SAR Base PASS (версия 2017) в среднем одно соединение имеет в спектре биологической активности менее трёх видов активности, хотя, например, для Топирамата их 239. При прогнозе из 5050 прогнозируемых видов активности будет около сотни активностей с Pa>0.5.

Описанный выше избыток ложноположительных предсказаний – следствие редкой удачи найти активное соединение, однако использование прогноза PASS может в десятки раз сократить объём необходимого экспериментального тестирования по сравнению со слепым поиском.

Обширный спрогнозированный спектр активности свидетельствует о том, что структура молекулы данного органического соединения довольно проста, не содержит каких-либо особенностей, обеспечивающих высокую селективность его биологического действия. Например, если дескрипторов MNA менее 20, то при пороге $Pa>Pi$ результат прогноза PASS (версия 2017) может содержать более тысячи видов активности, тогда как если дескрипторов MNA более 40, то, как правило, он будет включать менее двухсот активностей.

Если при прогнозе оказалось, что в структуре есть несколько новых по отношению к составу обучающей выборки дескрипторов MNA, то структура менее похожа на любую из структур в SAR Base, и результаты прогноза необходимо рассматривать как приблизительные оценки.

При анализе прогнозируемых PASS спектров биологической активности необходимо учитывать реальные возможности экспериментального тестирования. При этом общей рекомендацией является последовательное исследование различных прогнозируемых видов биологической активности, от наиболее вероятных к менее вероятным.

В таблице приведён пример прогноза спектра биологической активности для лекарственного препарата Топирамат (Topiramate).

Согласно сведениям, содержащимся в базе данных Integrity [51], Топирамат имеет следующие фармакотерапевтические показания: Treatment of Bipolar Disorder, Psychiatric Disorders (Not Specified), Treatment of Alcohol Dependency, Prophylactic Treatment of Migraine, Agents for Inflammatory Bowel Disease, Antiobesity Drugs, Antimigraine Drugs, Treatment of Cocaine Dependency, Treatment of Neuropathic Pain, Antiepileptic Drugs, Treatment of Substance Dependency, Treatment of Eating Disorders, Aid to Smoking Cessation, Treatment of Nutritional Disorders, Treatment of Cerebrovascular Diseases, Treatment of Obsessive-Compulsive Disorder (OCD). Как видно из приведённых в таблице результатов прогноза, большая часть этих эффектов успешно прогнозируется PASS.

В базе данных Integrity содержится информация о взаимодействии Топирамата с ферментами метаболизма лекарств CYP3A4 и CYP2C19, что также нашло свое отражение в результатах прогноза.

В Integrity приведена информация о следующих молекулярных механизмах действия Топирамата: Sodium Channel Blockers, Carbonic Anhydrase Type II Inhibitors, AMPA Antagonists, Kainate Antagonists. Блокирование Топираматом натриевых каналов прогнозируется с вероятностью Pa=0.710.

Прогноз указывает на возможность взаимодействия Топирамата с карбоангидразой II (Carbonic anhydrase II stimulant, Pa=0.988; Carbonic anhydrase II inhibitor, Pa=0.563), однако не позволяет прийти к заключению о направлении воздействия (стимуляция или ингибирование). В таких случаях, когда одновременно прогнозируется агонистическое или антагонистическое действие на рецепторы, стимуляция или ингибирование ферментов, открытие или блокада каналов, и т.п., требуется детальное

Таблица. Прогнозируемый на сайте PASS online спектр биологической активности препарата Топирамат (Topiramate, PubChem CID: 5284627) при пороге *Pa*>0.4

Pa	*Pi*	Activity	Known Activities
0.995	0.002	Anticonvulsant	+
0.988	0.000	Carbonic anhydrase II stimulant*	+
0.954	0.001	Antialcoholic	+
0.925	0.001	Bipolar disorder treatment	+
0.923	0.002	CYP2C19 inhibitor	+
0.920	0.002	CYP3A4 inducer	+
0.914	0.002	CYP3A inducer	+
0.907	0.001	Growth stimulant	
0.897	0.003	Ophthalmic drug	
0.893	0.006	CYP3A4 substrate	+
0.882	0.003	Antineurogenic pain	+
0.874	0.007	CYP3A substrate	+
0.862	0.003	Antiglaucomic	
0.855	0.002	Antismoking	+
0.826	0.001	Carbonic anhydrase V inhibitor	
0.810	0.001	Obsessive-compulsive disorder treatment	+
0.812	0.003	GABA receptor agonist	+
0.802	0.004	Antiepileptic	+
0.731	0.001	Carbonic anhydrase IX inhibitor	
0.710	0.004	Sodium channel blocker	+
0.660	0.004	Imidazoline I1 receptor agonist	
0.597	0.004	Antimigraine	+
0.589	0.001	Carbonic anhydrase inhibitor	
0.563	0.001	Carbonic anhydrase II inhibitor*	+
0.564	0.003	Gastric antisecretory	+
0.546	0.016	Antiobesity	+
0.489	0.030	Antiallergic	
0.468	0.026	CYP17 inhibitor	
0.445	0.009	Dependence treatment	+
0.425	0.001	Carbonic anhydrase I inhibitor	
0.463	0.048	Analgesic	+
0.434	0.042	Glucan 1,4-alpha-maltotriohydrolase inhibitor	

Примечание. Звёздочкой помечены потенциально противоречивые предсказания.

изучение характера воздействия в зависимости от дозы (известны ситуации, когда лекарственное вещество проявляет противоположные эффекты при разных дозах). Возможно, именно этим обстоятельством объясняется наличие в прогнозе антиглаукомного действия препарата, одним из известных механизмов которого является ингибирование карбоангидразы II. В то же время установлено, что у отдельных пациентов применение Топирамата приводит к возникновению глаукомы [52]. В прогнозе побочных эффектов PASS Online также имеется указание на возможность возникновения глаукомы.

Такие молекулярные механизмы действия Топирамата, как AMPA и Kainate Antagonists PASS не прогнозирует, что указывает на невысокое структурное сходство данного препарата с наиболее типичными молекулами, имеющими эти виды активности, в обучающей выборке PASS.

Для Топирамата также прогнозируется взаимодействие с ГАМК рецепторами, не указанное в Integrity [51], однако в литературе имеются указания на наличие такой биологической активности у препарата [53].

Для прогнозируемого эффекта Топирамата на имидазолиновые рецепторы (Imidazoline II receptor agonist), и антиаллергического действия (Antiallergic) нам не удалось найти экспериментального подтверждения в литературе. Эти виды биологической активности, наряду с возможным действием на другие формы карбоангидразы (Carbonic anhydrase V inhibitor, Carbonic anhydrase IX inhibitor, Carbonic anhydrase I inhibitor), целесообразно протестировать в эксперименте.

Перечень прогнозируемых PASS видов биологической активности включает в себя как фармакотерапевтические эффекты, так и механизмы действия, что существенным образом отличает данный подход от других упомянутых выше методов (SEA, TarPred и др.), которые, в основном, предсказывают действие вещества на молекулярные мишени. Это позволяет использовать PASS для решения различных задач при поиске и создании новых лекарственных препаратов:

- идентифицировать, какие соединения следует синтезировать в первую очередь или отобрать наиболее перспективные соединения среди доступных образцов, как потенциально обладающие требуемым видом биологической активности;
- определить, какие тесты являются наиболее релевантными для выборки соединений из конкретного химического класса;
- найти новые соединения с требуемым набором видов биологической активности среди доступных образцов из собственных или коммерчески доступных баз данных;
- выявить новые эффекты и/или механизмы действия для известных фармакологических веществ.

Рассмотрим некоторые примеры практического применения программы PASS в поиске и разработке новых фармакологических веществ.

ПРИМЕРЫ ПРИМЕНЕНИЯ PASS

Отбор на основе прогноза PASS наиболее перспективных соединений для синтеза и биологического тестирования. В рамках международного проекта INTAS был выполнен прогноз анксиолитического действия 5494 виртуальных структур из различных химических классов (тиазолы, пиразолы, изатины, имидазолы и др.), на основе которого было отобрано 8 наиболее перспективных соединений для синтеза и тестирования целевой активности [54]. Соединения были синтезированы и исследованы по стандартным фармакологическим тестам на лабораторных животных. Шесть из восьми исследованных соединений проявили анксиолитическую активность на уровне или выше препарата сравнения Медазепам. Структурные формулы пяти исследованных соединений существенно отличались от структуры известных анксиолитиков, что позволило отнести их к классу NCE (New Chemical Entities), то есть соединений, относящихся к химическим классам, в которых анксиолитическая активность ранее не была установлена [54].

С использованием прогноза PASS на основе виртуального скрининга химической библиотеки, содержащей 2648 органических молекул, было отобрано 32 «хита», для которых прогнозировалось ингибирование ксантиноксидазы, как потенциальных препаратов для лечения гиперурикемии [55]. 24 соединения были доступны в виде синтезированных образцов, и для них был проведен молекулярный докинг с помощью программы Glide XP (Schrodinger) по отношению к центру взаимодействия ксантиноксидазы с пираксостатом (идентификатор 3D структуры в PDB – 1VDV). Все 24 соединения были протестированы *in vitro*; обнаружено 3 ингибитора ксантиноксидазы; наиболее активное соединение имело значение IC_{50}=1.4 мкМ (для препарата сравнения, Аллопуринола, значение IC_{50}=5.7 мкМ). Анти-гиперурикемическое действие найденных соединений было подтверждено в экспериментах *in vivo* на крысах. Таким образом, авторы существенно сузили «пространство поиска» и докировали менее 1% соединений из всей химической библиотеки, отобранных на основе прогноза PASS. В конечном итоге, вместо полного скрининга всей библиотеки из 2648 соединений было протестировано 24 молекулы и найдено 3 активных вещества, то есть вероятность выявления активного соединения составила 12.5%, в то время как при случайном скрининге она составляет менее 1% [56].

Установление наиболее релевантных тестов для конкретных соединений. Несколько примеров практического применения PASS приведено в публикации Е.В. Бабаева [57].

Так, для новых производных 1-амино-4-(1,3-азолил-2)бутадиенов-1,3 наиболее высокой была вероятность противомикробной активности, причём прогноз был однотипен для всех синтезированных веществ. Антимикробное действие было подтверждено в экспериментах на грамположительных (*Staphylococcus Aureus*) и грамотрицательных (*Escherichia coli*) микроорганизмах.

В ходе изучения новых химических превращений были получены ранее неизвестные индолизины с донорными заместителями. Хотя по своей структуре полученные соединения напоминали психотропные индолы псилоцинового ряда, для некоторых молекул PASS прогнозирует с высокой вероятностью связывание с бета-2 адренорецепторами. Прогнозируемая активность была подтверждена в эксперименте на мембранах синаптосом мозга крыс.

Согласно прогнозу PASS, синтезированные по оригинальной методике новые производные имидазолов могли обладать антипротозойной активностью и быть эффективны против тропической лихорадки — лейшманиоза. Тестирование антилейшманиозной активности, проведённое в Университете г. Карачи (Пакистан), показало, что активность полученных соединений не уступает антилейшманиозному действию стандартного препарата Амфотерицина, недостатком которого является высокая токсичность [57].

Таким образом, приведённые выше примеры убедительно демонстрируют, каким образом предсказания PASS могут быть использованы для выбора наиболее перспективных тестов для изучения биологической активности конкретных соединений.

Выявление новых соединений с требуемым набором видов биологической активности. С целью поиска новых антигипертензивных веществ, обладающих дуальными механизмами действия,

с использованием PASS было выполнено прогнозирование ассоциированных с этим эффектом 30 молекулярных механизмов действия для 183462 молекул из баз данных компаний AsInEx и ChemBridge [58]. Тестирование *in vitro* четырёх соединений, для которых было предсказано ингибирование ангиотензинконвертирующего фермента (ACE) и нейтральной эндопептидазы (NEP), подтвердило наличие у них прогнозируемых видов активности (для ACE значения $IC_{50} = 10^{-7}$-10^{-9} M, для NEP значения $IC_{50} = 10^{-5}$ M). Также на основе прогноза были выявлены вещества с такими дуальными механизмами действия, комбинации которых ранее не были описаны в литературе.

На основе компьютерного прогноза ингибирования циклооксигеназ (COX) и липоксигеназы (LOX) для 573 виртуальных структур были отобраны для синтеза и биологического тестирования соединения, согласно прогнозу обладающие дуальными механизмами противовоспалительного действия [59]. Для синтеза и экспериментального тестирования было отобрано 9 производных 2-(тиазол-2-иламино)-5-фенилиден-4-тиазолидинона. Восемь изученных соединений проявили активность на широко используемой модели карагинанового воспаления; при этом 7 соединений ингибировали LOX, 7 соединений ингибировали COX, и 6 соединений ингибировали оба фермента [59].

Таким образом, продемонстрировано, что на основе анализа прогнозируемых с помощью PASS спектров биологической активности можно отбирать соединения, обладающие целевыми комбинациями механизмов действия и требуемыми фармакологическими эффектами.

Выявление новых эффектов или механизмов действия для известных веществ. Компанией «Oriflame Skin Research Institute» (Швеция) было показано, что ацетил аспарагиновая кислота (AAA) проявляет омолаживающие кожу свойства (anti-ageing action), однако механизм этого действия оставался неясным [60]. Прогноз спектра биологической активности с использованием PASS показал, что AAA может стимулировать регенерацию кератиноцитов благодаря ингибированию действия матричных металлопротеиназ 1-3 и экспрессии F-актина. Проведённые *in vitro* исследования полностью подтвердили результаты прогноза PASS, несмотря на то, что указанные выше виды биологической активности прогнозировались с невысокими вероятностями.

Прогноз спектров биологической активности для разрешённых к медицинскому применению лекарственных препаратов позволил предположить, что ряд антигипертензивных средств, ингибиторов ангиотензинконвертирующего фермента, включая Каптоприл, Эналаприл, Рамиприл и др., может обладать ноотропным действием [61]. Ноотропную активность трёх препаратов этого класса исследовали на мышах по тесту спонтанной ориентации (поведения патрулирования) в крестообразном лабиринте. Было показано, что Периндоприл в дозе 1 мг/кг, а Квинаприл и Моноприл в дозе 10 мг/кг вызывают улучшение показателей поведения патрулирования лабиринта, сходным образом с эффектами референсных ноотропных препаратов Пирацетама и Меклофеноксата (в дозах 300 мг/кг и 120 мг/кг соответственно). Установленное ноотропное действие некоторых ингибиторов АПФ, скорее всего, не связано с их антигипертензивным эффектом, поскольку ноотропное действие имело место лишь при относительно малых дозах периндоприла, квинаприла и моноприла, и исчезало при дальнейшем увеличении дозы. Выявление ноотропных свойств у антигипертензивных препаратов открывает возможности для их нового применения в медицинской практике, что было впоследствии подтверждено в клинике [62].

Обзор некоторых публикаций с результатами прогнозов биологической активности, полученных с использованием программы PASS, приведён в работах [36, 37].

ОТВЕТЫ НА ЧАСТО ЗАДАВАЕМЫЕ ВОПРОСЫ ПОЛЬЗОВАТЕЛЕЙ PASS

Выбор пороговых значений для дифференциации активных и неактивных молекул. В разделе про интерпретацию результатов прогноза PASS подчеркнуто, что в SAR Base PASS (версия 2017) для половины активностей априорная вероятность $P(A)$ менее 0.00003 и 0.00046 в среднем по всем активностям. Это может служить ориентиром ожидаемого успеха при слепом поиске активных соединений – необходимо выполнить биологические испытания до десятков тысяч соединений для обнаружения хотя бы одного активного соединения. Использование прогноза PASS может сократить необходимые объёмы экспериментальных исследований в десятки раз. Но выбрать одно какое-то пороговое значение *Pa* или *Pi* невозможно – мы рекомендуем следовать стратегии последовательного испытания соединений (активностей) в порядке убывания значений *Pa* (*Pa*-*Pi*) в результатах прогноза, при таком подходе будет самой высокой ***вероятность достижения первого успеха***.

Выше мы уже отмечали, что вероятность наличия активности *Pa* отражает сходство структуры молекул анализируемого вещества со структурами молекул «наиболее типичных» в соответствующем подмножестве активных соединений в обучающей выборке. Несмотря на то, что мы стараемся поддерживать актуальное состояние обучающей выборки, постоянно проводя поиск новой и уточнение имеющейся в обучающей выборке информации, очевидно, что размерность пространства известных биологически активных соединений существенно уступает размерности теоретически возможных органических молекул.

Чтобы проиллюстрировать, каким образом недостаток в обучающей выборке информации о конкретных химических классах соединений может повлиять на качество прогноза, были проведены специальные компьютерные эксперименты. Для этого мы извлекли из базы данных Integrity [51]

данные о соединениях, удовлетворяющих запросу: “Hypertension treatment” {*Therapeutic Group*} AND “Chemical Categories” {*Product Category*} AND “To 1250” {*Molecular Weight*}. Таких документов оказалось 1924. После фильтрации с помощью разработанной нами процедуры “CheckSDF” на основе вышеописанных критериев PASS для структур соединений, их осталось 1762: 162 документа было отсеяно, поскольку 112 записей содержали два и более компонентов; 45 записей содержали заряженные молекулы; 4 записи содержали молекулы, у которых было более разрешённого числа аминокислотных остатков; 1 запись содержала символ атома, не соответствующий таблице Менделеева. После проведения процедуры обучения программы PASS в базе знаний SAR Base оказалось 1655 молекул, имеющих 3967 разных дескрипторов MNA 1-го и 2-го уровней, с 809 наименованиями биологической активности. После проведения селекции мы оставили 78 видов активности, ассоциированных с антигипертензивным действием; при этом среднее качество прогноза по критерию IAP, оцененное по процедуре скользящего контроля с исключением по одному, составило 0.9840. Точность прогноза для отдельных видов активности варьировала от 0.8898 до 1.0000. В процедуре скользящего контроля со случайным разбиением выборки на 20 частей среднее по активностям значение IAP равно 0.9841, что свидетельствует о достаточно высокой точности и предсказательной способности полученной нами SAR Base. В дальнейшем мы сконцентрировались на единственной активности “Antihypertensive”.

С помощью имеющихся в Integrity возможностей информационного поиска мы проанализировали, какие химические классы (Product Category) составляют полученную нами выборку. Оказалось, что около половины молекул относятся к классу “Oligopeptides, less than 10 AA” (олигопептиды, содержащие менее 10 аминокислотных остатков). Далее мы разбили эту обучающую выборку на две подвыборки: первая включала в себя все молекулы, не отнесенные к классу “Oligopeptides, less than 10 AA” и была использована в качестве обучающей, а вторая, содержащая такие олигопептиды, была использована в качестве тестовой выборки. После обучения оказалось, что в SAR Base включено 792 молекулы, и для антигипертензивной активности IAP=0.8791.

Прогноз антигипертензивной активности на основе исходной обучающей выборки, содержащей все извлечённые из Integrity молекулы, при пороге *Pa*>*Pi* позволяет правильно классифицировать 889 из 939 молекул (~95%). На основе новой обучающей выборки, не содержащей молекул олигопептидов, прогноз антигипертензивной активности при пороге *Pa*>*Pi* позволяет правильно классифицировать 752 из 939 молекул (~80%). Среднее значение *Pa* для предсказанных активными молекул в первом случае составило 0.675, в то время как во втором случае это значение снизилось до 0.394.

Ещё более разительные отличия наблюдались при исключении из обучающей выборки веществ, отнесенных к категории “Nucleosides”. В этом случае после обучения SAR Base содержала 1508 молекул, а для антигипертензивной активности IAP=0.9354. Прогноз антигипертензивной активности на основе исходной обучающей выборки, содержащей все извлечённые из Integrity молекулы, при пороге *Pa*>*Pi* позволяет правильно классифицировать 136 из 166 молекул (~82%). Прогноз антигипертензивной активности на основе новой обучающей выборки, не содержащей молекулы нуклеозидов, при пороге *Pa*>*Pi* позволяет правильно классифицировать 89 из 166 молекул (~54%). При этом среднее значение Pa для предсказанных активными молекул в первом случае составило 0.332, в то время как во втором случае это значение снизилось до 0.245.

Полученные в ходе этих компьютерных экспериментов результаты указывают на существенную зависимость получаемых при прогнозе оценок значений *Pa* от того, насколько структура анализируемого соединения сходна со структурой включенных в обучающую выборку веществ, имеющих конкретный вид активности. Поскольку PASS прогнозирует несколько тысяч видов биологической активности и для каждого из них величины сходства анализируемого соединения с соответствующими соединениями обучающей выборки могут отличаться, ***нельзя предложить единый универсальный критерий для выбора «порога»***, позволяющего по результатам прогноза PASS отделить «активные» соединения от «неактивных». Именно поэтому «по умолчанию» в PASS принято пороговое значение *Pa*>*Pi*, особенности которого можно видеть на рисунке 5.

Мы рекомендуем использовать результат прогноза для референсного соединения (применяемого в фармакологических исследованиях препарата сравнения) как реперной точки. Так, например, в случае антигипертензивных препаратов олигопептидной природы референсным соединением может быть разрешенный к медицинскому применению препарат Moexipril Hydrochloride [51]. Как видно из результата прогноза PASS на рисунке 6, при использовании SAR Base, не содержащей олигопептидов, для этой молекулы значение *Pa* равно 0.299. Если выбрать это значение в качестве порогового, то при использовании данной SAR Base из 752 соединений, для которых *Pa*>*Pi*, можно будет выявить 466 с *Pa*>0.299.

Дополнительные погрешности при рассмотрении прогностических оценок могут вносить так называемые “Activity cliffs”, когда сравнительно небольшие изменения структуры приводят к резкому падению (либо возрастанию) величины активности [63, 64]. Очевидно, что существующие методы анализа зависимостей «структура-активность», как правило, основаны на предположении о существовании «гладких зависимостей», что не позволяет в большинстве случаев идентифицировать такого рода ситуации.

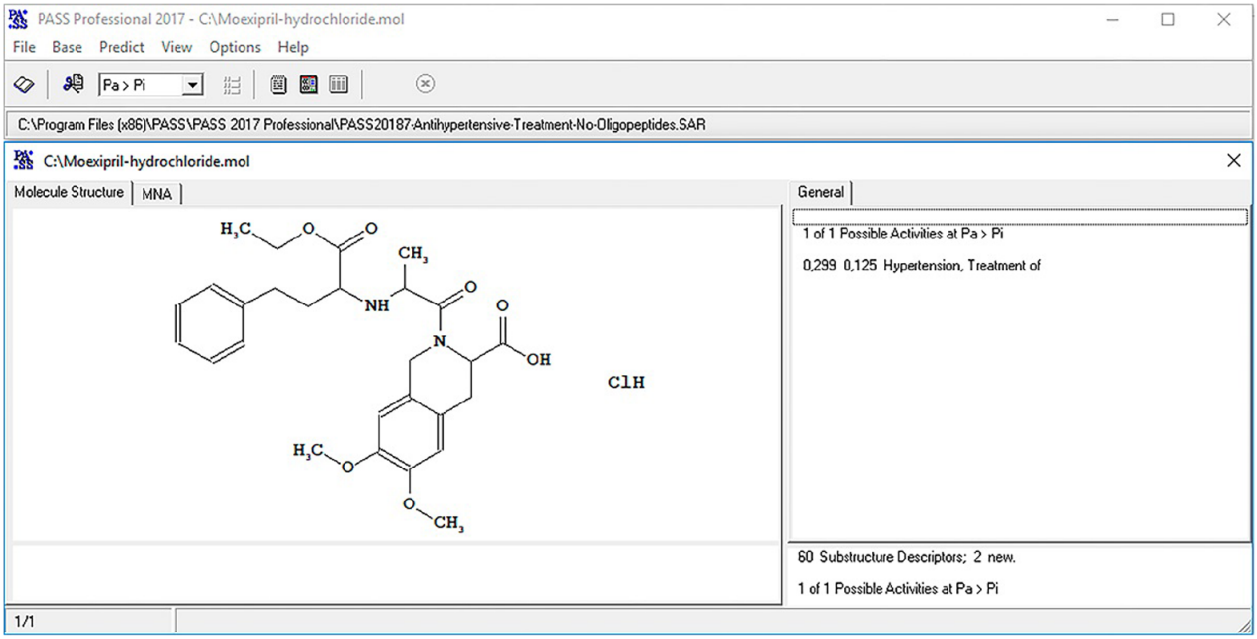


Рисунок 6. Пример выбора пороговых значений *Pa* на основе прогноза антигипертензивной активности для референсного препарата Moexipril Hydrochloride.

В этой связи, даже если в прогнозе нет интересующей пользователя биологической активности, но он по каким-либо причинам уверен, что такая активность должна присутствовать, имеет смысл проверить экспериментально её наличие.

Ограничение перечня прогнозируемых видов активности. В программе PASS реализована процедура селекции интересующих пользователя видов биологической активности ещё до проведения прогноза. Это дает возможность снизить временные затраты на анализ результатов прогноза, заранее ограничив их либо интересующей фармакотерапевтической областью, либо теми видами активности, для которых пользователь может организовать проведение экспериментального тестирования. Выбор интересующих видов активности может быть выполнен либо вручную, по одному, либо путем загрузки списка целевых видов активности. Так, например, на рисунке 7 представлена загрузка списка видов активности, ассоциированных с антигипертензивным действием.

В результате, вместо 5050 видов биологической активности, прогнозируемых по умолчанию PASS (2017) со средней точностью IAP=0.9652, пользователь получит прогноз 93 интересующих его видов активности с чуть более высокой средней точностью IAP=0.9738.

Границы применимости прогноза PASS. Возможности PASS ограничены перечнем прогнозируемых видов активности, который составлен с учётом современного состояния фармакологической науки, и источниками данных о результатах экспериментального исследования органических соединений на биологическую активность. Создание новых биологически активных веществ – динамично развивающаяся область, постоянно претерпевающая как количественные, так и качественные изменения. Появляются новые термины, характеризующие биологическую активность, например, “Drugs Modulating Gene Expression”, “Transcription Factor Ligands”, “Translation Initiation Factor Inhibitors” [51] или “Modulators of Alternative Splicing” [65], и др. Усложняется наше понимание связей между молекулярными механизмами действия и вызываемыми ими фармакотерапевтическими эффектами, что привело к появлению и развитию «сетевой фармакологии» [66]. Соответственно, необходимы постоянные усилия, направленные не только на то, чтобы пополнять обучающую выборку PASS новыми данными о структуре и биологической активности органических соединений, но и существенным образом уточнять понятийный аппарат описания химико-биологических взаимодействий в ряду «лиганд – мишень – биологический процесс – болезнь» [67].

Кроме того, необходимо отметить, что критерии отнесения соединений к «активным» и «неактивным» также изменяются со временем. В случае исследования новых мишеней, для которых лиганды либо не известны, либо их активность сравнительно невысока, соединения со значениями IC_{50}<100 мкМ могут рассматриваться как активные. Если для рассматриваемой мишени уже известны вещества, действующие в микромолярных или даже субмикромолярных концентрациях, этот порог снижается до IC_{50}<10 мкМ, или даже до IC_{50}<1 мкМ. Сложности интерпретации количественных данных об активности также связаны с различием методик экспериментального тестирования [68-70]. Отчасти вносимые этими факторами погрешности компенсируются статистической устойчивостью используемого в PASS подхода [38].

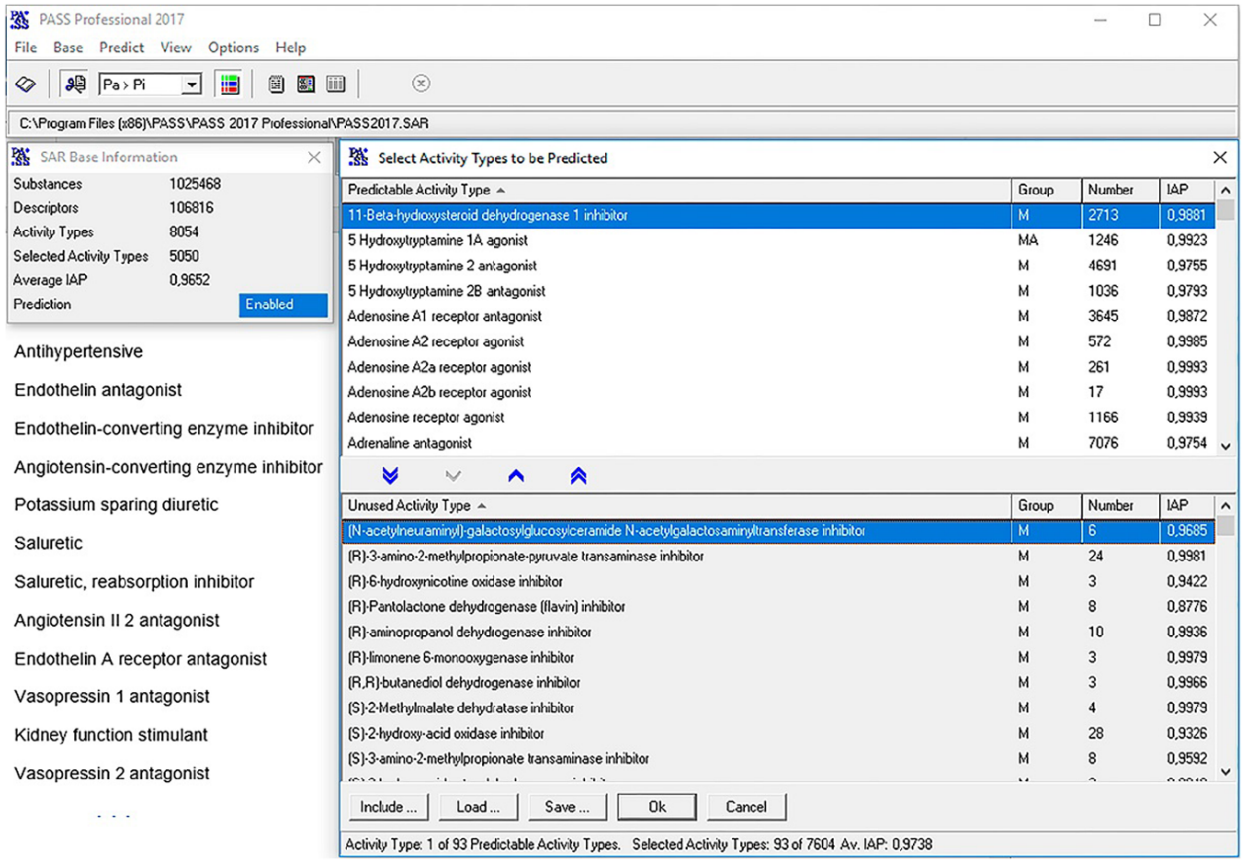


Рисунок 7. Ограничение прогноза спектра биологической активности только перечнем механизмов и эффектов, связанных с антигипертензивным действием.

Необходимо помнить, что PASS прогнозирует возможность проявления биологической активности конкретным соединением, однако не позволяет сделать каких-либо умозаключений относительно величины активности и условий экспериментального тестирования (доза, путь введения, биологический объект, пол, возраст и т.п.), при которых эта активность может проявиться. Таким образом, PASS позволяет сузить область экспериментального тестирования в отношении конкретных соединений, однако любой прогноз необходимо подтверждать экспериментом.

Особую осторожность следует проявлять при интерпретации прогнозируемых PASS побочных или токсических эффектов, поскольку эти эффекты могут не только проявляться при существенно более высоких, в сравнении с терапевтическими, дозах, но также могут наблюдаться у сравнительно небольших групп пациентов (известно, что многие побочные эффекты возникают вследствие идиосинкратических реакций на приём лекарств) [71].

Следует также подчеркнуть, что PASS не может предсказать, станет ли конкретное вещество лекарственным препаратом, поскольку это зависит от ряда различных факторов. Предсказание, однако, может помочь определить, на какие виды биологической активности следует протестировать анализируемое соединение в первую очередь, и какие вещества с наибольшей вероятностью могут проявить требуемые виды активности.

Необходимость нормализации структуры молекул до выполнения прогноза. Как отмечалось выше, во всех современных исследованиях, направленных на компьютерный анализ зависимостей «структура-активность», необходимо предварительно «нормализовать структуру» (убрать солевой компонент, нейтрализовать заряды, заменить координационные связи простыми и т.д.) [42-44]. Одной из широко используемых для нормализации структуры молекул является компьютерная программа Standardizer фирмы ChemAxon [72]. Не все применяемые при этом операции интуитивно понятны химикам-синтетикам, однако их выполнение необходимо для того, чтобы обеспечить однородность представления структурной химической информации как в обучающей выборке PASS, так и в структурах, направляемых на прогноз их биологической активности.

Как уже говорилось выше, в PASS имеются некоторые дополнительные ограничения на анализируемые структуры: наличие не менее трёх атомов углерода и молекулярный вес, не превосходящий 1250 а.е.м. Эти ограничения связаны с необходимостью обеспечить соответствие между молекулами, направляемыми на прогноз, и молекулами, содержащимися в обучающей выборке (прогноз должен выполняться для молекул, попадающих в область применимости зависимостей «структура-активность», представленных в SAR Base).

ПРИМЕРЫ НЕТОЧНОСТЕЙ ПРИ ИНТЕРПРЕТАЦИИ ПРОГНОЗА PASS

Отсутствие экспериментального подтверждения прогноза. Наиболее типичной ситуацией, с которой приходится сталкиваться при рассмотрении некоторых опубликованных работ, является отсутствие экспериментального подтверждения прогнозируемых PASS видов биологической активности. Это нельзя назвать «ошибкой», поскольку, по-видимому, у исследователя просто нет возможности проведения соответствующих экспериментальных исследований. В то же время, ошибочным является утверждение, что вещества, для которых получен прогноз, обладают прогнозируемыми видами активности. Так, например, в работе [73] осуществлён синтез N-(2,5-диметил-4-нитрофенил)-4-метилбензолсульфонамида (NDMPMBS), для которого выполнен прогноз спектра биологической активности с использованием PASS. Ряд активностей прогнозируется с достаточно высокой вероятностью, включая Arylsulfate sulfotransferase inhibitor (*Pa*=0.889), Polyporopepsin inhibitor (*Pa*=0.888), Glutamyl endopeptidase II inhibitor (*Pa*=0.860), Phospholipid-translocating ATPase inhibitor (*Pa*=0.850), и др. Авторы делают вывод о перспективности синтезированного вещества для использования в фармацевтических целях («Results provided … indicate great potential of the newly synthesized NDMPMBS molecule for application in pharmaceutical applications»). Очевидно, что это – слишком оптимистичное утверждение, которое не обосновано полученными в цитируемой работе результатами.

«Подтверждение» прогноза PASS на основе молекулярного докинга. Иногда встречаются публикации, в которых для некоторых молекулярных мишеней, взаимодействие с которыми прогнозируется PASS, проводится молекулярный докинг, по результатам которого утверждается, что изучаемые соединения обладают конкретным видом биологической активности. Так, например, в работе [74], авторы выполнили с помощью PASS прогноз спектров биологической активности для некоторых алкалоидов из растений рода *Strychnos*. Одним из прогнозируемых для стризонобразилина (Strychnobrasiline) видов активности является противоопухолевое действие (*Pa*=0.396). Используя AutoDock-Vina, авторы выполнили докинг этого вещества к комплексу ДНК с топоизомеразой II, проанализировали возможности взаимодействия стризонобразилина с этой мишенью. Аналогичный докинг был выполнен для 12-гидрокси-10,11-стризонобразилина, в результате чего авторы пришли к заключению о более высокой перспективности этого производного по сравнению с исходной молекулой (интересно, что для 12-гидрокси-10,11-стризонобразилина противоопухолевое действие прогнозируется с более высокой вероятностью *Pa*=0.622). В то же время, эти умозаключения требуют экспериментальной проверки, которая в цитируемой работе [74] не была осуществлена.

Интересно, что схожие выводы были сделаны авторами работы [75], которые на основе прогноза с помощью Swiss Target Prediction выбрали в качестве молекулярной мишени серотониновый транспортер. Докинг с использованием AutoDock 4.0, по мнению авторов цитируемой работы, позволяет прийти к заключению, что молекула 3-(1,8-дихлоро-9,10-дигидро-9,10-этаноантрацен-11-yl)акрилальдегида является эффективным антидепрессантом («Docking study indicated that compound 2 is a good antidepressant-like compound»). Понятно, что это умозаключение не вполне корректно, учитывая известные ограничения методов докинга [13].

Рассмотрение полного спектра биологической активности в качестве «подтверждения» перспективности изучаемых веществ. Как было указано выше, для «простых» веществ, не имеющих существенных структурных особенностей, прогнозируемый спектр биологической активности может оказаться чрезвычайно широким. Так, например, в работе [76] приведён прогнозируемый спектр биологической активности для бета-элемена, включающий 629 видов биологической активности (последний в списке – Retinoic acid receptor antagonist, для которого значение *Pa*=0.022). Несмотря на то, что для отдельных предсказанных видов активности автор приводит литературные данные, подтверждающие их наличие, вывод относительно перспективности дальнейшего исследования этого соединения в качестве фармакологического вещества может оказаться неоправданно оптимистичным, поскольку относительно простые молекулы, действительно могут связываться со многими молекулярными мишенями в биологических системах, но с низкими значениями аффинности и специфичности [10].

Таким образом, при анализе результатов предсказаний PASS необходимо учитывать сложность понятия биологической активности и возможные неоднозначности установления зависимостей «структура-активность», что требует достаточно высокой квалификации исследователя, и, несомненно, верификации прогноза в эксперименте.

ЗАКЛЮЧЕНИЕ

В настоящей работе мы рассмотрели используемый в PASS подход к прогнозированию биологической активности, основанный на анализе информации около 1 млн органических соединений с установленной биологической активностью, и привели рекомендации по корректной интерпретацию результатов предсказаний.

Необходимо подчеркнуть, что эти рекомендации применимы как к стандартной версии PASS, прогнозирующей несколько тысяч видов биологической активности [36, 77], так и к специализированным версиям программы: PASS Targets [78, 79], DIGEP Pred [80, 81], PASS CLC Pred [82-84], SMP [85, 86], SOMP [87, 88], RA [89, 90], MetaTox [91-93], ADVER-Pred [94, 95], ROSC-Pred [96, 97], KinScreen [98].

Также необходимо указать, что если полученные пользователем результаты прогноза представляются ему не соответствующими известным литературным (или персональным) данным, у него имеется

возможность добавления веществ соответствующего химического класса к обучающей выборке PASS с использованием веб ресурса SAR Creator [99]. Поскольку периодически мы осуществляем обновление SAR Base, проводя заново процедуру обучения, в следующих версиях программы PASS прогностические возможности для соединений данного химического класса будут расширены.

Если и после обновления SAR Base прогноз будет неудовлетворительным, у пользователя будет возможность построения (Q)SAR моделей на основе подготовленной им обучающей выборки с использованием специализированных программ, например, программы GUSAR, основанной на применении QNA дескрипторов и самосогласованной регрессии [100-110].

БЛАГОДАРНОСТИ

Работа выполнена в рамках Программы фундаментальных научных исследований государственных академий наук на 2013-2020 годы. Авторы выражают искреннюю признательность компании Clarivate Analytics за предоставление лицензии на доступ к базе данных Integrity и компании ChemAxon за предоставление лицензии на JChem.

ЛИТЕРАТУРА

1. Barenboim, G.M. & Malenkov, A.G. (1986) Biologically active substances. New techniques of discovery. Moscow: Science (Rus).

2. Czerepak, E. & Ryser, S. (2008) Drug approvals and failures: implications for alliances. Nature Reviews Drug Discovery, 7, 197-198. DOI: 10.1038/nrd2531

3. ChemNavigator, Retrieved March 24, 2018, from http://www.chemnavigator.com/

4. CAS. Retrieved March 24, 2018, from http://www.cas.org/

5. SAVI. Retrieved March 24, 2018, from https://cactus.nci.nih.gov/download/savi_download/

6. Ruddigkeit, L., Blum, L. C. & Reymond, J.-L. (2013) Visualization and virtual screening of the chemical universe database GDB-17. Journal of Chemical Information and Modeling, 53(1), 56-65. DOI: 10.1021/ci300535x

7. Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I. & Overington, J. P. (2017) A comprehensive map of molecular drug targets. Nature Reviews Drug Discovery, 16(1), 19-34. DOI: 10.1038/nrd.2016.230

8. Li, Y. H., Yu, C. Y., Li, X.X., Zhang, P., Tang, J., Yang, Q., Fu, T., Zhang, X., Cui, X., Tu, G., Zhang, Y., Li, S., Yang, F., Sun, Q., Qin, C., Zeng, X., Chen, Z., Chen, Y. Z. & Zhu, F. (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of' targeted therapeutics. Nucleic Acids Research, 46(D1), D1121-D1127. DOI: 10.1093/nar/gkx1076

9. Chisholm-Burns, M.A., Schwinghammer, T.L., Wells, B.G., Malone, P.M., DiPiro, J.T., & Kolesar, J.M. (2015). Pharmacotherapy Principles and Practice, Fourth Edition. NY: McGraw Hill Professional.

10. Lipinski, C. & Hopkins, A. (2004). Navigating chemical space for biology and medicine. Nature, 432, 855-861. DOI: 10.1038/nature03193

11. Ivanov, A.S., Poroikov, V.V., & Archakov, A.I. (2003) Bioinformatics: way from genome to drug *in silico*. Bulletin of RSMU, 4(30), 19-23 (Rus)..

12. Jorgensen, W.L. (2004) The many roles of computation in drug discovery. Science, 303(5665), 1813-1818. DOI: 10.1126/science.1096361

13. Chen, Y.C. (2015) Beware of docking! Trends in Pharmacological Science, 36(2), 78-95. DOI: 10.1016/j.tips.2014.12.001

14. Luo, H., Zhang, P., Cao, X. H., Du, D., Ye, H., Huang, H., Li, C., Qin, S., Wan, C., Shi, L., He, L. & Yang, L. (2016) DPDR-CPI, a server that predicts drug positioning and drug repositioning via chemical-protein interactome. Scientific Reports, 6, 35996. DOI: 10.1038/srep35996

15. Martin, Y.C., Kofron, J.L. & Traphagen, L.M. (2002) Do structurally similar molecules have similar biological activity? Journal of Medicinal Chemistry, 45(19), 4350-4358. DOI: 10.1021/jm020155c

16. Bender, A. (2010) How similar are those molecules after all? Use two descriptors and you will have three different answers. Expert Opinion on Drug Discovery, 5(12), 1141-1151. DOI: 10.1517/17460441.2010.517832

17. PubChem. Retrieved March 24, 2018, from https://pubchem.ncbi.nlm.nih.gov/

18. ChEMBL. Retrieved March 24, 2018, from https://www.ebi.ac.uk/chembl/

19. DrugBank. Retrieved March 24, 2018, from https://www.drugbank.ca/

20. ChemProt. Retrieved March 24, 2018, from http://potentia.cbs.dtu.dk/ChemProt/

21. SEA. Retrieved March 24, 2018, from http://sea.bkslab.org/

22. SuperPred. Retrieved March 24, 2018, from http://prediction.charite.de/

23. SwissTargetPrediction. Retrieved March 24, 2018, from http://www.swisstargetprediction.ch/

24. TarPred. Retrieved March 24, 2018, from http://202.127.19.75:5555/

25. TargetHunter. Retrieved March 24, 2018, from http://www.cbligand.org/TargetHunter/

26. Mervin, L.H., Afzal, A.M., Drakakis, G., Lewis, R., Engkvist, O. & Bender, A. (2015) target prediction utilising negative bioactivity data covering large chemical space. Journal of Cheminformatics, 7, 51. DOI: 10.1186/s13321-015-0098-y

27. Burov, Yu.V., Poroikov, V.V., Korolchenko, L.V. (1990) National system for registration and biological testing of chemical compounds: facilities for new drugs search. Bulletin of National Center for Biologically Active Compounds, 1, 4-25 (Rus)

.28. Poroikov, V.V., Filimonov, D.A. & Boudunova, A.P. (1993) Comparison of the Results of Prediction of the Spectra of Biological Activity of Chemical Compounds by Experts and the PASS System. Automatic Documentation and Mathematical Linguistics, 27 (3), 40-43.

29. Filimonov, D.A., Poroikov V.V., Karaicheva, E.I., Kazaryan, R.K., Boudunova, A.P., Mikhailovskiy, E.M., Rudnitskih, A.V., Goncharenko, L.V. & Burov, Yu.V. (1995) Computer-Aided Prediction of Biological Activity Spectra of Chemical Substances on the Basis of Their Structural Formulae: Computerized System PASS. Experimental and Clinical Pharmacology, 58(2), 56-62 (Rus).

30. Filimonov, D.A. & Poroikov, V.V. (1996) PASS: computerized prediction of biological activity spectra for chemical substances. In Bioactive Compound Design: Possibilities for Industrial Use (pp. 47-56). Oxford, UK: BIOS Scientific Publishers.

31. Poroikov, V.V. (1999) Computer-aided prediction of biological activity for chemical substances: possibilities and limitations. Chemistry in Russia, 2, 8-12 (Rus).

32. Lagunin, A., Stepanchikova, A., Filimonov, D. & Poroikov, V. (2000) PASS: prediction of activity spectra for biologically active substances. Bioinformatics, 16(8), 747-748. DOI: 10.1093/bioinformatics/16.8.747

33. Poroikov, V.V., Filimonov, D.A., Ihlenfeldt, W.-D., Gloriozova, T.A., Lagunin, A.A., Borodina, Yu.V., Stepanchikova, A.V. & Nicklaus, M.C. (2003). PASS Biological Activity Spectrum Predictions in the Enhanced Open NCI Database Browser. Journal of Chemical Information and Computer Sciences, 43(1) 228-236. DOI: 10.1021/ci020048r

34. Filimonov, D.A. & Poroikov, V.V. (2006) Prediction of biological activity spectra for organic compounds. Russian Chemical Journal, 50(2), 66-75 (Rus).

35. Filimonov, D.A. & Poroikov V.V. (2008). Probabilistic approach in activity prediction. In A. Varnek & A. Tropsha (Eds.) Chemoinformatics Approaches to Virtual Screening (pp. 182-216). Cambridge, UK: RSC Publishing.

36. Filimonov, D.A., Lagunin, A.A., Gloriozova, T.A., Rudik, A.V., Druzhilovskiy, D.S., Pogodin, P.V. & Poroikov V.V. (2014) Prediction of the biological activity spectra of organic compounds using the PASS online web resource. Chemistry of Heterocyclic Compounds, 50(3), 444-457. DOI: 10.1007/s10593-014-1496-1

37. Druzhilovskiy, D.S., Rudik, A.V., Filimonov, D.A., Gloriozova, T.A., Lagunin, A.A., Dmitriev, A.V., Pogodin, P.V., Dubovskaja, V.I., Ivanov, S.M., Tarasova, O.A., Bezhentsev, V.M., Murtazalieva, K.A., Semin, M.I., Maiorov, I.S., Gaur, A.S., Sastry, G.N. & Poroikov, V.V. (2017). Computational platform Way2Drug: from the prediction of biological activity to drug repurposing. Russian Chemical Bulletin, International Edition, 66(10), 1832-1841. DOI: 1066-5285/17/6610-1832

38. Poroikov, V.V., Filimonov, D.A., Borodina, Yu.V., Lagunin, A.A. & Kos, A. (2000) Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. Journal of Chemical Information and Computer Sciences, 40(6), 1349-1355. DOI: 10.1021/ci000383k

39. Filimonov, D., Poroikov, V., Borodina, Yu. & Gloriozova T. (1999) Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. Journal of Chemical Information and Computer Sciences, 39(4), 666-670. DOI: 10.1021/ci980335o

40. Höltje, H.-D., Folkers, G., Mannhold, R., Kubinyi, H. & Timmerman, H. (2008) Molecular modeling: Basic principles and applications. Weinheim: Wiley.

41. Dalby, A., Nourse, G.J., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A. & Laufer, J. (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. Journal of Chemical Information and Computer Science, 32(3), 244-255. DOI: 10.1021/ci00007a012

42. Fourches, D., Muratov, E. & Tropsha A. (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. Journal of Chemical Information and Modeling, 50(7), 1189-1204. DOI: 10.1021/ci100176x

43. Fourches, D., Muratov, E. & Tropsha A. (2015) Curation of chemogenomics data. Nature Chemical Biology, 11(8), 535. DOI: 10.1038/nchembio.1881

44. Fourches, D., Muratov, E. & Tropsha A. (2016) Trust, But Verify II: A Practical Guide to Chemogenomics Data Curation. Journal of Chemical Information and Modeling, 56(7), 1243-1252. DOI: 10.1021/acs.jcim.6b00129

45. Townsend, J.A., Glen, R.C. & Mussa, H.Y. (2012) Note on naive Bayes based on binary descriptors in cheminformatics. Journal of Chemical Information and Modeling, 52(10), 2494-2500. DOI: 10.1021/ci200303m

46. Mussa, H.Y., Mitchell, J.B.O. & Glen, R.C. (2013) Full “Laplacianised” posterior Naive Bayesian algorithm. Journal of Cheminformatics, 5, 37. DOI: 10.1186/1758-2946-5-37

47. Mussa, H.Y., Marcus, D., Mitchell, J.B. & Glen, R.C. (2015) Verifying the fully "Laplacianised" posterior Naive Bayesian approach and more. Journal of Cheminformatics, 7, 27. DOI: 10.1186/s13321-015-0075-5

48. Geronikaki, A., Druzhilovsky, D., Zakharov, A. & Poroikov, V. (2008) Computer-aided predictions for medicinal chemistry via Internet. SAR and QSAR in Environmental Research, 19(1-2), 27-38. DOI: 10.1080/10629360701843649

49. Druzhilovskiy, D.S., Rudik, A.V., Filimonov, D.A., Lagunin, A.A., Gloriozova, T.A. & Poroikov V.V. (2016) Online resources for the prediction of biological activity of organic compounds. Russian Chemical Bulletin, International Edition, 65(2), 384-393.

50. Murtazalieva, K.A., Druzhilovskiy, D.S., Goel, R.K., Sastry, G.N. & Poroikov V.V. (2017). How good are publicly available web services that predict bioactivity profiles for drug repurposing? SAR and QSAR in Environmental Research, 28(10), 843-862. DOI: 10.1080/1062936X.2017.1399448

51. Integrity. Retrieved March 24, 2018, from https://integrity.thomson-pharma.com/

52. SleepDisorders. Retrieved March 24, 2018, from http://sleepdisorders.about.com/od/sleepdisorderstreatment/a/What-Is-Topamax.htm

53. Bandini, F., Arena, E. & Mauro, G. (2012) Pre-orgasmic sexual headache responsive to topiramate: a case report. Cephalalgia, 32(10), 797-798. DOI: 10.1177/0333102412452046

54. Geronikaki, A., Babaev, E., Dearden, J., Dehaen, W., Filimonov, D., Galaeva, I., Krajneva, V., Lagunin, A., Macaev, F., Molodavkin, G., Poroikov, V., Saloutin, V., Stepanchikova, A. & Voronina, T. (2004) Design of new anxiolytics: from computer prediction to synthesis and biological evaluation. Bioorganic & Medicinal Chemistry, 12(24), 6559-6568. DOI: 10.1016/j.bmc.2004.09.016

55. B-Rao, C., Kulkarni-Almeida, A., Katkar, K.V., Khanna, S., Ghosh, U., Keche, A., Shah, P., Srivastava, A., Korde, V., Nemmani, K.V., Deshmukh, N.J., Dixit, A., Brahma, M.K., Bahirat, U., Doshi, L., Sharma, R. & Sivaramakrishnan H. (2012) Identification of novel isocytosine derivatives as xanthine oxidase inhibitors from a set of virtual screening hits. Bioorganic & Medicinal Chemistry, 20(9), 2930-2839. DOI: 10.1016/j.bmc.2012.03.019

56. Folmer, R.H.A. (2016) Integrating biophysics with HTS-driven drug discovery projects. Drug Discovery Today, 21(3), 491-498. DOI: 10.1016/j.drudis.2016.01.011

57. Babaev, E.V. (2009) Combinatorial chemistry in high school: ten-year experience of scientific, educational and organizational projects. Russian Chemical Journal, 53(5), 140-152 (Rus).

58. Lagunin, A.A., Gomazkov, O.A., Filimonov, D.A., Gureeva, T.A., Dilakyan, E.A., Kugaevskaya, E.V., Elisseeva, Yu.E., Solovyeva, N.I. & Poroikov, V.V. Computer-aided selection of potential antihypertensive compounds with dual mechanisms of action. (2003) Journal of Medicinal Chemistry, 46(15), 3326-3332. DOI: 10.1021/jm021089h

59. Geronikaki, A.A., Lagunin, A.A., Hadjipavlou-Litina, D I., Elefteriou, P.T., Filimonov, D.A., Poroikov, V.V., Alam, I. &

Saxena A.K. (2008) Computer-aided discovery of anti-inflammatory thiazolidinones with dual cyclooxygenase/lipoxygenase inhibition. Journal of Medicinal Chemistry, 51(6), 1601-1609. DOI: 10.1021/jm701496h

60. Gillbro, J.M., Lundahl, M., Westman, M., Baral, R., Al-Bader, T. & Mavon A. (2015) Structural activity relationship analysis (SAR) and *in vitro* testing reveal the anti-ageing potential activity of acetyl aspartic acid. International Journal of Cosmetic Science, 37(S1), 15-20. DOI: 10.1111/ics.12253

61. Kryzhanovskii, S.A., Salimov, R.M., Lagunin, A.A., Filimonov, D.A., Gloriozova, T.A. & Poroikov V.V. (2012) Nootropic action of some antihypertensive drugs: computer predicting and experimental testing. Pharmaceutical Chemistry Journal, 45(10), 605-611..

62. Gao, Y., O'Caoimh, R., Healy, L., Kerins, D.M., Eustace, J., Guyatt, G., Sammon, D. & Molloy, D.W. (2013) Effects of centrally acting ACE inhibitors on the rate of cognitive decline in dementia. BMJ Open, 3, 1-8. from http://doi.org/10.1136/bmjopen-2013-002881

63. Cruz-Monteagudo, M., Medina-Franco, J.L., Perez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. & Borges, F. (2014) Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discovery Today, 19(8), 1069-1080. DOI: 10.1016/j.drudis.2014.02.003

64. Bajorath, J. (2017) Representation and identification of activity cliffs. Expert Opinion on Drug Discovery, 12(9), 879-883. DOI: 10.1080/17460441.2017.1353494

65. Bates, D.O., Morris, J.C., Oltean, S. & Donaldson, L.F. (2017) Pharmacology of modulators of alternative splicing. Pharmacological Reviews, 69(1), 63-79. DOI: 10.1124/pr.115.011239

66. Hopkins, A.L. (2007) Network pharmacology. Nature Biotechnology, 25(10), 1110-1111. DOI: 10.1038/nbt1007-1110

67. Poroikov, V. (2015) 20th EuroQSAR: Understanding Chemical-Biological Interactions. Molecular Informatics, 34(6-7), 340. DOI: 10.1002/minf.201580631

68. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. (2012) The experimental uncertainty of heterogeneous public K(i) data. Journal of Medicinal Chemistry, 55(11), 51655173. DOI: 10.1021/jm300131x

69. Williams, A.J., Ekins, S. & Tkachenko, V. (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discovery Today, 17(13-14), 685701. DOI: 10.1016/j.drudis.2012.02.013

70. Tarasova, O.A., Urusova, A.F., Filimonov, D.A., Nicklaus, M.C., Zakharov, A.V. & Poroikov, V.V. (2015) QSAR modeling using large-scale databases: case study for HIV-1 reverse transcriptase inhibitors. Journal of Chemical Information and Modeling, 55(7), 1388-1399. DOI: 10.1021/acs.jcim.5b00019

71. Ivanov, S.M., Lagunin, A.A. & Poroikov, V.V. (2016) *In silico* assessment of adverse drug reactions and associated mechanisms. Drug Discovery Today, 21(1), 58-71. DOI: 10.1016/j.drudis.2015.07.018

72. Standardizer. Retrieved March 24, 2018, from https://chemaxon.com/products/chemical-structure-representation-toolkit

73. Murthy, P.K., Suneetha, V., Armakovic, S., Armakovic, S.J., Suchetan, P.A., Giri, L. & Rao, R.S. (2018) Synthesis, characterization and computational study of the newly synthetized sulfonamide molecule. Journal of Molecular Structure, 1153, 212-229. DOI: 10.1016/j.molstruc.2017.10.028

74. Costa, R.A., Oliveira, K.M.T., Costam E.V. & Pinheiro, M.L.B. (2017) Vibrational, structural and electronic properties investigation by DFT calculations and molecular docking studies with DNA topoisomerase II of strychnobrasiline type alkaloids: A theoretical approach for potentially bioactive molecules. Journal of Molecular Structure, 1145, 254-267. DOI: 10.1016/j.molstruc.2017.05.087

75. Sultan, M.A., Almansour, A.I., Pillai, R.R., Kumar, R.S., Arumugam, N., Armakovic, S., Armakovic, S.J. & Soliman, S.M. (2017) Synthesis, theoretical studies and molecular docking of a novel chlorinated tetracyclic: (Z/E)-3-(1,8-dichloro-9,10-dihydro-9,10-ethanoanthracen-11-yl) acrylaldehyde. Journal of Molecular Structure, 1150, 358-365. DOI: 10.1016/j.molstruc.2017.08.101

76. Riju, Aikkal (2016) Phytochemical analysis, carminative, enzyme inhibitor, and anticancer activities of beta-elemene. Retrieved March 24, 2018, from DOI: 10.13140/rg.2.1.2004.9048/

77. PASS Online. Retrieved March 24, 2018, from http://www.way2drug.com/passonline/

78. Pogodin, P.V., Lagunin, A.A., Filimonov, D.A. & Poroikov V.V. (2015) PASS' targets: ligand-based multi-target computational system based on public data and naive Bayes approach. SAR and QSAR in Environmental Research, 26(10), 783-793. DOI: 10.1080/1062936X.2015.1078407

79. PASS targets. Retrieved March 24, 2018, from http://www.way2drug.com/passtargets/

80. Lagunin, A., Ivanov, S., Rudik, A., Filimonov, D. & Poroikov, V. (2013) DIGEP-Pred: web-service for *in-silico* prediction of drug-induced expression profiles based on structural formula. Bioinformatics, 29(16), 2062-2063. DOI: 10.1093/bioinformatics/btt322

81. DIGEP Pred. Retrieved March 24, 2018, from http://www.way2drug.com/ge/

82. Konova, V., Lagunin, A., Pogodin, P., Kolotova, E., Shtil, A. & Poroikov V. (2015) Virtual screening of chemical compounds active against breast cancer cell lines based on cell cycle modeling, prediction of cytotoxicity and interaction with' targets. SAR and QSAR in Environmental Research, 26(7-9), 595-604. DOI: 10.1080/1062936X.2015.1076516

83. Lagunin, A.A., Dubovskaja, V.I., Rudik, A.V., Pogodin, P.V., Druzhilovskiy, D.S., Gloriozova, T.A., Filimonov, D.A., Sastry, G.N. & Poroikov, V.V. (2018) CLC-Pred: a freely available web-service for in silico prediction of human cell line cytotoxicity for drug-like compounds. PLOS One, 13(1), e0191838. DOI: 10.1371/journal.pone.0191838

84. PASS CLC Pred, Retrieved March 24, 2018, from http://www.way2drug.com/cell-line/

85. Rudik, A.V., Dmitriev, A.V., Lagunin, A.A., Filimonov, D.A. & Poroikov V.V. (2014) Metabolism sites prediction based on xenobiotics structural formulae and PASS prediction algorithm. Journal of Chemical Information and Modeling, 54(2), 498-507. DOI: 10.1021/ci400472j

86. SMP. Retrieved March 24, 2018, from http://www.way2drug.com/SMP/

87. Rudik, A., Dmitriev, A., Lagunin, A., Filimonov, D. & Poroikov, V. (2015) SOMP: web-service for *in silico* prediction of sites of metabolism for drug-like compounds. Bioinformatics, 31(12), 2046-2048. DOI: 10.1093/bioinformatics/btv087

88. SOMP. Retrieved March 24, 2018, from http://www.way2drug.com/SOMP/

89. Rudik, A.V., Dmitriev, A.V., Lagunin, A.A., Filimonov, D.A. & Poroikov, V.V. (2016) Prediction of reacting atoms for the major biotransformation reactions of organic xenobiotics. Journal

of Cheminformatics, 8, 68. DOI: 10.1186/s13321-016-0183-x

90. RA, Reacting Atoms. Retrieved March 24, 2018, from http://www.way2drug.com/RA/

91. Dmitriev, A., Rudik, A., Filimonov, D., Lagunin, A., Pogodin, P., Dubovskaja, V., Bezhentsev, V., Ivanov, S., Druzhilovsky, D., Tarasova, O. & Poroikov, V. (2017) Integral estimation of xenobiotics' toxicity with regard to their metabolism in human organism. Pure and Applied Chemistry, 89(10), 1449-1458. DOI: 10.1515/pac-2016-1205

92. Rudik, A.V., Bezhentsev, V.M., Dmitriev, A.V., Druzhilovskiy, D.S., Lagunin, A.A., Filimonov, D.A. & Poroikov, V.V. (2017) MetaTox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites. Journal of Chemical Information and Modeling, 57(4), 638642. DOI: 10.1021/acs.jcim.6b00662

93. MetaTox. Retrieved March 24, 2018, from http://way2drug.com/mg/

94. Ivanov, S.M., Lagunin, A.A., Rudik, A.V., Filimonov, D.A. & Poroikov, V.V. (2018) ADVER-Pred - web service for prediction of adverse effects of drugs. Journal of Chemical Information and Modeling, 58(1), 8-11. DOI: 10.1021/acs.jcim.7b00568

95. ADVER-Pred. Retrieved March 24, 2018, from www.way2drug.com/adverpred.

96. Lagunin, A., Rudik, A., Filimonov, D., Druzhilovsky, D. & Poroikov, V. (2018) ROSC-Pred: web-service for rodent organ-specific carcinogenicity prediction. Bioinformatics, 34(4), 710-712. DOI: 10.1093/bioinformatics/btx678

97. ROSC-Pred. Retrieved March 24, 2018, from http://www.way2drug.com/ROSC/

98. KinScreen, Retrieved March 24, 2018, from http://www.way2drug.com/KinScreen/

99. SAR Creator. Retrieved March 24, 2018, from http://www.way2drug.com/dr/substance.php/

100. Filimonov, D.A., Zakharov, A.V., Lagunin, A.A. & Poroikov V.V. (2009) QNA based "Star Track" QSAR approach. SAR and QSAR in Environmental Research, 20(7-8), 679-709. DOI: 10.1080/10629360903438370

101. Lagunin, A., Zakharov, A., Filimonov, D. & Poroikov, V. (2011) QSAR modelling of rat acute toxicity on the basis of PASS Prediction. Molecular Informatics, 30(2-3), 241-250. DOI: 10.1002/minf.201000151

102. Kokurkina, G.V., Dutov, M.D., Shevelev, S.A., Popkov, S.V., Zakharov, A.V. & Poroikov V.V. (2011) Synthesis, antifungal activity and QSAR study of 2-arylhydroxynitroindoles. European Journal of Medicinal Chemistry, 46(9), 4374-4382. DOI: 10.1016/j.ejmech.2011.07.008

103. Zakharov, A.V., Lagunin, A.A., Filimonov, D.A. & Poroikov, V.V. (2012) Quantitative prediction of antitarget interaction profiles for chemical compounds. Chemical Research in Toxicology, 25(11) 2378-2385. DOI: 10.1021/tx300247r

104. Zakharov, A.V., Peach, M.L., Sitzmann, M. & Nicklaus, M.C. (2014) QSAR Modeling of imbalanced high-throughput screening data in PubChem. Journal of Chemical Information and Modeling, 54(3), 705-712. DOI: 10.1021/ci400737s

105. Fedorova, E.V., Buryakina, A.V., Zakharov, A.V., Filimonov, D.A., Lagunin, A.A. & Poroikov V.V. (2014). Design, synthesis and pharmacological evaluation of novel vanadium-containing complexes as antidiabetic agents. PLOS One, 9(7), e100386. DOI: 10.1371/journal.pone.0100386

106. Hadjikakou, S.K., Ozturka, I.I., Banti, C.N., Kourkoumelis, N. & Hadjiliadis, N. (2015). Recent advances on antimony (III/V) compounds with potential activity against tumor cells. Journal of Inorganic Biochemistry, 153, 293-305. DOI: 10.1016/j.jinorgbio.2015.06.006

107. Ajeet, Verma, M., Rani, S. & Kumar, A. (2016) Antitarget interaction, acute toxicity and protein binding studies of quinazolinedione sulphonamides as GABA1 antagonists. Indian Journal of Pharmaceutical Sciences, 78(1), 4853.

108. Unnissa, S.H. & Rajan, D. (2016) Drug design, development and biological screening of pyridazine derivatives. Journal of Chemical and Pharmaceutical Research, 8(8), 999-1004. Retrieved March 24, 2018, from http://www.jocpr.com/articles/drug-design-development-and-biological-screening-of-pyridazine-derivatives.pdf

109. Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., Zakharov, A., Worth, A., Richard, A.M., Grulke, C.M., Trisciuzzi, D., Fourches, D., Horvath, D., Benfenati, E., Muratov, E., Wedebye, E.B., Grisoni, F., Mangiatordi, G.F., Incisivom G.M., Hong, H., Ng, H.M., Tetko, I.V., Balabin, I., Kancherla, J., Shen, J., Burton, J., Nicklaus, M., Cassotti, M., Nikolov, N.G., Nicolotti, O., Andersson, P.L., Zang, Q., Politi, R., Beger, R.D., Todeschini, R., Huang, R., Farag, S., Rosenberg, S. A., Slavov, S., Hu, X. & Judson R S. (2016) CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environmental Health Perspectives, 124(7), 1023-1033. DOI: 10.1289/ehp.1510267

110. Ozturk, I.I., Yarar, S., Banti, C.N., Kourkoumelis, N., Chrysouli, M.P., Manoli, M., Tasiopoulos, A.J. & Hadjikakou, S.K. (2017). QSAR studies on antimony (III) halide complexes with N-substituted thiourea derivatives. Polyhedron, 123, 152-161. DOI: 10.1016/j.poly.2016.11.008

Поступила: 19. 03. 2018.
Принята к публикации: 24. 03. 2018.

COMPUTER-AIDED PREDICTION OF BIOLOGICAL ACTIVITY SPECTRA FOR CHEMICAL COMPOUNDS: OPPORTUNITIES AND LIMITATIONS

D.A. Filimonov[1], D.S. Druzhilovskiy[1], A.A. Lagunin[1,2], T.A. Gloriozova[1], A.V. Rudik[1], A.V. Dmitriev[1], P.V. Pogodin[1], V.V. Poroikov[1]*

[1]Institute of Biomedical Chemistry,
10 Pogodinskaya str., 10 bldg. 8, Moscow, 119121 Russia; *e-mail: vladimir.poroikov@ibmc.msk.ru
[2]Pirogov Russian National Research Medical University, 1 Ostrovityanova str., 1, Moscow, 117997 Russia

An essential characteristic of chemical compounds is their biological activity since its presence can become the basis for the use of the substance for therapeutic purposes, or, on the contrary, limit the possibilities of its practical application due to the manifestation of side action and toxic effects. Computer assessment of the biological activity spectra makes it possible to determine the most promising directions for the study of the pharmacological action of particular substances, and to filter out potentially dangerous molecules at the early stages of research. For more than 25 years, we have been developing and improving the computer program PASS (Prediction of Activity Spectra for Substances), designed to predict the biological activity spectrum of substance based on the structural formula of its molecules. The prediction is carried out by the analysis of structure-activity relationships for the training set, which currently contains information on structures and known biological activities for more than one million molecules. The structure of the organic compound is represented in PASS using Multilevel Neighborhoods of Atoms descriptors; the activity prediction for new compounds is performed by the naive Bayes classifier and the structure-activity relationships determined by the analysis of the training set. We have created and improved both local versions of the PASS program and freely available web resources based on PASS (http://www.way2drug.com). They predict several thousand biological activities (pharmacological effects, molecular mechanisms of action, specific toxicity and adverse effects, interaction with the unwanted targets, metabolism and action on molecular transport), cytotoxicity for tumor and non-tumor cell lines, carcinogenicity, induced changes of gene expression profiles, metabolic sites of the major enzymes of the first and second phases of xenobiotics biotransformation, and belonging to substrates and/or metabolites of metabolic enzymes. The web resource Way2Drug is used by over 19000 researchers from more than 100 countries around the world, which allowed them to obtain over 600000 predictions and publish about 500 papers describing the obtained results. The analysis of the published works shows that in some cases the interpretation of the prediction results presented by the authors of these publications requires an adjustment. In this work, we provide the theoretical basis and consider, on particular examples, the opportunities and limitations of computer-aided prediction of biological activity spectra.

Key words: analysis of structure-activity relationships; biological activity spectra; computer-aided prediction; PASS; accuracy; predictivity; web-resource Way2Drug