КРУПНОМАСШТАБНОЕ ПРЕДСКАЗАНИЕ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ С ИСПОЛЬЗОВАНИЕМ СИСТЕМЫ ACTIVE-IT

В.Л. Алмейда[1,2], О.Д.Х. дос Сантос[3], Х.С.Д. Лопес[1]*

[1]Chemoinformatics Group — NEQUIM, Departamento de Quimica, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais (UFMG), av. Pres. Antônio Carlos, 6627, Pampulha, 31.270-901, Belo Horizonte, MG, Brazil; *e-mail: jlopes.ufmg@gmail.com
[2]Servico de Fitoquimica e Prospeccao Farmaceutica, Fundacao Ezequiel Dias (FUNED), Belo Horizonte, MG, Brazil
[3]Departamento de Farmacia, Escola de Farmacia, Universidade Federal de Ouro Preto (UFOP), Brazil

Традиционные методы тестирования при разработке новых фармацевтических препаратов являются трудоёмкими и дорогими, однако инструменты *in silico* оценки могут помочь в решении этой проблемы. Система Active-IT — "инструмент" для проведения виртуального скрининга на основе структуры лигандов, который был разработан нами для предсказания биологической активности малых органических молекул. Включает в себя четыре независимых модуля: модуль генерации молекулярных дескрипторов (3D-Pharma); модуль машинного обучения (ExCVBA); базу данных о биологических активностях; модуль предсказания. Данные о биологических активностях были получены из базы данных PubChem BioAssay. Для построения моделей машинного обучения использованы метод опорных векторов и наивный байесовский классификатор. Модели были сконструированы с использованием случайного рекурсивного стратифицированного разбиения, их валидацию проводили путём рандомизации данных по активности (Y-random). Были построены модели для 3500 биологических тест-систем, каждая из которых состоит из: (i) 30 моделей, построенных с использованием метода опорных векторов; (ii) 30 моделей, построенных по наивному байесовскому алгоритму; (iii) 60 рандомизированных моделей для валидации. Биологические тест-системы, обладающие низкой производительностью или невысокой дискриминационной способностью, были исключены. С использованием системы Active-IT в данной работе была проведена оценка трёх биоактивных компонентов чая Аяуаска. Прогнозы были проверены с использованием известных мишеней, описанных в нескольких общедоступных базах данных. Результаты внешней валидации показали, что 16 из 33 (48,5%, p<0,0001) известных мишеней были предсказаны верно. Такой уровень точности при крупномасштабном виртуальном скрининге является удовлетворительным, и демонстрирует эффективность методологии Active-IT в прогнозировании биологической активности для малых органических молекул.

Ключевые слова: виртуальный скрининг на основе структуры лигандов; предсказание биологической активности; машинное обучение; случайное рекурсивное стратифицированное разбиение; фармакофорные фингерпринты; 3D молекулярные структуры

DOI: 10.18097/PBMC20247006435

ВВЕДЕНИЕ

Традиционные методы тестирования при разработке фармацевтических препаратов являются времяёмкими и дорогостоящими. В связи с этим оценка биологической активности *in silico* представляет многообещающую альтернативу. Система Active-IT — это современная платформа, которая включает в себя улучшенную генерацию молекулярных дескрипторов, машинное обучение, предсказательные модели и специальный модуль прогнозирования [1]. Она может быть эффективно использована в процессе поиска и разработки лекарств [1–5].

В Active-IT для описания структуры используются мультиконформационные бинарные фармакофорные дескрипторы. Предложенный инновационный подход учитывает конформационную динамику трёхмерных структур взамен традиционных 2D-вычислений, обычно используемых при виртуальном скрининге [6]. Отличительной чертой Active-IT является отказ от единой "лучшей модели" в пользу рекурсивного построения и оценки нескольких моделей. Эти модели используются на этапе прогноза, что позволяет получать устойчивые результаты.

Основная цель нашей работы — представление методологии Active-IT и её валидации при оценке биологической активности алкалоидов из стебля *Banisteriopsis caapi* и листьев *Psychotria viridis*. Эти два растения входят в состав чая Аяуаска (используемого в местных религиозных ритуалах), известного своими психоактивными свойствами [7]. Наша система Active-IT была использована для прогноза потенциальных видов биологической и фармакологической активности алкалоидов **1**–**3** (рис. 1), входящих в состав этих растений [8, 9].

МЕТОДИКА

Система NEQUIM Active-IT — это инновационная платформа для прогноза биологической активности на основе структуры лигандов, содержащая четыре основных компонента (рис. 2): (1) генерацию дескрипторов веществ (3D-Pharma), (2) современные

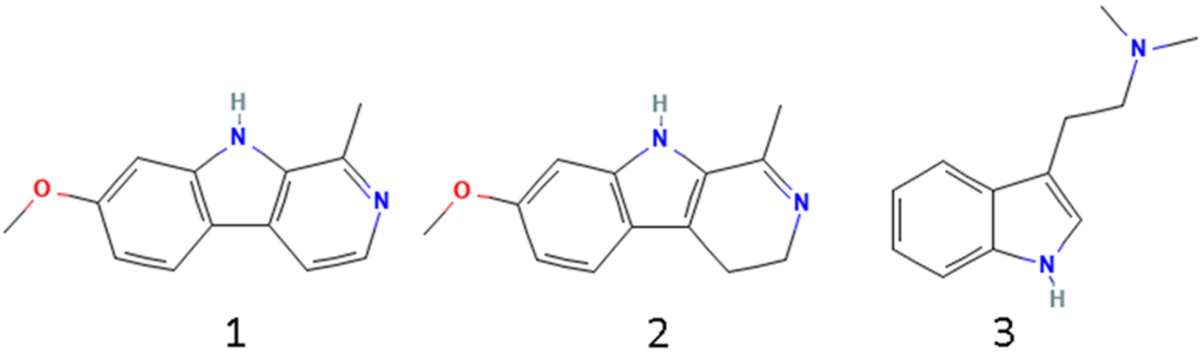

Рисунок 1. Химическая структура 7-метокси-1-метил-9H-пиридо[3,4-b]индола (гармин, **1**), 7-метокси-1-метил-4,9-дигидро-3H-пиридо[3,4-b]индола (гармалин, **2**) из стеблей *B. caapi* и 2-(1H-индол-3-ил)-N,N-диметилэтанамин (N,N-диметилтриптамин, **3**) (ДМТ) из листьев *P. viridis*.

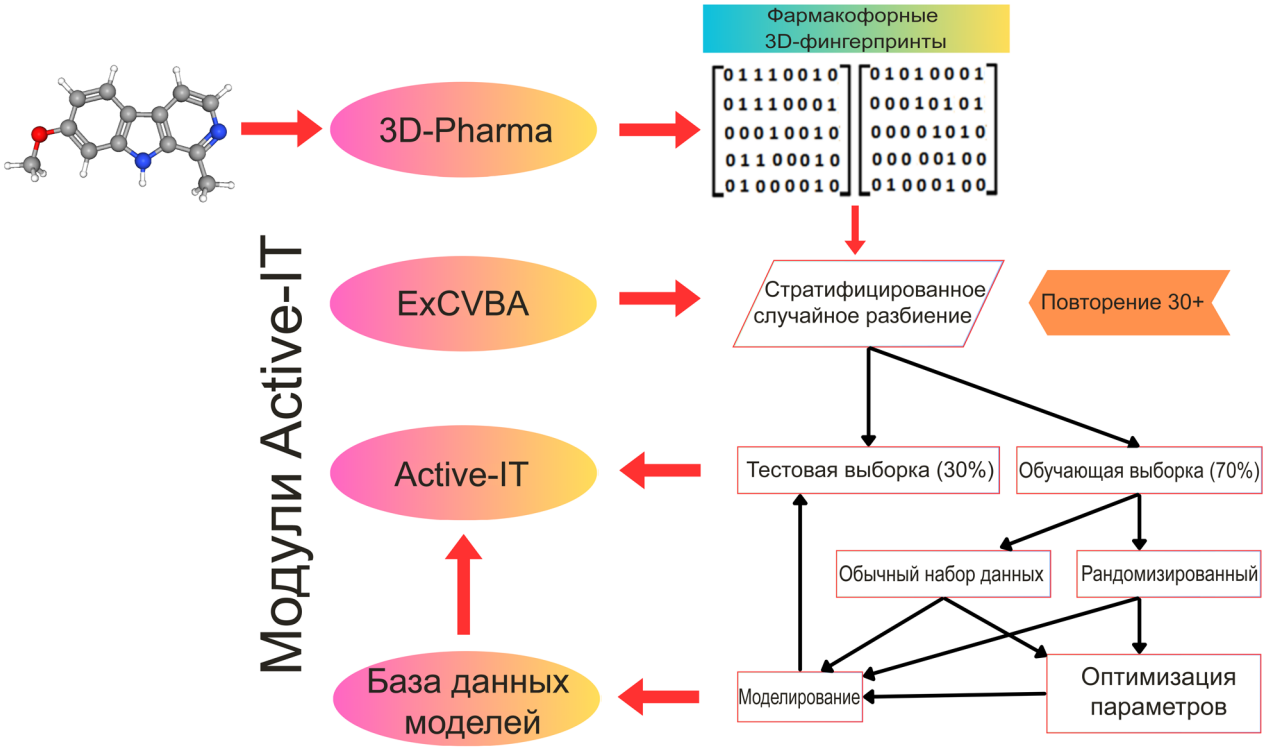

Рисунок 2. Блок-схема системы Active-IT с её четырьмя модулями и показан протокол моделирования, использующий рекурсивный метод стратифицированного случайного разбиения, управляемый модулем ExCVBA. Цветной вариант рисунка доступен в электронной версии статьи на сайте журнала.

методы машинного обучения (ExCVBA), (3) обширную базу данных предсказательных моделей, (4) робастный модуль прогнозирования (Active-IT). На рисунке 2 показана упрощённая схема системы Active-IT и выполняемых с её помощью процессов моделирования.

Модуль 3D-Pharma использует фармакофорные 3D-фингерпринты для кодирования молекулярных структур с учётом множества конформаций [10]. Эти фингерпринты позволяют идентифицировать ключевые особенности, ответственные за биологическую активность соединений. Каждая трёхмерная структура тщательно анализируется; при этом все атомы, кроме водорода, подразделяются на шесть групп: положительные, отрицательные, доноры водорода, акцепторы водорода, гидрофобные и ароматические [11]. Кроме того, модуль 3D-Pharma рассчитывает все возможные комбинации трёх фармакофоров с учётом расстояний (дискретизированных по десяти ячейкам). Каждый триплет фармакофоров представлен строкой из 6 символов (3 символа для объекта в вершинах и 3 — для ячеек расстояний по ребрам), которая однозначно идентифицирует этот триплет [12]. Каждая конформация соединения связана с вектором трёхточечных потенциальных фармакофоров (3PPP). Для получения одного вектора, описывающего соединение, векторы, полученные для каждой конформации, объединяют в уникальный мультиконформационный бинарный модальный вектор [13]. Этот вектор содержит информацию о структурных деталях и динамическом поведении соединения в четырёх измерениях [14].

Система Active-IT использует обширные данные, извлечённые из базы данных PubChem BioAssay [15, 16]. Системный подход к отбору данных обеспечивает их достоверность. Структуры (до 10 конформаций) каждого классифицированного соединения из каждой тест-системы были загружены из базы данных соединений PubChem [17] и обработаны, как описано выше. Мы объединили в нашей программе, написанной на языке Perl, информацию о более чем 3500 тест-системах, объединив их с модулем ExCVBA [18], в котором реализованы методы машинного обучения, включая метод опорных векторов (реализован через LibSVM Chang и Lin [19]) и наивный Байес (разработан Williams [20] и доступен в репозитории CPAN).

При построении моделей для прогноза биологической активности использовали строгий метод случайного рекурсивного стратифицированного разбиения. Этот метод обеспечивает согласованное распределение активных и неактивных соединений как в тестовой (или валидационной), так и в обучающей выборках. Наш подход заключался в случайном разделении каждого набора соединений на две подгруппы с разделением на 30% и 70% без замены, соответствующие тестовой и обучающей выборкам. Этот рекурсивный процесс повторяли минимум 30 раз (как показано на рисунке 2) для каждой биологической тест-системы, в результате чего было получено 30 различных моделей для каждого метода машинного обучения. Для оценки эффективности наших моделей, каждый набор данных был подвергнут процессу рандомизации данных по активности (Y-random). В ходе этого процесса были сгенерированы случайные выборки, что позволило оценить достоверность моделей с точки зрения риска переобучения [21]. В результате для каждой проанализированной биологической тест-системы было разработано 120 моделей (60 обычных и 60 рандомизированных).

На рисунке 3 представлено распределение кривых AUC-ROC для тест-систем PubChem AID 1194 (мутагенность сальмонеллы DSSTox). Показаны все четыре набора моделей (как обычный, так и рандомизированный SVM и наивный Байес) и представлена наглядная картина процесса рандомизации. Если модели, сгенерированные с помощью рандомизации, обладали такой же прогностической точностью, как и обычные модели (подтверждено наложением кривых распределения AUC), то обычные модели отвергались как непригодные для надёжного прогнозирования потенциальной биологической активности. Кроме того, модели, созданные для каждой тест-системы, были оценены с использованием тестовых выборок, что позволило определить общее значение AUC для каждого метода. Впоследствии модели для некоторых тест-систем были исключены, поскольку их точность была ниже порогового значения AUC > 0,6. Около 22% наборов данных, с использованием которых были построены модели при помощи SVM и 39% наборов данных, использованных для построения моделей при помощи наивного байесовского алгоритма, были исключены на основе этих двух фильтров.

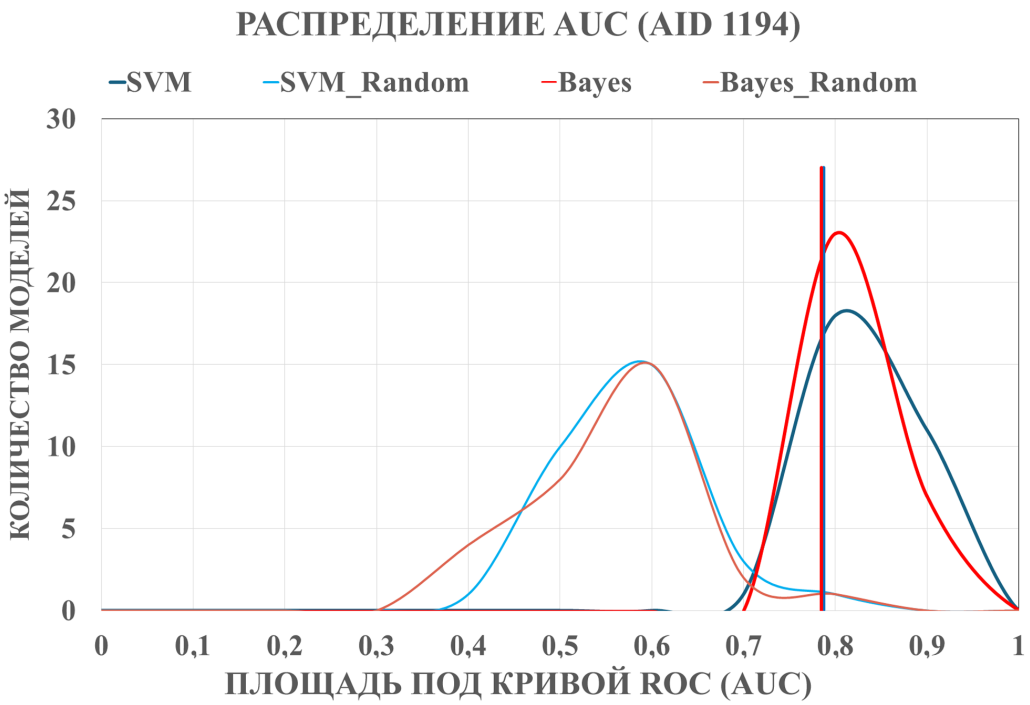

Рисунок 3. Распределение значений площади под кривой ROC (AUC) для моделей, сгенерированных для каждого метода (SVM и наивный Байес), и процесса рандомизации активности (Y-random) для тест-системы PubChem 1194 (мутагенность сальмонеллы DSSTox). Вертикальные линии представляют собой значение общей AUC, рассчитанное по полному набору моделей с использованием тестовой выборки. В случае данной тест-системы общая AUC превышает пороговое значение 0,6 и считается удовлетворительной. Цветной вариант рисунка доступен в электронной версии статьи на сайте журнала.

Метод опорных векторов (SVM) с линейным ядром требует точной настройки параметра издержек (cost parameter), которую мы выполнили путём 5-кратной кросс-валидации в обучающей выборке, используя разработанную нами ранее метрику мощности (PM) в качестве целевой функции [22, 23]. Для прогнозирования потенциальной биологической активности новых соединений их структуры извлекали из базы данных PubChem и подвергали указанным выше этапам обработки. В тех случаях, когда структуры были недоступны, их строили с помощью программного обеспечения ChemAxon (https://www.chemaxon.com), а конформации получали с использованием программы OpenEye OMEGA [24].

В ходе прогнозирования мультиконформационный модальный фингерпринт нового соединения вносили во все модели, разработанные для каждой биологической тест-системы. Каждая модель выдаёт исходную оценку, которая сравнивается с исходным распределением оценок “активный/неактивный” для соединений, тестовой выборки. В результате получают две новые оценки, соответствующие вероятностям того, что соединение либо проявляет активность (Pa), либо не проявляет активность (Pi) (как показано на рисунке 4) [25]. В нашем предыдущем исследовании [5] мы установили пороговые величины 0,5 и 0,8 для значений разности Pa и Pi (Pa – Pi) в методах SVM и наивного Байеса соответственно, которые могут считаться перспективными в контексте наличия активности у исследуемого соединения.

Ограничительные линии на рисунке 4 рассчитаны на основе дисперсии Pa – Pi (σ^2_{Pa-Pi}, оцененной аналитически, как обсуждалось ранее в работе Rocha и соавт. [2], в соответствии с уравнением (1), где Na и Ni — количество активных и неактивных соединений.

$$\sigma^2_{Pa-Pi} = \frac{Pa\times(1-Pa)}{Na} + \frac{Pi\times(1-Pi)}{Ni} \quad (1)$$

Доверительный интервал Pa – Pi был рассчитан на основе дисперсии и *t*-критерия Стьюдента для 95% уровня значимости (уравнение (2)) [26].

$$(Pa-Pi)_{estimate} = (Pa-Pi)_{mean} \pm t_{stat} \times \sqrt{\sigma^2_{Pa-Pi}} \quad (2)$$

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Потенциальные виды биологической активности для соединений **1–3** чая Аяуаски были предсказаны с помощью системы Active-IT. Их фармакофорные фингерпринты были внесены в систему Active-IT для прогнозирования их потенциальной активности с использованием SVM (2782 биологических тестов) и наивного Байеса (2176 биологических тестов). С учётом только биологических тестов, связанных с конкретными биологическими мишенями, их число составляет 1550 для SVM и 1111 для наивного Байеса.

Для внешней валидации прогнозов были использованы мишени, найденные в базе данных PubChem Compound (“Chemical-Target Interactions”) для соединений гармина (**1**) (77 мишеней), гармалина (**2**) (30 мишеней) и N,N-диметилтриптамина (**3**) (13 мишеней). Эта информация была агрегирована из нескольких баз данных: таких как DrugBank, IUPHAR/BPS, базы данных терапевтических мишеней (TTD) и базы данных сравнительной токсикогеномики (CTD).

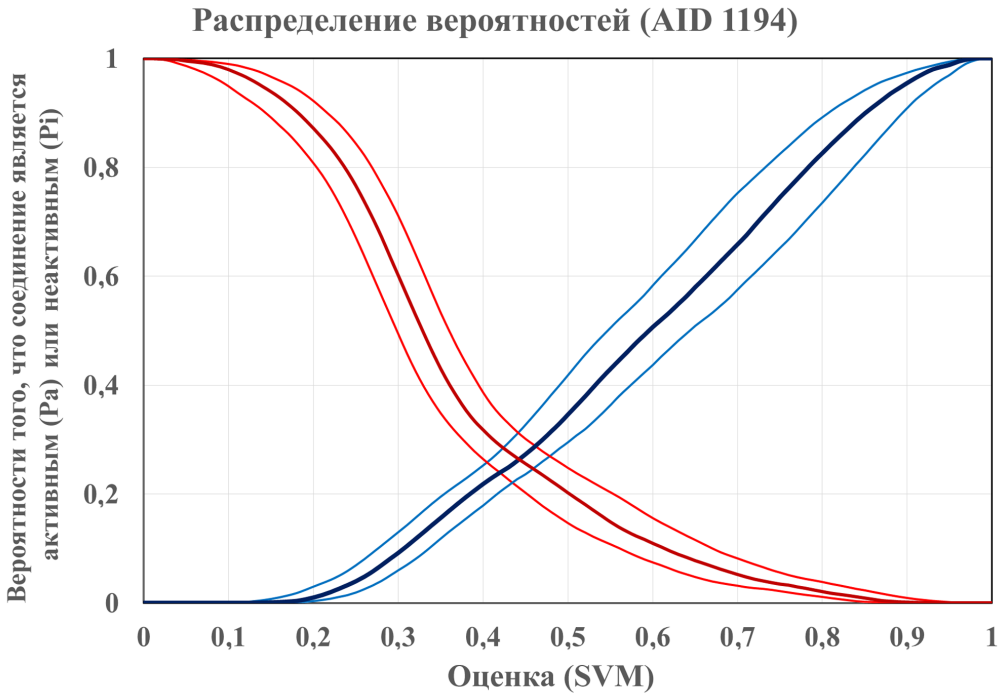

Рисунок 4. Распределение вероятностей того, что соединение является активным (Pa), синие линии по возрастанию, и неактивным (Pi), красные линии по убыванию, в зависимости от исходной оценки SVM (или наивного Байеса) для биологической тест-системы PubChem AID 1194 (мутагенность DSSTox Salmonella). Цветной вариант рисунка доступен в электронной версии статьи на сайте журнала.

Учитывая только мишени, доступные в системе Active-IT, осталось 33 прогнозируемых мишени. Анализируя только те случаи, в которых значения Pa – Pi превышают вышеупомянутые пороговые значения, методы SVM и наивного Байеса правильно предсказали 16 мишеней с высокой степенью комплементарности (табл. 1). Таким образом, около 48% мишеней были предсказаны правильно.

Для оценки статистической значимости полученных результатов, мы использовали значение *p*, которое в данном случае соответствует вероятности получения результата, равного или существенно превышающего случайный. В системе Active-IT было оценено значение *p* с использованием гипергеометрического распределения (3), где *N* — количество смоделированных биологических тест-систем, использованных в прогнозе, *k* — количество известных мишеней, *M* — количество отобранных биологических тестов и *n* — количество известных мишеней, предсказанных среди отобранных биологических тестов.

$$p\text{-value} = \left(\frac{k}{n}\right)\left(\frac{N-k}{M-n}\right) \quad (3)$$

В таблице 2 показано количество биологических тестов, выбранных для каждого соединения в обоих методах прогнозирования, которые использовались для оценки *p*-значений прогнозов. Результаты прогноза для 16 мишеней, полученные с помощью обоих методов, имеют *p*-value<0,0001, что свидетельствует о высокой значимости полученных оценок точности наших прогнозов.

ЗАКЛЮЧЕНИЕ

Наше исследование посвящено разработке и оценке системы Active-IT — современного “инструмента” для виртуального скрининга на основе структуры лигандов, предназначенного для прогнозирования биологической активности малых органических молекул. Active-IT включает мультиконформационные фармакофорные бинарные фингерпринты для генерации молекулярных дескрипторов, рекурсивное стратифицированное случайное разбиение набора данных для разработки моделей машинного обучения и робастный модуль прогнозирования биологических активностей.

Предсказательная точность Active-IT была продемонстрирована при оценке свойств трёх биологически активных соединений, содержащихся в чае Аяуаска. 48,5% (*p*-value<0,0001) известных мишеней были предсказаны точно. Этот уровень точности при крупномасштабном виртуальном скрининге является удовлетворительным. Результаты внешней валидации показывают, что система Active-IT эффективна в прогнозировании биологической активности и может внести значительный вклад в поиск и разработку лекарственных средств.

БЛАГОДАРНОСТИ

Авторы выражают благодарность Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) и Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) за финансовую поддержку.

Таблица 1. Количество известных мишеней для анализируемых соединений и количество мишеней, правильно предсказанных системой Active-IT

Соединение	Известные мишени	PubChem Bioassays		
		Мишени, отобранные с помощью SVM	Мишени, отобранные с помощью наивного Байеса	Мишени, отобранные с помощью SVM или наивного Байеса
Гармин (**1**)	19	5 (26%)	5 (26%)	8 (42%)
Гармалин (**2**)	11	4 (36%)	4 (36%)	6 (55%)
N,N-Диметилтриптамин (ДМТ) (**3**)	3	1 (33%)	1 (33%)	2 (67%)
Всего	**33**	**10 (30%)**	**10 (30%)**	**16 (48%)**

Таблица 2. Количество смоделированных тестов в PubChem Bioassays, связанных с биологическими мишенями, использованными в прогнозах, и количество тестов, выбранных каждым методом). В последнем столбце приведены значения *p*, рассчитанные для прогнозов с использованием обоих методов (SVM или наивного Байеса)

Соединение	Известные мишени	PubChem Bioassays			*p*
		Мишени, отобранные с помощью SVM	Мишени, отобранные с помощью наивного Байеса	Мишени, отобранные с помощью SVM или наивного Байеса	
Гармин (**1**)	19	137	78	193	<0,00005
Гармалин (**2**)	11	125	49	160	<0,00002
N,N-Диметилтриптамин (ДМТ) (**3**)	3	109	32	132	<0,01000
Все отобранные мишени	33	371	159	485	<0,00010
Все предсказанные мишени	—	1550	1111	2661	—

ФИНАНСИРОВАНИЕ

Исследование было поддержано бразильской программой “Наука без границ” (CNPq стипендия 202407/2014-4 to JCDL и 249299/2013-5 to VLA) и FAPEMIG стипендия BIP-00213-24 to VLA.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Эта статья не содержит исследований с участием людей или с использованием животных в качестве экспериментальных объектов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. *Rocha M.P., Campana P.R.V., Scoaris D.O., Almeida V.L., Lopes J.C.D., Shaw J.M.H., Silva C.G.* (2018) Combined *in vitro* studies and *in silico* target fishing for the evaluation of the biological activities of *Diphylleia cymosa* and *Podophyllum hexandrum*. Molecules (Basel), **23**(12), 3303. DOI: 10.3390/molecules23123303
2. *Rocha M.P., Campana P.R.V., Scoaris D.O., Almeida V.L., Lopes J.C.D., Silva F.A., Pieters L., Silva G.C.* (2018) Biological activities of extracts from *Aspidosperma subincanum* Mart. and *in silico* prediction for inhibition of acetylcholinesterase. Phytother. Res., **32**(10), 2021–2033. DOI: 10.1002/ptr.6133
3. *Briñez-Ortega E., Almeida V.L., Lopes J.C.D., Burgos A.E.* (2020) Partial inclusion of bis(1,10-phenanthroline)silver(I) salicylate in β-cyclodextrin: Spectroscopic characterization, *in vitro* and *in silico* antimicrobial evaluation. Anais da Academia Brasileira de Ciências, **92**(3), e20181323. DOI: 10.1590/0001-3765202020181323
4. *da Silva R.G., Almeida T.C., Reis A.C.C., Filho S.A.V., Brandão G.C., da Silva G.N., de Sousa H.C., de Almeida V.L., Lopes J.C.D., de Souza G.H.B.* (2021) *In silico* pharmacological prediction and cytotoxicity of flavonoids glycosides identified by UPLC-DAD-ESI-MS/MS in extracts of *Humulus lupulus* leaves cultivated in Brazil. Nat. Prod. Res., **35**(24), 5918–5923. DOI: 10.1080/14786419.2020.1803308
5. *Sudan C.R.C., Pereira L.C., Silva A.F., Moreira C.P.S., de Oliveira D.S., Faria G., dos Santos J.S.C., Leclercq S.Y., Caldas S., Silva C.G., Lopes J.C.D., de Almeida V.L.* (2021) Biological activities of extracts from *Ageratum fastigiatum*: Phytochemical study and *in silico* target fishing approach. Planta Medica, **87**(12–13), 1045–1060. DOI: 10.1055/a-1576-4080
6. *Axen S.D., Huang X.P., Cáceres E.L., Gendelev L., Roth B.L., Keiser M.J.* (2017) A simple representation of three-dimensional molecular structure. J. Med. Chem., **60**(17), 7393–7409. DOI: 10.1021/acs.jmedchem.7b00696
7. *Gonçalves J., Luís Â., Gallardo E., Duarte A.P.* (2023) A systematic review on the therapeutic effects of Ayahuasca. Plants, **12**(13), 2573. DOI: 10.3390/plants12132573
8. *Pires A.P., de Oliveira C.D., Moura S., Dörr F.A., Silva W.A., Yonamine M.* (2009) Gas chromatographic analysis of dimethyltryptamine and beta-carboline alkaloids in Ayahuasca, an Amazonian psychoactive plant beverage. Phytochem. Anal., **20**(2), 149–153. DOI: 10.1002/pca.1110
9. *Callaway J.C., McKenna D.J., Grob C.S., Brito G.S., Raymon L.P., Poland R.E., Andrade E.N., Andrade E.O., Mash D.C.* (1999) Pharmacokinetics of Hoasca alkaloids in healthy humans. J. Ethnopharmacology, **65**(3), 243–256. DOI: 10.1016/s0378-8741(98)00168-8
10. *Domingues B.F., Martins-José A., Lopes J.C.D.* (2024) 3D-Pharma, a ligand-based virtual screening tool using 3D pharmacophore fingerprints. ChemRxiv (Preprint), **2024**, DOI: 10.26434/chemrxiv-2024-dkvf8
11. *Sud M.* (2016) Mayachemtools: An open source package for computational drug discovery. J. Chem. Inf. Model., **56**(12), 2292–2297. DOI: 10.1021/acs.jcim.6b00505
12. *Abrahamian E., Fox P.C., Naerum L., Christensen I.T., Thøgersen H., Clark R.D.* (2003) Efficient generation, storage, and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. J. Chem. Inf. Comput. Sci., **43**(2), 458–468. DOI: 10.1021/ci025595r
13. *Shemetulskis N.E., Weininger D., Blankley C.J., Yang J.J., Humblet C.* (1996) Stigmata: An algorithm to determine structural commonalities in diverse datasets. J. Chem. Inf. Comput. Sci., **36**(4), 862–871. DOI: 10.1021/ci950169
14. *Domingues B.F., Lopes J.C.D.* (2012) 3D-Pharma: Uma Ferramenta para Triagem Virtual Baseada em Fingerprints de Farmacyforos. [Doctoral dissertation, Universidade Federal de Minas Gerais]. UFMG Institutional Repository. (in Portuguese) Retrieved September 29, 2024 from: http://hdl.handle.net/1843/BUBD-9DKHDA
15. *Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q.,Shoemaker B.A., Thiessen P.A., Yu B., Zaslavsky L., Zhang J., Bolton E.E.* (2023) PubChem 2023 update. Nucleic Acids Res., **51**(D1), D1373–D1380. DOI: 10.1093/nar/gkac956
16. *Kim S., Bolton E.E.* (2024) PubChem: A Large-Scale Public Chemical Database For Drug Discovery. In: Open Access Databases and Datasets for Drug Discovery (Daina A., Przewosny M., Zoete V., eds.). pp. 39–66. DOI: 10.1002/9783527830497.ch2
17. *Bolton E.E., Chen J., Kim S., Han L., He S., Shi W., Simonyan V., Sun Y., Thiessen P.A., Wang J., Yu B., Zhang J., Bryant S.H.* (2011) PubChem3D: A new resource for scientists. J. Cheminformatics, **3**(1), 32. DOI: 10.1186/1758-2946-3-32
18. *Santos F.M., de Winter H., Augustyns K., Lopes J.C.D.* (2015) Use of extensive cross-validation and bootstrap application (ExCVBA) for molecular modeling of some pharmacokinetics properties. Poster presented at OPENTOX EURO 2015 — OpenTox InterAction Meeting — Innovation in Predictive Toxicology, Dublin, Ireland. DOI: 10.13140/RG.2.1.2274.8888
19. *Chang C., Lin C.* (2011) LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., **2**(3), 27. DOI: 10.1145/1961189.196119
20. *Williams K.* (2004) Naïve Bayes algorithm at comprehensive perl archive network. Retrieved September 29, 2024 from: https://metacpan.org/pod/Algorithm::NaiveBayes
21. *Tropsha A.* (2010) Best practices for QSAR model development, validation, and exploitation. Mol. Inform., **29**(6–7), 476–488. DOI: 10.1002/minf.201000061
22. *Lopes J.C.D., dos Santos F.M., Martins-José A., Augustyns K., de Winter H.* (2017) The power metric: A new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. J. Cheminformatics, **9**, 7. DOI: 10.1186/s13321-016-0189-4

23. *de Winter H., Lopes J.C.D.* (2018) Reply to the comment made by Šicho, Vorśilák and Svozil on “The power metric: A new statistically robust enrichment-type metric for virtual screening applications with early recovery capability”. J. Cheminformatics, **10**, 14. DOI: 10.1186/s13321-018-0262-2
24. *Hawkins P.C., Nicholls A.* (2012) Conformer generation with OMEGA: Learning from the data set and the analysis of failures. J. Chem. Inf. Model., **52**(11), 2919–2936. DOI: 10.1021/ci300314k
25. *Filimonov D.A., Lagunin A.A., Gloriozova T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V.* (2014) Prediction of the biological activity spectra of organic compounds using the PASS online web resource. Chem. Heterocycl. Compd., **50**(3), 444–457. DOI: 10.1007/s10593-014-1496-1
26. *Nicholls A.* (2014) Confidence limits, error bars and method comparison in molecular modeling. Part 1: The calculation of confidence intervals. J. Comput.-Aided Mol. Des., **28**(9), 887–918. DOI: 10.1007/s10822-014-9753-z

Поступила в редакцию: 07. 10. 2024.
После доработки: 01. 11. 2024.
Принята к печати: 03. 11. 2024.

LARGE-SCALE PREDICTION OF BIOLOGICAL ACTIVITIES WITH ACTIVE-IT SYSTEM

V.L. Almeida[1,2], O.D.H. dos Santos[3], J.C.D. Lopes[1]*

[1]Chemoinformatics Group — NEQUIM, Departamento de Quimica, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais (UFMG), av. Pres. Antônio Carlos, 6627, Pampulha, 31.270-901, Belo Horizonte, MG, Brazil; *e-mail: jlopes.ufmg@gmail.com
[2]Serviço de Fitoquímica e Prospecção Farmacêutica, Fundação Ezequiel Dias (FUNED), Belo Horizonte, MG, Brazil
[3]Departamento de Farmácia, Escola de Farmácia, Universidade Federal de Ouro Preto (UFOP), Brazil

Traditional testing methods in pharmaceutical development can be time-consuming and costly, but *in silico* evaluation tools can offer a solution. Our *in-house* Active-IT system, a Ligand-Based Virtual Screening (LBVS) tool, was developed to predict the biological and pharmacological activities of small organic molecules. It includes four independent modules for generating molecular descriptors (3D-Pharma), machine learning modeling (ExCVBA), a database of bioactivity models, and a prediction module. Activity data collected from the PubChem BioAssay database was used for modelling SVM and Naïve Bayes machine learning methods. Models have been constructed using a recursive stratified partition method and validated through an activity randomization (Y-random) process. Over 3500 bioassays were modeled, each comprising 30 SVM and 30 Naïve Bayes models and 60 randomized models. Bioassays with low performance or discrimination between regular and randomized were discarded. Using the Active-IT system we have evaluated three bioactive compounds of Ayahuasca tea. The predictions were thoroughly validated using known targets described in several public databases. The external validation results are noteworthy, with 16 of 33 (48.5% with p-value<0.0001) known targets correctly predicted. This level of accuracy in large-scale virtual screening methods is very significant and demonstrates the effectiveness of the Active-IT methodology in predicting the potential biological activities of small organic molecules.

The whole English version is available at http://pbmc.ibmc.msk.ru.

Key words: ligand-based virtual screening; bioactivity prediction; machine learning modeling; recursive stratified random partition; pharmacophore fingerprint; 3D molecular structures

Funding. This study was supported by the Brazilian Science Without Border program (CNPq fellowships 202407/2014-4 to JCDL and 249299/2013-5 to VLA) and FAPEMIG fellowship BIP-00213-24 to VLA.

Received: 07.10.2024; revised: 01.11.2024; accepted: 03.11.2024.