УДК 577.29

# СРАВНЕНИЕ КОМПЬЮТЕРНЫХ ПРЕДИКТОРОВ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ (АНАЛИТИЧЕСКИЙ ОБЗОР)

**Адылова Ф.Т.**
д.т.н., зав. лабораторией Института математики при Национальном Университете Узбекистана,
тел.: +(99871) 292-98-78, e-mail: fatima_adilova@rambler.ru

Как правило, в виртуальном скрининге для построения компьютерных предикторов большинство пользователей используют QSAR-моделирование или подход на основе химического сходства в зависимости от их опыта и/или доступности инструмента. Целью работы является сравнение этих двух основных подходов на одном эталонном наборе данных, где прогностическая эффективность сопоставлялась бы с учетом точности прогнозирования как на обучающих, так и на внешних наборах данных. В обзоре представлены два метода прогноза биологической активности, реализованные в виде онлайн-пакетов, - SEA, PASS и схема KNN QSAR. Результаты вычислительных экспериментов по этим трем подходам показали преимущество схемы KNN QSAR. Методы, рассмотренные в работе, представляют интерес для химиков и экспериментальных биологов, работающих в области биологического скрининга химических библиотек.

**Ключевые слова:** виртуальный скрининг; QSAR моделирование; валидация моделей; пакеты PASS, SEA.

## COMPARISON OF THE COMPUTER PREDICTORS OF ORGANIC COMPOUNDS BIOLOGICAL ACTIVITY

Adilova F.T.

As a rule, in the virtual screening to build predictors most users use QSAR-modeling approach or approach based on chemical similarity; it's depend on their experience and / or tools available. The purpose was to review the investigations where compared to a reference dataset of these two basic approaches, predictive effectiveness of which was compared to that based on the prediction of both the training, and external data sets. We present two methods of biological activity prediction realized in the form of on-line packages, - SEA, PASS and KNN QSAR scheme. The results of computational experiments have shown the advantage of the latter scheme. The methods discussed in the work should be useful for chemists and experimental biologists working in the field of biological screening of chemical libraries.

**Keywords:** virtual screening, QSAR modeling, model validation, PASS, SEA.

## ORGANIK BIRIKMALARNING BIOLOGIK AKTIVLIGINI KOMPYUTERLI PREDIKTORLARINI TAQQOSLASH

Adilova F.T.

Odatda, Virtual skriningda ko'pchilik foydalanuvchilar kompyuterli prediktor qurish uchun QSAR modellashtirishdan foydaniladi yoki ularning tajriba va/yoki vositalari mavjudligiga qarab kimyoviy o'xshashlik asosida yondashiladi. Ishning maqsadi bu ikkita asosiy yondashuv bilan bitta etalondagi ma'lumotlarni solishtirish bo'yicha ishlarning tavsiflanishi, bunda o'rganilayotgan va tashqi ma'lumotlar to'plamlaridagi bashoratlash aniqligini hisobga olish bilan bashoratlash samaradorligi taqqoslandi. SEA, PASS va KNN QSAR- on-layn paketlar ko'rinishida amalga oshirilib, biologik aktivlikni prognoz qiladigan ikkita metod taqdim etilgan. Hisoblash natijalari ikkinchi sxema afzalligini ko'rsatdi. Mazkur ishda qaralgan metodlar kimyoviy biblotekalarni biologik skriningi sohasida ishlaydigan kimyogarlar va eksprimental biologlarga foydali bo'lishi kerak.

**Tayanch iboralar:** virtual skrining, QSAR modellashtirish, modellarni tekshirish, PASS, SEA paketlar.

## 1. Введение

Виртуальный скрининг (VS) является распространенным и эффективным подходом к открытию новых соединений. VS-методы классифицируются как методы на основе лиганда (LBVs) и структуры (SBVS) в зависимости от наличия кристаллических структур для интересующей цели. SBVS - наиболее популярный подход к выявлению предполагаемых активных соединений в химических библиотеках, но он требует знания трехмерной (3D) структуры белка-мишени. Если же структура белка-мишени неизвестна, что является более распространенным случаем, то часто используют LBVs подходы. Любой инструмент LBVs основан на принципе сходства, т.е. соединения со сходными химическими структурами, как ожидается, имеют сходные биологические свойства. Тогда можно прогнозировать

специфическую биологическую активность молекулы химически подобных соединений, для которых уже известны активности [1].

В последние годы, с ростом изученности сложных заболеваний, внимание привлекло новое направление в стратегии виртуального скрининга - полифармакология («много целей-много лекарств»), учитывающая системную регуляцию нескольких целей. Идея полифармакологии состоит в том, что пространства данных нового препарата, его цели в организме и болезни могут быть взаимосвязаны, что делает необходимым изучить функции препаратов в этих разных пространствах, и взаимосвязи последних. Это позволит использовать эти знания для разработки лекарств или их смесей, которые эффективно адресуются одной или нескольким болезням [2]. Для реализации исследований по виртуальному скринингу на основе лиганда в полифармакологии необходимо иметь сеть «лекарство-цель», чтобы прогнозировать большое число вероятных видов биологической активности вещества на основе его структурной формулы с использованием единообразного описания химической структуры и универсального алгоритма построения модели «структура-активность» (QSAR).

Сеть «лекарство-цель» (Drug-Target Network, или DTN) представляет собой двудольный граф, в котором каждое ребро указывает, например, на связь препарата с белком, если белок является известной целью препарата. Сегодня доступно большое количество молекулярных баз данных, которые постоянно растут в размерах и числе. Они интегрируют разнообразную информацию о молекулярных путях, кристаллических структурах, об экспериментах по связыванию, побочных эффектах и целях лекарственных препаратов.

Yildirim et. al. [5] описали ряд существенных особенностей, связанных с сетевой топологией сети DTN. В настоящее время применительно к небольшим молекулам представлена сетевая концепция химического пространства, идею которой впервые озвучил G.Maggiora [6]. Сегодня известны только пять работ, где использовали сетевую парадигму химических пространств [7-11], поэтому область сетей химического пространства (Chemical Space Networks или CSNs) является открытой для новых возможностей. Почему именно сети нужны для представления химического пространства? Помимо простоты аннотации, существует три причины, по которым формальные CSNs обеспечивают желательное представление химического пространства: 1) сети дают «естественное представление» химических пространств, они не страдают от «проклятия размерности», многие сети обладают фрактальной размерностью [12]; 2) сети обеспечивают соответствующую понятийную основу для статистического анализа многих аспектов химических пространств; 3) сегодня есть эффективные алгоритмы для анализа многих типов сетевых функций и для количественной оценки характеристик сети. Несмотря на убедительные основания к использованию *сетевого представления химического пространства, необходимы дополнительные исследования, поскольку* наш небольшой вычислительный эксперимент показал, что не всегда сети имеют преимущество перед традиционным описанием соединений в виртуальном скрининге [3].

## 2. Постановка задачи

Как правило, большинство пользователей LBVs используют QSAR или подход на основе химического сходства в зависимости от их опыта и/или доступности инструмента. Поскольку до 2016 г. не было опубликованных исследований, где бы на одном эталонном наборе данных сравнивались эти два основных LBVs подхода, а прогностическая эффективность нескольких методов LBVs сопоставлялась бы с учетом точности прогнозирования как на их внутренних (обучающих), так и на внешних наборах данных, то представляется интересным показать результаты исследования [13], в котором дополнительно включена схема KNN-QSAR.

Цель данного обзора - проанализировать преимущества и недостатки обозначенных выше подходов в решении конкретных задач оптимального выбора предикторов биологической активности. Обзор еще раз подчеркивает важность решения двух теоретических проблем в разработке прогнозирующих дескрипторов (предикторов) - оценки сходства и представления исходного набора данных дескрипторами (признаками описания соединения), которые могут улучшить качество прогнозов.

## 3. Материал и методы

Сравнивалась предсказательная сила PASS и SEA методов, доступных онлайн, на базе наборов данных соединений против нескольких мишеней, - G-белок спаренных рецепторов (GPCRs). Несколько наборов данных для GPCR лигандов использованы в анализе для следующих мишеней: серотониновые 5-HT1A, 5-HT1B, 5-HT1D, 5-HT2B, 5-HT6, 5-HT7 и Д5 дофаминовые рецепторы. Данные получены из Всемирной молекулярной базы данных по биоактивности (WOMBAT), Национального института психического здоровья (NIMH), программы скрининга психоактивных лекарств (PDSP), Ki базы данных и базы данных ChEMBL.

Изучалась производительность моделей, построенных с помощью методов SEA и PASS на фоне KNN QSAR моделей, построенных по схеме, разработанной в течение многих лет группой А.Тропши [14,15].

*Пакет PASS.* При наличии достаточно богатой коллекции разнообразных химических соединений страны СНГ обладают крайне ограниченными возможностями для их экспериментального тестирования, что требует тщательного отбора потенциально перспективных веществ уже на ранних стадиях исследования. Оказалось, что такой отбор

может быть осуществлен на основе компьютерного прогноза спектра биологической активности химических соединений. В основе пакета лежит понятие спектра биологической активности: совокупность фармакологических эффектов, биохимических механизмов действия и видов специфической токсичности, которые вещество может проявить при взаимодействии с биологическими объектами. В рамках этого определения рассматривается биологическая активность как «внутреннее» свойство вещества, которое проявляется при соответствующих условиях в эксперименте или клинике. При этом биологическая активность определяется лишь качественным образом (наличие/отсутствие), что является достаточно грубым описанием действительной ситуации, но при таком приближении в аналитических и прогностических целях можно использовать значительный объём информации о биологически активных соединениях, накопленный к настоящему времени.

В течение многих лет была создана компьютерная программа PASS, позволяющая прогнозировать большое число вероятных видов биологической активности вещества на основе его структурной формулы с использованием единого описания химической структуры и универсального математического алгоритма установления зависимостей «структура-активность» [4]. В PASS версии 2006 г. предсказание основано на QSAR анализе обучающей выборки, содержащей более 60.000 соединений. Здесь используют «многоуровневые соседства атомов» (Multilevel Neighborhoods of Atoms, MNA), как дескрипторы химической структуры, и алгоритм оценивания спектра активности как процедуру обучения. Алгоритм прогноза PASS отобран среди сотен различных вариантов и базируется на классическом байесовском подходе. Для любого нового соединения результаты даются в виде спектра активности, который является ранжированным списком вероятностей «быть активным - $P_a$ », или «быть неактивным - $P_i$ » и вида активности. Соединение считается активным, если значение ( $P_a$ - $P_i$ ) превышает пороговое значение, например, по умолчанию $(P_a - P_i) > 0$ . Кроме того, утверждается, что если $P_a > 0,7$ , то весьма вероятно, что экспериментальное значение активности доступно, и в этом случае вероятность того, что такое соединение является аналогом известного фармацевтического агента, также высока; если $0,5 < P_a < 0,7$ , то соединение, возможно, было экспериментально охарактеризовано, но вероятность этого низка, и соединение не очень похоже на известные фармацевтические агенты. Если $P_a < 0,5$ , соединение (или его близкий аналог) вряд ли экспериментально охарактеризованы, и если наличие этой активности подтверждается в эксперименте, соединение рассматривается в качестве кандидата в препараты.

*Пакет SEA.* В LBVs также популярен и подход на основе сходства, где требуется иметь, по крайней мере, один известный хит, в то время как модели QSAR могут быть разработаны при наличии достаточно большого набора данных биологически активных соединений. Процедуры оценки сходства сравнивают химические структуры непроверенных молекул, обычно кодируемые как отпечатки пальцев, с известными биологически активными молекулами, в то время как для построения и проверки моделей QSAR нужны специальные условия, прежде чем они могут быть использованы для виртуального скрининга. Этот подход реализован пакетом SEA (Similarity Ensemble Approach), который является средством для прогнозирования активности конкретной мишени, а также в качестве связывания лигандов вне мишени [16]. С помощью алгоритма BLAST он сравнивает белковые мишени по сходству соответствующих наборов лигандов, которые связываются с этими целями; точность прогнозирования выражается в виде средних значений. Коэффициенты Танимото ($T_C$) вычисляются для каждой пары лигандов из разных наборов. Необработанные оценки сходства между любыми двумя наборами лигандов рассчитываются как сумма Тс пары лигандов по всем парам с $T_c \geq 0,57$ , и значение этой оценки вычисляется после коррекции случайного ожидания. Таким образом, предсказание, будет ли лиганд связан с конкретной целью, может быть выполнено путем вычисления структурного сходства лиганда против всего набора лигандов для данной цели. Прогноз считается надежным, когда значение $p \leq 10^{-10}$ .

## 4. Результаты и обсуждение

Точность прогнозирования сравнивалась как для внутреннего (обучающего) набора соединений, так и для внешнего набора (из ChEMBL) для всех семи GPCRs. Число активных веществ для семи выбранных биологических мишеней в PDSP, WOMBAT и ChEMBL показано на рис. 1.

Неожиданно было обнаружено, что уровень точности на обучающей выборке пакета SEA находится около 70%, за исключением случая рецептора допамина *D5*, который содержал только 14 соединений. Причина низкой точности SEA, возможно, в том, что использовали только простое сравнение сходства при предсказании. Результаты прогнозирования были бы предвзятыми, если бы набор известных лигандов для рецептора был неравномерно распределен в химическом пространстве.

В предсказании на внешней выборке соединений дополнительно проявились преимущества использования мишень-специфичных, тщательно разработанных и проверенных моделей (KNN QSAR) по сравнению с SEA и PASS. Метод SEA показал самую низкую точность, поскольку смог точно предсказать менее половины набора соединений внешней проверки, в то время как модели KNN QSAR дали точность около 90% для большинства из

семи случаев. Точность предсказания для PASS намного выше, чем SEA, но ниже, чем у KNN QSAR (рис. 2). Это, скорее всего, потому, что в PASS использовали простой линейный регрессионный метод для внутреннего процесса построения модели в интересах автоматизации и скорости. При моделировании набор соединений для PASS недоступен (встроен в программное обеспечение), поэтому в прямом сравнении PASS с SEA и PASS с QSAR нет необходимости.

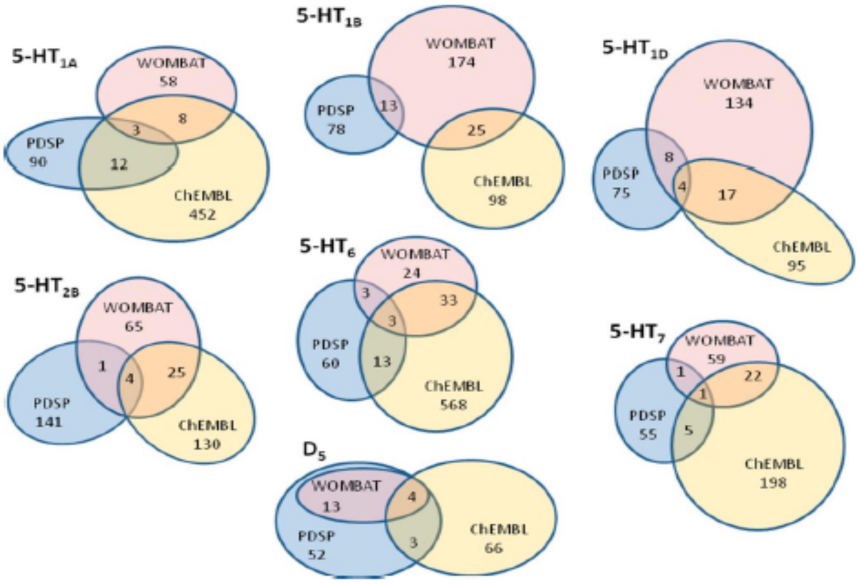


*Рис. 1. Диаграммы Венна, показывающие перехлест активных соединений для семи различных мишеней GPCR в PDSP, WOMBAT и ChEMBL*

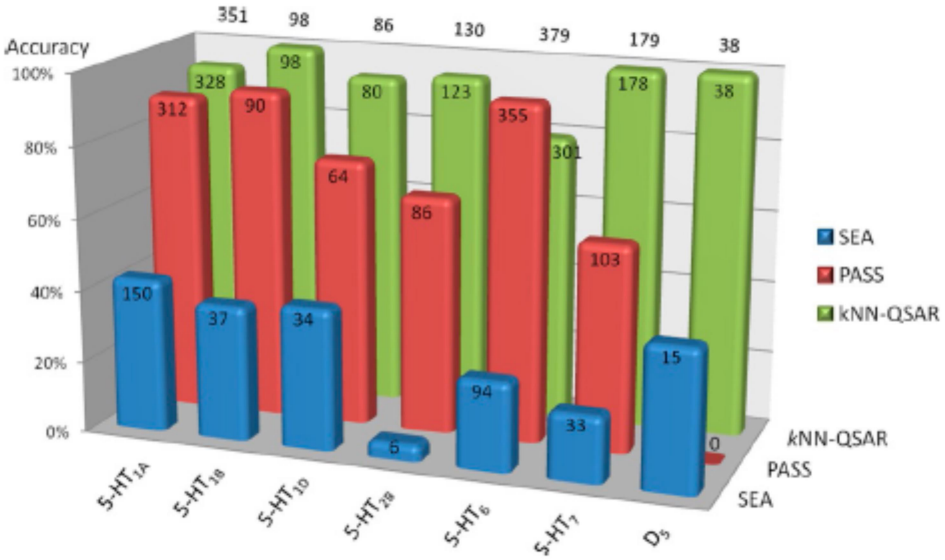


*Рис. 2. Сравнение внешней предсказательной силы для SEA, PASS и KNN-QSAR. KNN-QSAR модели для семи выбранных целей GPCR получены с использованием активных и неактивных веществ из PDSP. Одни и те же наборы уникальных соединений из ChEMBL для каждого целевого GPCR использованы в качестве набора внешней проверки для сравнения с числом соединений, показанных на верхней части графика. Число соединений, которые правильно предсказаны различными подходами, показаны на вершине каждого столбца*

Стоит отметить, что модели QSAR, сгенерированные для активов и неактивов из последовательного источника данных (PDSP), имеют более высокую внешнюю предсказательную силу, чем у моделей, генерируемых для активов и неактивов из смешанных источников (активов от WOMBAT и неактивов от PDSP).

Причина этого наблюдения неясна, но это может быть связано с областью применимости Applicability Domain (AD). D самом деле, внешняя точность прогнозирования для моделей, построенных с данными из WOMBAT, резко увеличилась с применением AD. По-видимому, соединения в наборах лигандов из WOMBAT очень похожи, и потому модели, построенные с использованием этих соединений, имеют очень плотный AD. Количество соединений, исключенных из предсказания из-за AD, варьирует среди различных мишеней GPCR, и в

среднем составляет 80%. Этот высокий уровень исключения продемонстрировал разницу химического пространства, представленную WOMBAT и PDSP соединениями.

## 5. Выводы

QSAR-методы оказались лучше химического подхода, основанного на сходстве как по внутренней, так и по внешней точности прогнозирования. Строго построенные и проверенные специфические модели QSAR показали самую высокую прогностическую силу для почти всех протестированных наборов данных GPCR. Программное обеспечение PASS, которое опирается на простой автоматизированный QSAR-подход, продемонстрировало умеренную прогностическую силу. Подход SEA на основе химического сходства показал низкую точность предсказания для всех испытанных семи случаях GPCR. Возможно, дополнительные исследования с использованием большего количества целей и дальнейшего совершенствования методологий оценки сходства помогут устранению недостатков в методе SEA.

Обзор ещё раз подчеркивает важность решения двух теоретических проблем в разработке предикторов биологической активности органических соединений, а именно, оценки сходства (1) и представления исходного набора данных (дескрипторы) (2), которые могут улучшить качество прогнозов, однако эта тема выходит за рамки настоящей работы.

## Литература


[1] *Адылова Ф.Т.*, *Давронов Р.Р.* Компьютерные технологии в создании новых лекарств: виртуальный скрининг // Журнал теоретической и клинической медицины. – 2013. – № 5. – С. 51-55.

[2] *Адылова Ф.Т.*, *Давронов Р.Р.* Полифармакология: новые стратегии виртуального скрининга // Журнал теоретической и клинической медицины. – 2015. – № 1. – С. 53-57.

[3] *Адылова Ф.Т.*, *Икрамов А.А.* Новая парадигма описания химического пространства в компьютерных приложениях: миф или реальность? // Узбекский химический журнал. – 2016. – № 4. – С. 95-100.

[4] *Филимонов Д.А.*, *Поройков В.В.* Прогноз спектра биологической активности органических соединений // Российский химический журнал. – 2006. – Т. 50, № 2. – С. 66-75.

[5] *Yildirim M.A. et. al.* Drug-target network // Nature Biotechnology. – 2007. – Vol. 25. – Pp. 1119-1126. – URL: https://goo.gl/4iP03k.

[6] *Maggiora G.*, *Bajorath J.* Chemical space networks: a powerful new paradigm for the description of chemical space // Journal of Computer-Aided Molecular Design. – 2014. – Vol. 28. – Issue 8. – Pp. 795-802. – DOI: 10.1007/s10822-014-9760-0.

[7] *Tanaka N. et. al.* Small-world phenomena in chemical library networks: application to fragment-based drug discovery // Journal of Chemical Information and Modeling. – 2009. – Vol. 49. – Issue 8. – Pp. 2677-2686. – DOI: 10.1021/ci900123v.

[8] *Krein M.P.*, *Sukumar N.* Exploration of the topology of chemical spaces with network measures // The Journal of Physical Chemistry A. – 2011. – Vol. 115. – Issue 45. – Pp. 12905-12918. – DOI: 10.1021/jp204022u.

[9] *Wawer M. et. al.* Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices // Journal of Medicinal Chemistry. – 2008. – Vol. 51. – Issue 19. – Pp. 6075-6084. – DOI: 10.1021/jm800867g.

[10] *Ripphausen P. et. al.* Rationalizing the role of SAR tolerance for ligand-based virtual screening // Journal of Chemical Information and Modeling. – 2011. – Vol. 51. – Issue 4. – Pp. 837-842. – DOI: 10.1021/ci200064c.

[11] *Stumpfe D.*, *Dimova D.*, *Bajorath J.* Composition and topology of activity cliff clusters formed by bioactive compounds // Journal of Chemical Information and Modeling. – 2014. – Vol. 54. – Issue 2. – Pp. 451-461. – DOI: 10.1021/ci400728r.

[12] *Cohen R.*, *Havlin S.* Scaling properties of complex networks and spanning trees // Handbook of large-scale random networks. – New York: Springer, 2009. – Pp. 143-169.

[13] *Man L. et. al.* Comparative Analysis of QSAR-based vs. Chemical Similarity Based Predictors of GPCRs Binding Affinity // Molecular Informatics. – 2016. – Vol. 35. – Issue 1. – Pp. 36-41. – DOI: 10.1002/minf.201500038.

[14] *Golbraikh A., Tropsha A.* Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection // Journal of Computer-Aided Molecular Design. – 2002. – Vol. 16. – Issue 5-6. – Pp. 357369.

[15] *Todd M. et. al.* Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? // Journal of Chemical Information and Modeling. – 2012. – Vol. 52. – Issue 10. – Pp. 2570-2578.

[16] *Keiser M.J.* Predicting new molecular targets for known drugs // Nature. – 2009. – No. 462(7270). – Pp. 175-181. – DOI: 10.1038/nature08506.