УДК 544.165

ВЫЯВЛЕНИЕ ВЫБРОСОВ ПРИ QSAR-МОДЕЛИРОВАНИИ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ СОЕДИНЕНИЙ НА ОСНОВЕ АНАЛИЗА КАРТ «СТРУКТУРА – АКТИВНОСТЬ – ПОДОБИЕ»

Л.Д. Григорьева[1]*, В.Ю. Григорьев[2], А.В. Ярков[2]

*([1]Факультет фундаментальной физико-химической инженерии МГУ имени М.В. Ломоносова; [2]Институт физиологически активных веществ РАН; *e-mail: ldg@)physchem.msu.ru)*

На основе анализа карт «структура – активность – подобие» (SAS) разработан новый метод для выявления выбросов в обучающих выборках при конструировании QSAR-моделей. Он заключается в эмпирической оценке вероятности появления химического соединения в той или иной области SAS. В качестве выбросов предложено рассматривать соединения, которые имеют максимальную вероятность появления в области «скачков активности» и минимальную вероятность в «гладкой» области. Предложенный метод может быть использован в области медицинской химии для поиска новых перспективных биологически активных химических соединений.

Ключевые слова: QSAR, SAR, выбросы, карты «структура – активность – подобие».

В настоящее время широкое распространение получили исследования с использованием методологии, основанной на существовании количественной связи между структурой и активностью (QSAR) [1–3]. В основе этого междисциплинарного подхода лежат фундаментальные принципы: 1) структура соединений определяет их свойства, 2) подобные вещества действуют подобным образом [4, 5].

Основная цель QSAR заключается в конструировании статистических моделей зависимости активности молекул от их структуры на основе методов машинного обучения [6]. Один из необходимых элементов этого обучения – формирование обучающих выборок.

QSAR-исследование представляет собой многоступенчатый процесс [7], в ходе которого проводится анализ на выбросы [8]. Под выбросами в широком плане подразумеваются некоторые выделенные данные, существенно отличающиеся от оставшихся. Анализ на выбросы представляет собой составную часть исследований, проводимых в области компьютерных наук [9]. Цель этого этапа – выявление химических соединений, отличающихся по определенным критериям от других. Как правило, удаление таких соединений из обучающей выборки приводит к улучшению статистических характеристик количественных моделей. Нужно отметить, что выявление выбросов может происходить как до, так и после конструирования моделей. Для этого используются разные подходы. Достаточно широкое распространение получили статистические методы анализа на Y-, X-, XY-выбросы [10]. В их основе лежит использование разных статистических величин. Один из самых простых подходов для анализа Y-выбросов состоит в изучении стандартизованных остатков. При этом в качестве выбросов рассматриваются точки, имеющие два-три стандартных отклонения [11]. Более усложненная процедура – графический анализ нормальных вероятностей стандартизованных остатков. X-выбросы связаны с дескрипторами, которые не попадают в область допустимых значений обучающего ряда. Для их детектирования достаточно широко применяются проекционные методы, основанные на анализе X-остатков, определении расстояния до модельного центра и использовании статистики Стьюдента [12]. Появление X/Y-выбросов обусловлено наличием в тестовой выборке химических соединений, у которых соотношения между зависимыми и независимыми переменными не такие, как в обучающей выборке. Скорее всего, X-выбросы можно рассматривать как X/Y-выбросы.

Для выявления потенциальных выбросов могут быть использованы методы, основанные на анализе карт «структура – активность – подобие» (Structure – Activity – Similarity, SAS) [13, 14]. В основу создания этих карт положен принцип попарного сравнения молекул с использованием количественных характеристик подобия. При этом предполагается, что пары молекул, которые

близки по структуре, но сильно отличаются по активности (так называемые «скачки активности» (activity cliffs, AC)), могут быть потенциальными выбросами при QSAR-моделировании. С одной стороны, такие молекулы препятствуют созданию адекватных QSAR-моделей, но с другой стороны, представляют интерес для медицинских химиков как объекты для дальнейших исследований [15]. Обнаружить «скачки активности» можно с помощью таких методов, как индексы SALI (Structure-Activity Landscape Index), SARI (Structure – Activity Relationship Index), анализ MMP (Matched Molecular Pair) и др. [16]. Следует обратить внимание на количественный подход, основанный на индексе SALI [17], который находит широкое применение при анализе карт «структура – активность – подобие».

Нужно отметить, что основное внимание исследователей направлено на изучение области AC, однако потенциальным источником информации о возможных выбросах могут служить и другие области SAS. В этом плане интересным представляется исследование «гладкой области» (smooth region, SR), в которой находятся пары молекул, имеющие высокую степень подобия как по структуре, так и по активности. Низкая вероятность появления того или иного соединения в этой области может служить индикатором того, что оно станет выбросом.

Цель настоящего исследования – разработка статистического метода анализа карт «структура – активность – подобие» в области скачков активности и гладкой области для выявления потенциальных выбросов в QSAR-моделях.

Биологические и структурные данные

В работе использованы два ряда литературных данных по биологической активности. Первый ряд состоит из 32 бензимидазолов (рис. 1, табл. 1), обладающих антипаразитарной активностью по отношению к *Trichomonas vaginalis* [18]. Данные по биологической активности выражены в виде $\log(1/IC_{50})$, где IC_{50} – концентрация (M), вызывающая 50%-е ингибирование.

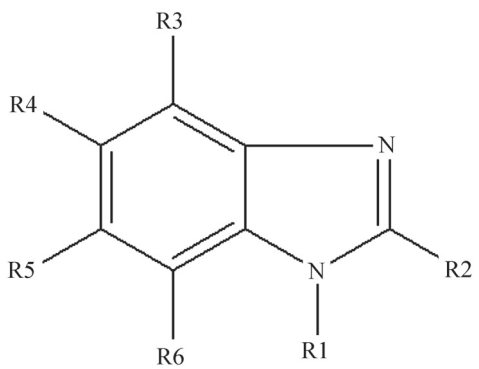

Рис. 1. Структурная формула бензимидазолов

Второй исследуемый ряд соединений включает 47 различных органических растворителей, для которых по отношению к крысам были экспериментально определены величины нейротоксичности, представленные в виде EC_{30} (концентрация (мкМ), вызывающая нейротоксический эффект у 30% животных). [19] Минимальная величина $\log(1/EC_{30})$ составляла –2,939, а максимальная –0,568.

Меры структурного сходства, индекс SALI и дескрипторы соединений

При создании карт «структура – активность – подобие» для описания молекулярной структуры использовали молекулярные отпечатки двух видов: MACCS (166 бит) и ECFP4 (1024 бит). Их величины рассчитаны с помощью программы OpenBabel [20]. Оценку структурного сходства молекул проводили на основе бинарных коэффициентов Танимото (T_c) [21]. Подобие молекул по активности ($S_{акт.}$) рассчитывали по формуле [13]:

$$S_{акт.}(A, B) = 1 - |Act(A) - Act(B)|/(Act_{макс} - Act_{мин}),$$

где Act(A) и Act(B) – значения активности соединений A и B, $Act_{макс}$ и $Act_{мин}$ – максимальные и минимальные значения активности соответственно.

Индекс SALI оценивали по формуле, представленной в публикации [17]:

$$SALI(A, B) = |Act(A) - Act(B)|/(1 - T_c).$$

Для пар соединений, у которых $T_c = 0$, величину SALI принимали равной максимальному значению индекса оставшихся соединений. Такого рода пары ранжировали по величинам |Act(A) – Act(B)|.

Для QSAR-моделирования рассчитали два ряда дескрипторов: 27 физико-химических дескрипторов с использованием компьютерной программы HYBOT [22] и 52 дескриптора (электронные, топологические, физико-химические) с использованием прграммы DNESTR [23]. Отбор дескрипторов осуществляли на основе анализа величин парных коэффициентов корреляции при пороговом значении 0,80–0,95. В случае необходимости проводили автошкалирование переменных по формуле:

$$Z^* = (Z - Z_m)/SD,$$

Таблица 1

Химические структуры и биологическая активность $\log(1/IC_{50})$ (IC_{50}, М) бензимидазолов [18]

Номер соединения	R_1	R_2	R_3	R_4	R_5	R_6	$\log(1/IC_{50})$
1	H	CF_3	H	H	H	H	5,50
2	CH_3	CF_3	H	CF3	H	H	5,39
3	CH_3	CF_3	H	H	CF_3	H	5,27
4	CH_3	CF_3	H	$CH_3CH_2CH_2S$	H	H	6,70
5	CH_3	CF_3	H	H	$CH_3CH_2CH_2S$	H	5,59
6	CH_3	CF_3	H	$C_6H_5C(O)$	H	H	4,53
7	CH_3	CF_3	H	H	$C_6H_5C(O)$	H	4,97
8	H	CF_3	H	Br	Br	H	6,66
9	H	CF_3	Br	Br	Br	Br	8,70
10	H	C_2F_5	H	Cl	Cl	H	6,52
11	H	CF_3	H	NO_2	NO_2	H	6,24
12	H	C_2F_5	Br	Br	Br	Br	5,00
13	CH_3	$CONH_2$	H	H	Cl	H	6,96
14	CH_3	$CONHCH_3$	H	H	Cl	H	6,98
15	CH_3	$CON(CH_3)_2$	H	H	Cl	H	6,63
16	CH_3	$COOCH_2CH_3$	H	H	Cl	H	7,72
17	CH_3	$CONH_2$	H	Cl	H	H	6,73
18	CH_3	$CONHCH_3$	H	Cl	H	H	6,45
19	CH_3	$CON(CH_3)_2$	H	Cl	H	H	6,68
20	CH_3	$COOCH_2CH_3$	H	Cl	H	H	7,57
21	CH_3	$CONH_2$	H	Cl	Cl	H	6,87
22	CH_3	$CONHCH_3$	H	Cl	Cl	H	6,65
23	CH_3	$CON(CH_3)_2$	H	Cl	Cl	H	7,12
24	CH_3	$COOCH_2CH_3$	H	Cl	Cl	H	7,53
25	CH_3	$CONH_2$	H	H	H	H	6,78
26	CH_3	$CONHCH_3$	H	H	H	H	6,98
27	CH_3	$CON(CH_3)_2$	H	H	H	H	6,37
28	CH_3	$COOCH_2CH_3$	H	H	H	H	7,07
29	CH_3	$COCH_3$	H	H	H	H	6,68
30	CH_3	$COCH_3$	H	Cl	H	H	6,88
31	CH_3	$COCH_3$	H	H	Cl	H	6,64
32	CH_3	$COCH_3$	H	Cl	Cl	H	7,20

где Z и Z^* – исходная и автошкалированная переменные соответственно, Z_m – среднее значение переменной, SD – стандартное отклонение.

Статистические методы

Для QSAR-моделирования применяли три метода: множественную линейную регрессию (MLR), метод случайного леса (RF) и метод опорных векторов (SVM). Для расчета коэффициентов линейных уравнений использовали компьютерную программу SVD [24]. Особенность этой программы заключается в применении сингулярного разложения матрицы данных, позволяющего эффективно выявлять линейные зависимости между переменными. Регрессионные модели на основе RF сконструированы с помощью оригинальной программы автора метода (Leo Breiman) [25]. В качестве контролируемых использовали такие параметры, как число деревьев (jbt = 500), число случайно выбираемых дескрипторов (mtry = mdim/3, где mdim – число дескрипторов), число объектов в узле, ниже которого не происходит расщепление деревьев (nthsize = 5). Расчеты SVM-моделей проводили с помощью программы flssvm [26]. Предварительно дескрипторы соединений автошкалировали. В качестве ядерной функции применяли радиальную базисную функцию.

Для оценки статистических характеристик QSAR-моделей использовали следующие параметры: n – число молекул; m – число дескрипторов; r^2 – квадрат коэффициента линейной корреляции; rmse – среднеквадратичная ошибка; r^2_{cv} – квадрат коэффициента линейной корреляции в условиях перекрестного контроля; $rmse_{cv}$ – среднеквадратичная ошибка в условиях перекрестного контроля; r^2_{rand} – квадрат рандомизированного коэффициента корреляции, который рассчитывали в соответствии с публикацией [27], используя 10 итераций: $r^2_{rand} = (r^2)^{0,5}\ (r^2 - r^2_r)^{0,5}$, где r^2_r представляет собой усредненное значение квадрата коэффициента линейной корреляции для рандомизированных моделей (таких моделей, где матрица дескрипторов остается неизменной, а зависимая переменная используется в случайном порядке); FIT – фитнесс-функция, представляющая собой модификацию критерия Фишера [28]. При конструировании RF-моделей для их тестирования использовали процедуру out-of-bag, представляющую собой составную часть алгоритма. Учитывая небольшой размер исходной выборки, тестовую выборку из нее не выделяли. Тестирование моделей проводили путем перекрестного контроля с выбором по 5 и использованием 10 итераций. В качестве основной стратегии выбора дескрипторов для конструирования моделей использовали метод полного перебора комбинаций из 1, 2, … 5 дескрипторов. Модели сравнивали по величине FIT обучающей выборки и выбирали модель с ее максимальным значением.

Результаты и их обсуждение

Антипаразитарная активность. В работе были использованы простые SAS-карты (рис. 2). Каждая точка соответствует паре молекул, которые имеют определенные величины подобия по структуре (T_c) и активности ($S_{акт.}$).

На основе границ подобия (пунктирные линии на рис. 2) все пространство карт может быть поделено на четыре области. В области I находятся молекулярные пары, имеющие высокую степень молекулярного подобия, но сильно отличающиеся по активности, скачки активности (activity cliffs, AC). Молекулы, имеющие низкие величины подобия как по структуре, так и по активности, попадают в область II – область неопределенного вида (nondescript region, NR). Область III содержит молекулярные пары, имеющие низкую степень сходства по структуре, но близкие по активности структурные скачки (structure cliffs, SC). В область IV попадают молекулы, подобные как по структуре, так и по активности – гладкая область (smooth region, SR).

Основные факторы, влияющие на распределение молекул:

1) количественные меры подобия,
2) способ описания структуры,
3) границы подобия.

В настоящей работе влияние первого фактора не исследовалось. Для описания структуры соединений использовали два популярных типа молекулярных отпечатков: линейный MACCS и сферический ECFP4. Что касается границ подобия, то в литературе приведены разные значения. Так, в случаях MACCS и ECFP4 рекомендуется применять величины T_c, равные 0,85 и 0,55 соответственно [29]. Очевидно, что эти рекомендации носят эмпирический характер. Чтобы сделать картину более объективной, мы использовали комбинаторный подход, заключающийся в применении девяти границ подобия по структуре и девяти границ подобия по активности в интервале от 0,1 до 0,9 с шагом 0,1, т.е. общее число комбинаций границ подобия по структуре и активности должно быть равным 81.

Следует отметить, что SAS-карты содержат данные о парном распределении молекул. Для достижения цели нашей работы необходимо извлечь информацию об индивидуальных молекулах. Для этого рассчитаем относительные частоты

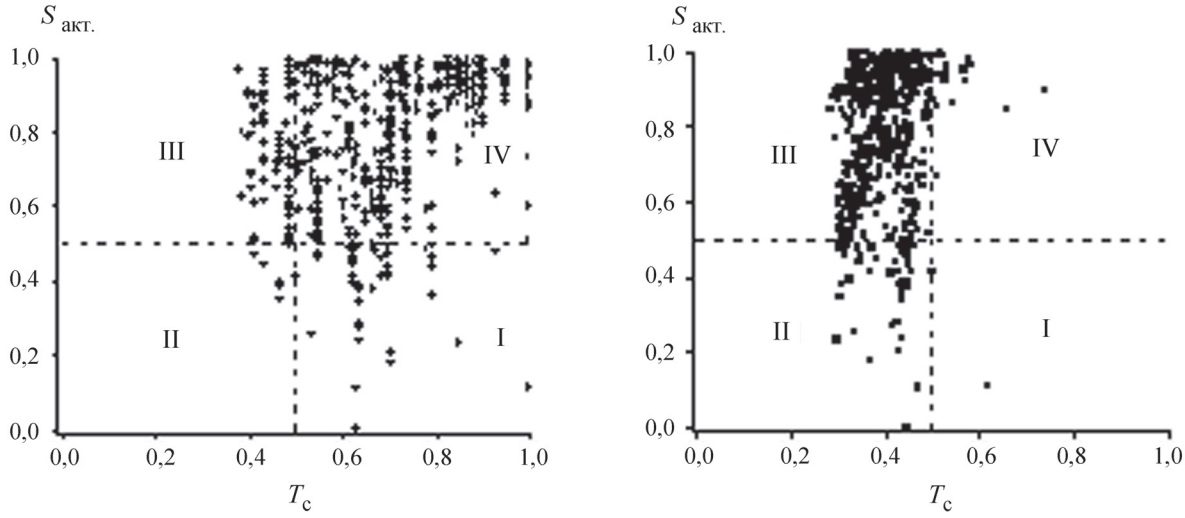

Рис. 2. SAS карты на основе a) MACCS и b) ECFP4 молекулярных «отпечатков»

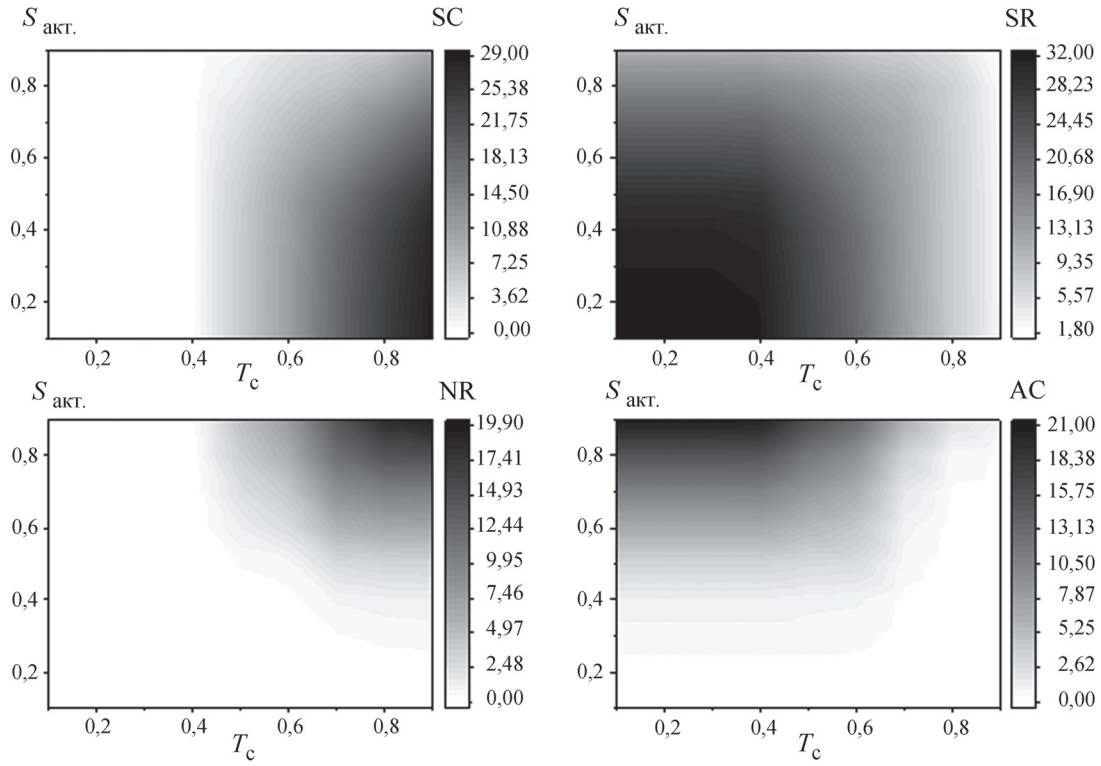

Рис. 3. Двумерные функции распределения суммарных относительных частот встречаемости молекул в различных областях SAS в зависимости от порогов сходства по структуре и активности при использовании молекулярных отпечатков MACCS

встречаемости молекул в разных областях SAS (F_i^k) по формуле:

$$F_i^k = M_i^k/(n-1);\ i = 1, 2, \ldots n;\ k = 1, 2, \ldots 4,$$

где n – общее число молекул; M_i^k – число молекулярных пар, содержащих i-ю молекулу в k-й области. Очевидно, что полученные величины F зависят от границ подобия. Общее представление об этой зависимости можно получить из двумерного распределения суммарных частот встречаемости молекул (ΣF) для четырех областей SAS (рис. 3).

На рис. 3 величине распределения соответствует цвет, который может меняться от белого (минимум) до черного (максимум) через градации серого. При этом значение ΣF может варьировать от 0 до 32. Из представленных данных следует, что

Таблица 2

Относительные частоты встречаемости соединений (F^{I}, F^{II}, F^{III}, F^{IV}) в соответствующих квадрантах SAS-карт при использовании двух видов молекулярных «отпечатков» (MACCS, ECFP4)

Номер соединения	MACCS				ECFP4			
	F^{I}	F^{II}	F^{III}	F^{IV}	F^{I}	F^{II}	F^{III}	F^{IV}
1	0,141	0,128	0,338	0,393	0,102	0,167	0,450	0,282
2	0,215	0,089	0,187	0,509	0,106	0,198	0,454	0,241
3	0,224	0,095	0,181	0,500	0,109	0,210	0,442	0,239
4	0,083	0,042	0,348	0,526	0,049	0,076	0,519	0,356
5	0,156	0,102	0,288	0,454	0,105	0,153	0,442	0,300
6	0,352	0,160	0,137	0,350	0,229	0,284	0,261	0,226
7	0,275	0,126	0,171	0,427	0,180	0,222	0,323	0,276
8	0,088	0,037	0,375	0,499	0,048	0,077	0,571	0,303
9	0,326	0,194	0,219	0,262	0,193	0,327	0,315	0,165
10	0,088	0,038	0,364	0,511	0,049	0,077	0,532	0,342
11	0,074	0,087	0,472	0,366	0,057	0,104	0,566	0,272
12	0,215	0,165	0,247	0,373	0,135	0,245	0,411	0,209
13	0,083	0,049	0,191	0,677	0,048	0,085	0,546	0,321
14	0,084	0,049	0,177	0,691	0,053	0,080	0,511	0,356
15	0,075	0,043	0,211	0,670	0,046	0,072	0,515	0,366
16	0,185	0,102	0,199	0,514	0,117	0,170	0,411	0,303
17	0,078	0,044	0,196	0,682	0,047	0,074	0,521	0,358
18	0,095	0,041	0,185	0,679	0,053	0,083	0,526	0,338
19	0,080	0,049	0,206	0,665	0,048	0,081	0,528	0,343
20	0,165	0,093	0,208	0,534	0,100	0,158	0,448	0,294
21	0,082	0,047	0,193	0,678	0,046	0,083	0,548	0,323
22	0,083	0,043	0,183	0,691	0,048	0,077	0,507	0,368
23	0,106	0,063	0,192	0,640	0,066	0,103	0,496	0,336
24	0,157	0,090	0,211	0,542	0,092	0,155	0,472	0,281
25	0,071	0,051	0,243	0,635	0,047	0,074	0,521	0,358
26	0,078	0,054	0,225	0,642	0,052	0,081	0,518	0,350
27	0,084	0,049	0,252	0,615	0,052	0,080	0,507	0,360
28	0,092	0,059	0,292	0,557	0,059	0,092	0,511	0,339
29	0,084	0,045	0,245	0,626	0,049	0,080	0,508	0,363
30	0,088	0,041	0,192	0,679	0,051	0,078	0,499	0,372
31	0,087	0,035	0,198	0,680	0,047	0,075	0,524	0,354
32	0,123	0,053	0,180	0,644	0,065	0,111	0,505	0,319

распределение носит различный характер в зависимости от области SAS. В качестве общей черты можно отметить, что плотность распределения минимальна при комбинации T_c и $S_{акт.}$, которая соответствует физическому смыслу данной области SAS, и максимальна для противоположной комбинации переменных. Так, в области АС находятся пары молекул, имеющие высокую степень подобия по структуре и низкую степень подобия по активности. Величина распределения при этом минимальна ($\Sigma F = 0{,}00$) при $T_c = 0{,}9$ и $S_{акт.} = 0{,}1$. Напротив, При $T_c = 0{,}1$ и $S_{акт.} = 0{,}9$ величина ΣF, равная 20,97, максимальна. Очевидно, что в первом случае число скачков активности будет очень мало, а во втором – очень велико. Для учета влияния всех комбинаций порогов структурного сходства и сходства по активности мы использовали простое (без взвешивания) суммирование величин F_i^k с дальнейшим усреднением. Сходные данные были получены и при применении для описания структуры соединений молекулярных отпечатков ECFP4. Результаты соответствующих расчетов приведены в табл. 2.

Каждая строка в табл. 2 содержит относительные частоты встречаемости соединения в той или иной области SAS-карт. При этом сумма четырех величин F равна 1. Эти величины в определенной степени можно рассматривать как оценку вероятности попадания молекулы в разные области SAS-карты. Минимальная величина для всех значений F (MACCS) составила 0,035, максимальная – 0,691, среднее значение 0,250, стандартное отклонение 0,206. Для ECFP4 эти величины были равны 0,046; 0,571; 0,250 и 0,171 соответственно. Очевидно, что в случае использования MACCS размах и рассеяние величин F выше по сравнению с ECFP4.

В табл. 3 представлены коэффициенты корреляции Пирсона (r_{xy}) для четырех величин F.

Для MACCS данных сильная ($|r_{xy}| > 0{,}8$) межгрупповая линейная связь наблюдается между величинами F^I/F^{II} и F^{II}/F^{IV}. В первом случае она является положительной, во втором – отрицательной. Данные для ECFP4 характеризуется большим разнообразием. Кроме отмеченных двух пар сильная корреляция существует между величинами F^I/F^{III}, F^I/F^{IV} и F^{II}/F^{III}. При этом величины $|r_{xy}|$ имеют более высокое значение по сравнению с соответствующими MACCS-данными. Можно также отметить существование межвидовой положительной линейной связи при сравнении соответствующих данных для MACCS и ECFP4 в случае F^I, F^{II} и F^{IV}.

Полученная из SAS-карт информация о распределении индивидуальных молекул может быть использована различным образом. В частности, появляется возможность выявления выбросов при QSAR-моделировании на основе фундаментального принципа «подобные соединения действуют подобным образом». Причем этот процесс не

Таблица 3

Коэффициенты корреляции Пирсона (r_{xy}) для двух видов молекулярных отпечатков (MACCS, ECFP4) и четырех относительных частот встречаемости соединений (F^I, F^{II}, F^{III}, F^{IV})

Номер соединения	r_{xy}							
	MACCS				ECFP4			
	F^I	F^{II}	F^{III}	F^{IV}	F^I	F^{II}	F^{III}	F^{IV}
1	1,000	0,891	–0,322	–0,744	0,981	0,973	–0,954	–0,860
2		1,000	–0,075	–0,863	0,921	0,959	–0,870	–0,901
3			1,000	–0,379	–0,263	–0,220	0,378	0,007
4				1,000	–0,779	–0,812	0,674	0,851
5					1,000	0,958	–0,975	–0,827
6						1,000	–0,913	–0,930
7							1,000	0,710
8								1,000

Таблица 4

Автошкалированные относительные частоты встречаемости соединений в области скачков активности (F^{I*}) и в «гладкой области» (F^{IV*})

Номер соединения	MACCS		ECFP4	
	F^{I*}	F^{IV*}	F^{I*}	F^{IV*}
1	0,115	–1,373	0,470	–0,598
2	1,095	–0,423	0,562	–1,366
3	1,210	–0,495	0,612	–1,412
4	–0,636	–0,278	–0,643	0,824
5	0,313	–0,877	0,536	–0,248
6	2,883	–1,729	3,138	–1,655
7	1,877	–1,094	2,101	–0,712
8	–0,568	–0,498	–0,660	–0,187
9	2,544	–2,462	2,385	–2,819
10	–0,578	–0,403	–0,651	0,558
11	–0,750	–1,597	–0,467	–0,773
12	1,084	–1,541	1,156	–1,975
13	–0,636	0,965	–0,668	0,163
14	–0,630	1,081	–0,568	0,824
15	–0,740	0,913	–0,710	1,022
16	0,694	–0,377	0,779	–0,195
17	–0,709	1,012	–0,677	0,855
18	–0,479	0,985	–0,559	0,474
19	–0,672	0,870	–0,668	0,573
20	0,433	–0,212	0,436	–0,362
21	–0,656	0,979	–0,702	0,193
22	–0,641	1,087	–0,660	1,045
23	–0,338	0,660	–0,292	0,436
24	0,329	–0,146	0,260	–0,613
25	–0,792	0,620	–0,677	0,855
26	–0,698	0,679	–0,585	0,703
27	–0,625	0,456	–0,576	0,900
28	–0,526	–0,021	–0,434	0,497
29	–0,630	0,548	–0,643	0,961
30	–0,568	0,982	–0,593	1,121
31	–0,589	0,995	–0,685	0,794
32	–0,114	0,696	–0,317	0,117

связан с видом статистических моделей и предшествует этапу их разработки. С теоретической точки зрения, выбросы должны иметь максимальную вероятность появления в квадранте I (AC) и минимальную – в квадранте IV (SR). Учитывая характер распределения величин F, необходимо использовать граничные значения, которые позволили бы отделить выбросы от остальных молекул. Но принимая во внимание эмпирический характер SAS-карт, введение одинаковых границ для частот встречаемости соединений (F) не представляется возможным. В качестве одного из решений предлагается использование автошкалированных величин F* (табл. 4), которые, по сути, представляют собой отклонение F от среднего значения, выраженное в единицах стандартного отклонения.

В нашей работе в качестве потенциальных выбросов рассматривались соединения, у которых величина $F^{I*} \geq 2$, $F^{IV*} \leq -2$. В случае MACCS это соединения 6 и 9, а для ECFP4 – 6, 7 и 9. С учетом необходимости соответствия одновременно двум критериям в качестве единственного выброса может выступать только соединение 9. Полученный результат хорошо согласуется с данными публикации [18]. В этой работе на основе консенсусного анализа SAS-карт было установлено наличие двух молекулярных пар (соединения 1/9 и 9/12) в области AC, которые потенциально могут быть выбросами при QSAR-моделировании.

В табл. 5 представлены результаты QSAR-моделирования антипаразитарной активности бензимидазолов по отношению к *Trichomonas vaginalis*. Модели 1–3 получены для полного ряда из 32 соединений. Модели 4–6 не содержат соединения 9. Очевидно, что моделирование с учетом выброса улучшает статистические характеристики. Модели 4–6 по своим характеристикам удовлетворяют современным требованиям QSAR ($n/m \geq 4$; $r^2 > 0{,}6$; $r^2_{cv} > 0{,}5$; $r^2_{rand} > 0{,}5$) [27, 30, 31]). При этом интервал изменения биологической активности превышает три логарифмические единицы. В состав моделей входит небольшое число дескрипторов с ясной физико-химической интерпретацией. Это дескрипторы, которые связаны с разными видами межмолекулярного взаимодействия: электростатическим ($q^-_{макс}$, $\Sigma|q|$, $\Sigma(q^+)$), дисперсионным (α) и водородной связью (ΣC_a).

Для выявления скачков активности был также использован индекс SALI. В этом случае наиболее вероятными кандидатами на выбросы были соединения, имеющие максимальные значения этого индекса. При описании структуры соединений в виде двух типов молекулярных отпечатков

ими оказалась пара соединений 9 и 12. Величина SALI составляла 30,520 для MACCS и 9,709 для ECFP4. Полученные с учетом этих выбросов регрессионные модели 7–9 по своим статистическим характеристикам оказались близки к моделям 4–6. Однако это было достигнуто при меньшем числе соединений.

Нейротоксичность. Результаты, сходные с данными исследования антипаразитарной активности, были получены в результате моделирования нейротоксического эффекта 47 органических растворителей. Для этого ряда соединений были проведены вышеописанные процедуры, т.е. рассчитаны функции распределения суммарных относительных частот встречаемости молекул и их относительные частоты встречаемости в разных областях SAS, а также определены потенциальные выбросы.

В частности, на основе описанного выше метода оценки вероятности появления соединения в той или иной области карты SAS при использовании порога в 2,5 стандартных отклонения, было установлено наличие двух потенциальных выбросов. Одно из соединений (*н*-пентан) имеет величину $\log(1/EC_{30}) = -2{,}939$, а у другого (1,1,2,2–тетрахлорэтана) $\log(1/EC_{30}) = -0{,}568$. Регрессионные модели 10–15 (табл. 6) были получены с использованием набора дескрипторов, отличающихся от дескрипторов моделей 1–9, представленных в табл. 5. Главное отличие состоит в появлении топологических дескрипторов. Однако конечный эффект сходен: удаление потенциальных соединений-выбросов из обучающей выборки приводит к улучшению статистических характеристик регрессионных моделей.

Анализ величин SALI для исследованных соединений показал, что в качестве примера скачков активности нужно рассматривать две пары молекул: 1-пропанол ($\log(1/EC_{30}) = -2{,}713$) и 2-метил-1-пропанол ($\log(1/EC_{30}) = -2{,}188$) (SALI = 4,160; MACCS), а также *н*-пентан ($\log(1/EC_{30}) = -2{,}939$) и 2-ксилол ($\log(1/EC_{30}) = -1{,}176$) (SALI = 7,154; ECFP4). QSAR-модели с учетом этих выбросов представлены в табл. 6 (16–18). В целом, соответствующие модели 13–15 и 16–18 близки между собой. Однако, как и в случае с изучением антипаразитарной активности соединений, этот результат достигается за счет удаления бóльшего числа соединений из обучающей выборки.

Таким образом, на основе анализа карт «структура – активность – подобие» предложен новый метод для выявления выбросов в обучающих выборках при конструировании QSAR-моделей. По сути, он заключается в эмпирической оценке вероятности появления химического соединения в той или иной области SAS. В качестве выбросов предложено рассматривать соединения, имеющие максимальную вероятность появления в области скачков активности и минимальную вероятность появления в гладкой области. На примере исследования антипаразитарной активности ряда бензимидазолов по отношению к *Trichomonas vaginalis* выявлен один выброс. При использовании индекса SALI обнаружены два выброса. Их удаление

Таблица 5

Статистические характеристики QSAR-моделей, антипаразитарная активность

Номер модели	Модель	*m*	Дескрипторы*	*n*	r^2	rmse	FIT	r^2_{cv}	$rmse_{cv}$	r^2_{rand}
1	MLR	3	$q^-_{макс}$, $\Sigma\|q\|$, ΣC_a	32	0,629	0,53	1,16	0,464	0,63	0,585
2	RF	2	α, $\Sigma(q^+)/\alpha$	32	0,874	0,31	5,56	0,668	0,50	0,610
3	SVM	2	$\Sigma\|q\|$, $\Sigma(C_{ad})/\alpha$	32	0,901	0,27	7,34	0,555	0,58	0,598
4	MLR	3	$q^-_{макс}$, $\Sigma\|q\|$, ΣC_a	31	0,706	0,43	1,62	0,584	0,51	0,669
5	RF	2	α, $\Sigma(q^+)/\alpha$	31	0,881	0,27	5,92	0,659	0,46	0,633
6	SVM	2	$\Sigma\|q\|$, $\Sigma(C_{ad})/\alpha$	31	0,901	0,25	7,24	0,564	0,52	0,613
7	MLR	3	$q^-_{макс}$, $\Sigma\|q\|$, ΣC_a	30	0,710	0,41	1,64	0,593	0,48	0,679
8	RF	2	α, $\Sigma(q^+)/\alpha$	30	0,865	0,28	5,11	0,610	0,47	0,568
9	SVM	2	$\Sigma\|q\|$, $\Sigma(C_{ad})/\alpha$	30	0,930	0,20	10,49	0,546	0,51	0,591

* $q^-_{макс}$ – максимальный отрицательный парциальный атомный заряд; $\Sigma|q|$ – сумма абсолютных величин парциальных атомных зарядов; ΣC_a – сумма H-акцепторных свободноэнергетических дескрипторов; α – молекулярная поляризуемость; $\Sigma(q^+)$ – сумма положительных парциальных атомных зарядов; ΣC_{ad} – сумма H-акцепторных и H-донорных свободноэнергетических дескрипторов.

Таблица 6

Статистические характеристики QSAR-моделей, нейротоксичность

Номер модели	Модель	*m*	Дескрипторы*	*n*	r^2	rmse	FIT	r^2_{cv}	$rmse_{cv}$	r^2_{rand}
10	MLR	4	$q^+_{макс}$, SIC1, MR, ESS	47	0,684	0,30	1,44	0,613	0,33	0,647
11	RF	3	$q^+_{макс}$, C_a^{max}, *Q*	47	0,840	0,21	4,03	0,664	0,30	0,580
12	SVM	2	SPAC, ESS	47	0,713	0,28	2,14	0,561	0,35	0,540
13	MLR	4	$q^+_{макс}$, SIC1, MR, ESS	45	0,708	0,26	1,59	0,648	0,28	0,651
14	RF	3	$q^+_{макс}$, $C_a^{макс}$, *Q*	45	0,845	0,19	4,14	0,659	0,28	0,580
15	SVM	2	SPAC, ESS	45	0,721	0,25	2,22	0,552	0,32	0,513
16	MLR	4	$q^+_{макс}$, SIC1, MR, ESS	43	0,697	0,26	1,48	0,623	0,29	0,658
17	RF	3	$q^+_{макс}$, $C_a^{макс}$, *Q*	43	0,854	0,18	4,39	0,694	0,26	0,573
18	SVM	2	SPAC, ESS	43	0,735	0,25	2,36	0,570	0,31	0,400

* $q^+_{макс}$ – максимальный положительный парциальный атомный заряд; SIC1 – структурное информационное содержание 1-го порядка; MR – молекулярная рефракция; ESS – сумма квадратов средних значений и дисперсий атомных электроотрицательностей; $C_a^{макс}$ – максимальное значение H-акцепторного свободноэнергетического дескриптора; *Q* – комбинация суммы степеней и суммы квадратов степеней вершин; SPAC – сумма путей между активными центрами.

из обучающей выборки приводит к улучшению статистических характеристик QSAR-моделей. Аналогичный результат получен при исследовании нейротоксического эффекта у ряда органических растворителей. Предложенный метод может быть использован в медицинской химии для поиска новых перспективных биологически активных химических соединений. В этом качестве могут рассматриваться структуры, имеющие максимальную вероятность появления в областях скачков активности и структурных скачков.

СПИСОК ЛИТЕРАТУРЫ

1. *Hansch C., Hoekman D., Leo A., Weininger D., Selassie C.D.* // Chem. Rev. 2002. Vol. 102. P. 783.
2. *Lewis R.A., Wood D.* // WIREs Comput. Mol. Sci. 2014. Vol. 4. P. 505.
3. *Wang T., Wu M.B., Lin J.P., Yang L.R.* // Exp. Opin. Drug. Discov. 2015. Vol. 10. N 12. P. 1283.
4. *McKinney J.D., Richard A., Waller C., Newman M.C., Gerberick F.* // Toxicol. Sci. 2000. Vol. 56. N 1. P. 8.
5. *Rouvray D.H.* The Evolution of the Concept of Molecular Similarity // In: Concepts and Applications of Molecular Similarity / Johnson M.A., Maggiora G.M., Eds. New York: Wiley, 1990. P. 15.
6. *Devinyak O.T., Lesyk R.B.* // Curr. Comput. Aided Drug. Des. 2016. Vol. 12. N 4. P. 265.
7. *Yousefinejad S., Hemmateenejad B.* // Chem. Int. Lab. Sys. 2015. Vol. 149 Part B. P. 177.
8. *Begam B.F., Kumar J.S.* // Ind. J. Sci. Tech. 2016. Vol. 9. N 8. P. 1.
9. *Aggarval C.C.* Outlier Analysis. Springer Int. Publ., 2017. 466 p.
10. *Furusjo E., Svenson A., Rahmberg M., Andersson M.* // Chemosphere. 2006. Vol. 63. N 1. P. 99.
11. *Roy K., Ghosh G.* // Chemosphere. 2009. Vol. 77. N 7. P. 999.
12. *Eriksson L., Jaworska J., Worth A.P., Cronin M.T.D., McDowell R.M., Gramatica P.* // Environ. Health Perspect. 2003. Vol. 111. N 10. P. 1361.
13. *Guha R.* // WIREs Comput. Mol. Sci. 2012. Vol. 2. P. 829.
14. *Maggiora G.M.* // J. Chem. Inf. Model. 2006. Vol. 46. N 4. P. 1535.
15. *Cruz-Monteagudo M., Medina-Franco J.L., Pérez-Castillo Y., Nicolotti O., Cordeiro M.N.D.S., Borges F.* // Drug Disc. Today. 2014. Vol. 119. N 8. P. 1069.
16. *Radchenko E.V., Makhaeva G.F., Palyulin V.A., Zefirov N.S.* Chemical Similarity, Shape Matching and QSAR // In: Computational Systems Pharmacology and Toxicology / Richardson R.J., Johnson D.E., Eds. RCS, 2017. P. 120.
17. *Guha R., Van Drie J.H.* // J. Chem. Inf. Model. 2008. Vol. 48. P. 646.
18. *Pérez-Villanueva J., Santos R., Hernández-Campos A., Giulianotti M.A., Castillo R., Medina-Franco J.L.* // Bioorg. Med. Chem. 2010. Vol. 18. P. 7380.
19. *Cronin M.T.D.* // Toxicol. in Vitro. 1996. Vol. 10. P. 103.
20. URL: http://openbabel.org/wiki/Main_Page.
21. *Willett P.* // J. Chem. Inf. Comput. Sci. 1998. Vol. 38. P. 983.
22. *Раевский О.А., Григорьев В.Ю., Трепалин С.В.* Свидетельство об официальной регистрации программы для ЭВМ HYBOT (Hydrogen Bond

Thermodynamics) № 990090 от 26 февраля 1999 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.
23. *Raevsky O., Sapegin A., Zefirov N.* QSAR Discriminant-Regression Model // In: QSAR: Rational Approaches to the Design of Bioactive Compounds / Silipo C., Vittoria A., Eds., Amsterdam: Elsevier, 1991. 189.
24. *Forsythe G.E., Malcolm M.A., Moler C.B.* Computer Methods for Mathematical Computations. Prentice-Hall, 1977. 259 p.
25. URL: http://www.stat.berkeley.edu/~breiman/RandomForests/reg_examples/RFR.f.
26. URL: https://github.com/jbcolme/fortran-ls-svm.
27. *Mitra I, Saha A, Roy K.* // Mol. Simult. 2010. Vol. 36. N 13. P. 1067.
28. *Kubinyi H.* // Quant. Struct.-Act. Relat. 1994. Vol. 13. N 3. P. 285.
29. *Wassermann A.M., Dimova D., Bajorath J.* // Chem. Biol. Drug. Des. 2011. Vol. 78. P. 224.
30. *Kiralj R., Ferreira M.M.C.* // J. Braz. Chem. Soc. 2009. Vol. 20. N 4. P. 770.
31. *Tropsha A.* // Mol. Inf. 2010. Vol. 29. P. 476.

Поступила в редакцию 10.03.18
После доработки 01.04.18
Принята к публикации 05.09.18

OUTLIER DETECTION FOR QSAR MODELING OF BIOLOGICAL ACTIVITY OF CHEMICALS ON THE BASIS OF «STRUCTURE – ACTIVITY – SIMILARITY» MAPS

L.D. Grigoreva[1]*, V.Y. Grigorev[2], A.V. Yarkov[2]

*([1]Department of Fundamental Physical and Chemical Engineering, Moscow State University; [2]Institute of Physiologically Active Compounds, Russian Academy of Sciences; *e-mail: ldg@)physchem.msu.ru)*

A new method of outlier detection for QSAR was developed on the basis of "structure – activity – similarity" (SAS) maps analysis. It consists in an empirical assessment of the probability of appearance of a chemical compound in any area of the SAS. It is proposed to consider compounds with the maximum probability of appearance in the activity cliffs and a minimal probability in the smooth region as outlier. The proposed method can also be used in the field of medicinal chemistry to search for new biologically active compounds.

Key words: QSAR, SAR, outlier, Structure – Activity – Similarity maps.

Сведения об авторах: *Григорьева Людмила Дмитриевна* – доцент факультета фундаментальной физико-химической инженерии МГУ имени М.В.Ломоносова, канд. физ.-матем. наук (ldg@)physchem.msu.ru); *Григорьев Вениамин Юрьевич* – вед. науч. сотр. отдела компьютерного молекулярного дизайна ИФАВ РАН, докт. хим. наук (beng@ipac.ac.ru); *Ярков Александр Валентинович* – ст. науч. сотр. отдела компьютерного молекулярного дизайна ИФАВ РАН, канд. хим. наук (yarkov@ipac.ac.ru).