Консенсусный подход при выборе химических соединений для QSAR моделирования на основе карт «Структура – Активность – Подобие»

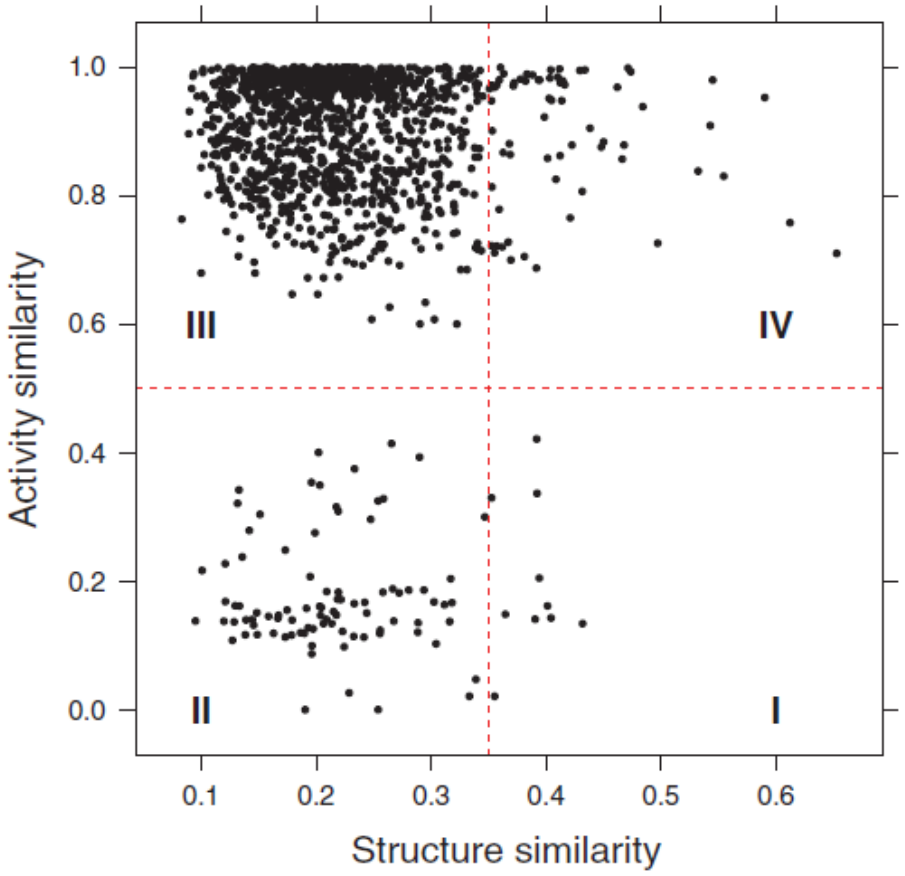
Григорьев В.Ю.

МО, Черноголовка, Институт физиологически активных веществ РАН
E-mail: beng@ipac.ac.ru

Причины появления «выбросов» в QSAR моделях

- Экспериментальные ошибки
- Нарушение статистических гипотез
- Несоблюдение принципа подобия

Пример 2D карты «структура – активность – подобие» (SAS)*



Activity similarity(AS):

$AS_{i,j} = 1 - abs(A_i - A_j)/(A_{max} - A_{min})$

Structure similarity(SS):

$SS_{i,j}$ = Tanimoto coefficient

I – Activity cliffs
II – Nondescript region
III – Structure cliffs
IV – Smooth region

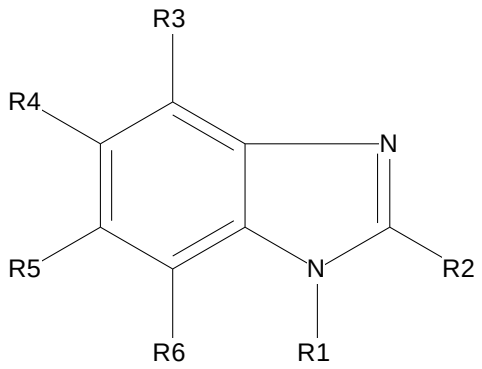
*Guha R. // WIREs Comput. Mol. Sci. 2012. Vol. 2. P. 829-841. (Антималярийная активность, 50 соединений).

Основные факторы, влияющие на вид SAS карт

- Способ описания структуры соединений
- Метрика структурного подобия
- Величины границ подобия

Данные для исследования I*

Биообъект: одноклеточный микроорганизм *Trichomonas vaginalis.*
Мера биоактивности (ингибирующая способность): pIC_{50} (IC_{50}, нМ), pIC_{50} (min) = 4.53; pIC_{50} (max) = 8.70.
Химические соединения: бензимидазолы (n=32).



R_1: H; CH_3
R_2: CF_3; C_2F_5; $CONH_2$; $CONHCH_3$; $CON(CH_3)_2$; $COOCH_2CH_3$; $COCH_3$
R_3: H; Br
R_4: H; CF_3; $SCH_2CH_2CH_3$; COC_6H_5; Br; Cl; NO_2
R_5: H; CF_3; $SCH_2CH_2CH_3$; COC_6H_5; Br; Cl; NO_2
R_6: H; Br

*Perez-Villanueva J. et al. // Bioorg. Med. Chem. 2010. Vol. 18. P. 7380-7391.

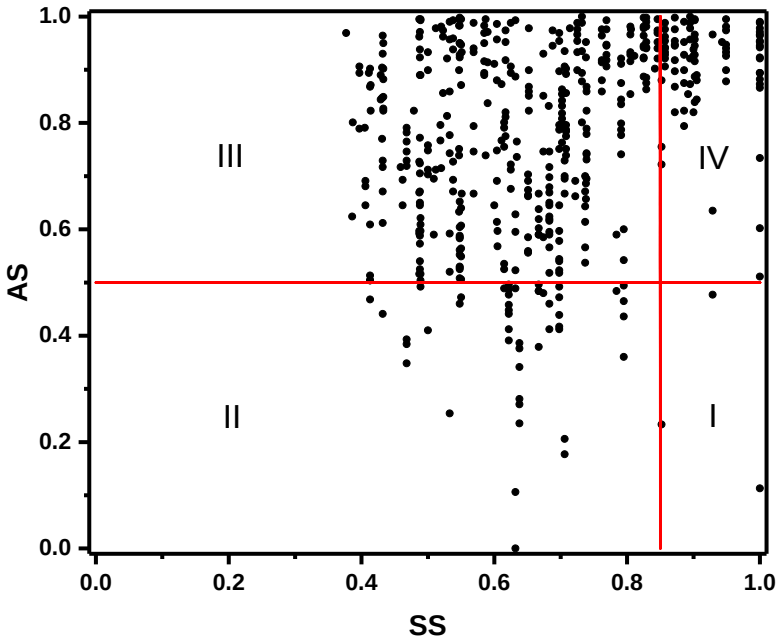
Данные для исследования II

SAS:
Молекулярные ключи: MACCS (166 бит); ECFP4 (1024 бит).
Мера структурного подобия: коэффициент Танимото.
Мера подобия по активности: величина на основе нормированных значений.
Граничные значения подобия (MACCS): 0.68 ÷ 0.85.
Граничные значения подобия (ECFP4): 0.40 ÷ 0.55.
Граничные значения подобия (активность): 0.50 ÷ 0.77.

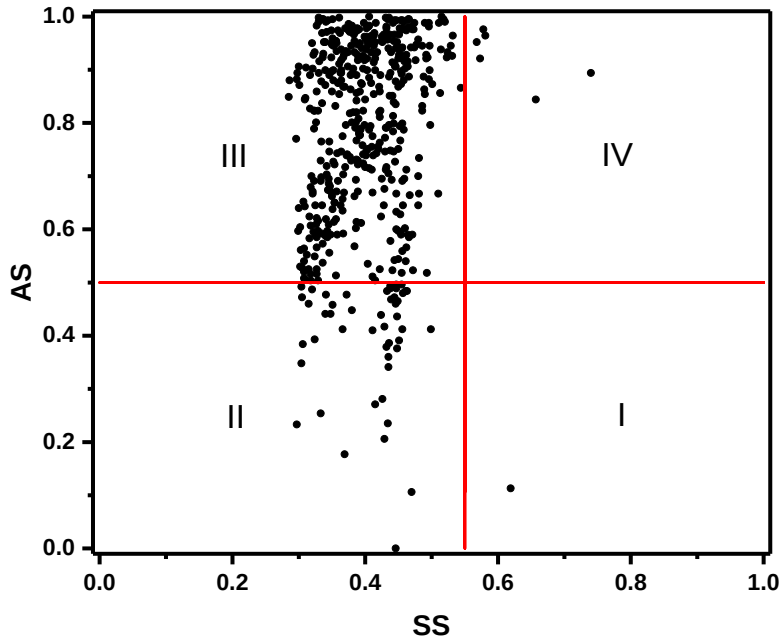
QSAR:
Дескрипторы (m=18): MW; α; $q^+(max)$; $q^-(max)$; Σq^-; $\Sigma q^+/\alpha$; $E_a(max)$; $C_a(max)$; $E_a(max)*E_d(max)$; $C_a(max)*C_d(max)$; ΣE_a; ΣE_{ad}; ΣC_a; ΣC_d; ΣC_{ad}; $\Sigma E_a/\alpha$; $\Sigma E_{ad}/\alpha$; $\Sigma C_a/\alpha$.
Модели: MLR; RF.
Стратегия моделирования: пошаговое включение дескрипторов.
Характеристика качества моделей: фитнес-функция Кубиньи (Kubinyi) FIT.
Критерий отбора моделей: max величина FIT в обучающей выборке.

SAS карты соединений

MACCS/0.85(SS)/0.50(AS)



ECFP4/0.55(SS)/0.50(AS)



Таблицы заселенности SAS карт

MACCS/0.85(SS)/0.50(AS)

№ соед.	P^{I}	P^{II}	P^{III}	P^{IV}
1	0.032	0.032	0.839	0.097
2	0.000	0.129	0.774	0.097
3	0.000	0.129	0.774	0.097
4	0.000	0.032	0.935	0.032
5	0.000	0.065	0.903	0.032
6	0.000	0.677	0.226	0.097
7	0.000	0.226	0.677	0.097
8	0.000	0.032	0.839	0.129
9	0.097	0.258	0.613	0.032
31	0.000	0.032	0.774	0.194
32	0.000	0.097	0.710	0.194

ECFP4/0.55(SS)/0.50(AS)

№ соед.	P^{I}	P^{II}	P^{III}	P^{IV}
29	0.000	0.032	0.903	0.065
30	0.000	0.032	0.903	0.065
32	0.000	0.097	0.871	0.032
16	0.000	0.226	0.742	0.032
21	0.000	0.032	0.935	0.032
17	0.000	0.032	0.935	0.032
11	0.000	0.032	0.968	0.000
12	0.032	0.161	0.806	0.000
10	0.000	0.032	0.968	0.000
20	0.000	0.161	0.839	0.000
9	0.032	0.323	0.645	0.000

$P_i^k = n_i^k/31$; i = 1,…32; k = 1,…4

P – заселенность

n – число молекулярных пар

Анализ заселенности SAS карт

Схема формирования 16-битовых векторов

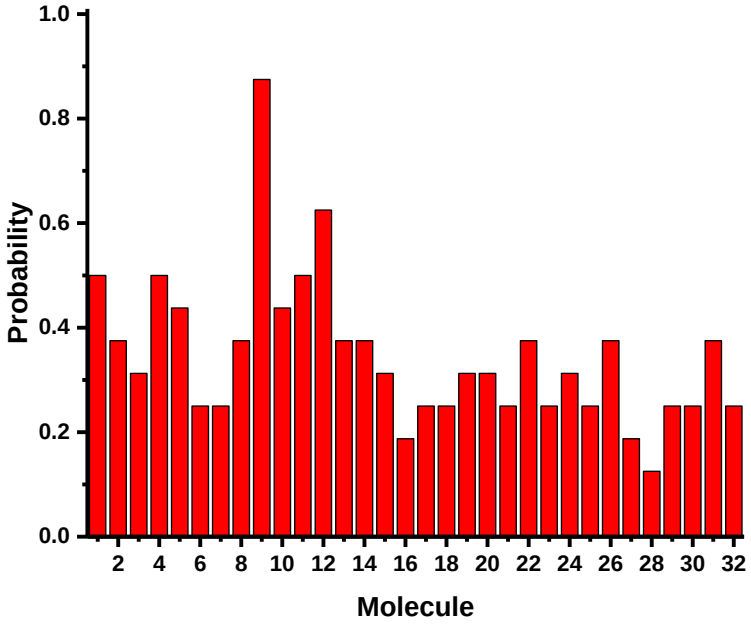
Уровень	Тип															
Fingerprint	MACCS								ECFP4							
Structure similarity	SS_{min}				SS_{max}				SS_{min}				SS_{max}			
Activity similarity	AS_{min}		AS_{max}		AS_{min}		AS_{max}		AS_{min}		AS_{max}		AS_{min}		AS_{max}	
SAS region	I	IV	I	IV	I	IV	I	IV	I	IV	I	IV	I	IV	I	IV

16-битовые вектора соединений и вероятность «выброса»*

№ соед	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ	Prob
9	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	14	0.88
12	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	10	0.63
11	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	1	8	0.50
4	0	0	1	1	0	0	1	1	1	0	1	0	0	1	0	1	8	0.50

*Граничное значение заселенности для формирования единичного бита: P>0 (SAS_I); P=0 (SAS_IV); Prob = Σ/16

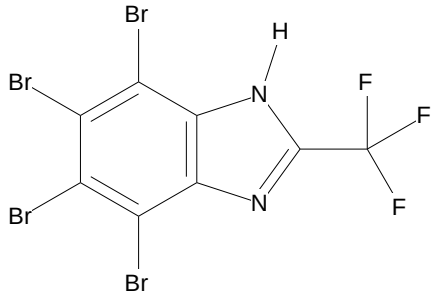
Вероятность проявления соединений в качестве «выбросов» при QSAR моделировании



Соединение 9:

pIC_{50} = 8.70(max);

$\Sigma C_a/\alpha$ = 0.086(min)



Результаты QSAR моделирования

NN	Модель	m	Дескрипторы	n	R^2	s	R^2_{cv}	s_{cv}	R^2_{rand}
1	MLR	3	$q^-(max)$; Σq^-; ΣC_a;	32	0.629	0.53	0.464	0.64	0.585
2	RF	2	α; $\Sigma q^+/\alpha$;	32	0.875	0.31	0.668	0.50	0.610
3	MLR	3	$q^-(max)$; Σq^-; ΣC_a;	31	0.707	0.43	0.585	0.51	0.670
4	RF	2	α; $\Sigma q^+/\alpha$;	31	0.873	0.31	0.646	0.52	0.588
5	MLR	2	$q^-(max)$; Σq^-;	31	0.692	0.44	0.614	0.49	0.668
6	RF	2	α; ΣC_a;	31	0.888	0.26	0.639	0.47	0.571

Выводы

1. Для оценки распределения соединений в различных областях карт «структура – активность –подобие» предложен простой способ расчета заселенности.
2. На основе анализа заселенности разработан консенсусный подход для выявления потенциальных «выбросов» в регрессионных моделях.
3. Установлено, что удаление потенциальных «выбросов» улучшает статистические характеристики QSAR моделей.