УДК 544.165

П. Г. Полищук [1], В. Е. Кузьмин [2], А. Г. Артеменко [1],
С. Г. Соболева [2], С. Ю. Макан [1]
[1] Физико-химический институт им. А. В. Богатского НАН Украины,
отдел медицинской химии, отдел молекулярной структуры
Украина, Одесса, 65080, Люстдорфская дорога, 86,
тел. 8-0482-66-30-41, e-mail: pavel_polishchuk@ukr.net

[2] Одесский национальный университет им. И. И. Мечникова,
химический факультет
Украина, Одесса, Дворянская, 2

QSAR АНАЛИЗ ЛИГАНДОВ 5-HT_{1A} РЕЦЕПТОРОВ МЕТОДОМ КЛАССИФИКАЦИОННЫХ ДЕРЕВЬЕВ

Проведен анализ связи структура-аффинитет к 5-HT_{1A} рецепторам методом классификационных деревьев для 38 соединений обучающей выборки. Разработана процедура, позволяющая в рамках этого алгоритма оценить относительное влияние структурных фрагментов на изучаемую активность. Полученная QSAR [1] модель позволяет достаточно надежно прогнозировать классы активности. Определены структурные фрагменты и характер их влияния на аффинитет.
Ключевые слова: QSAR, лиганды 5-HT_{1A} рецепторов, классификациионные деревья.

Исследования, посвященные изучению 5-HT_{1A} рецепторов и поиску новых перспективных лигандов, ведутся довольно продолжительное время [1]. Однако и по сей день, они остаются актуальными и важными задачами медицинской химии. Интерес к рецепторам 5-HT_{1A} обусловлен их вовлеченностью в регуляцию состояний тревоги и страха. В последние годы синтезировано много лигандов 5-HT_{1A} рецепторов, с целью дальнейшей разработки лекарственных препаратов для лечения различных психические расстройств. Для эффективного поиска наиболее перспективных соединений используются методы компьютерного моделирования и QSAR, по результатам которых можно проводить направленный синтез новых лекарственных препаратов.

Общие понятия о классификационных деревьях

В качестве альтернативы широко распространенным регрессионным методам для построения QSAR моделей мы предлагаем использовать метод классификационных деревьев [2]. Классификационные деревья применяются на сегодняшний день достаточно ши-

[1] QSAR — Quantitative structure-activity relationship.

роко и успешно в таких областях как финансы, бизнес, медицина и биология.

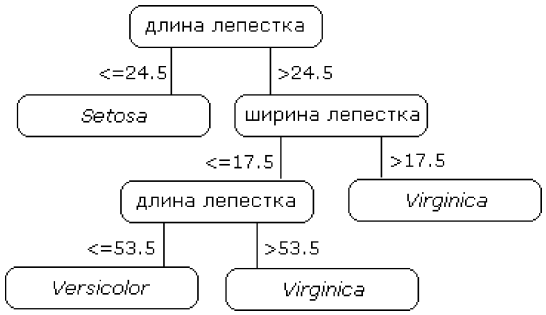


Рис 1. Пример модели классификационных деревьев для ирисов Фишера

На рис. 1 приведен пример модели дерева решений для классификации ирисов Фишера. Классификационное дерево состоит, как и обычное дерево, из:

— узлов (мест разветвления, называемых также вершинами), обозначенных прямоугольниками,

— ветвей, обозначенных отрезками, соединяющих узлы,

— корня — первой вершины дерева,

— листьев — терминальных вершин, которыми заканчиваются цепочки "корень-ветвь-вершина-...-вершина".

Построение классификационного дерева происходит следующим образом. На вход (в корень) подается некоторое обучающее множество, содержащее объекты, характеризуемые атрибутами, один из которых определяет принадлежность каждого объекта к определенному классу. Далее алгоритм вырабатывает общие критерии для объектов одного класса. Для этого в каждой вершине, начиная с корня, определяется правило, на основании которого и происходит дальнейшее разделение обучающего множества. Под правилом понимается логическая конструкция, представленная в виде "если... то...". Такой рост дерева происходит до тех пор, пока в вершине не останутся объекты, принадлежащие к одному классу, т. е. в результате мы получаем терминальную вершину. Иначе говоря, классификационные деревья — это способ представления правил классификации в иерархической, последовательной структуре, где каждому объекту соответствует единственная вершина, дающая решение.

В настоящий момент использование классификационных деревьев в QSAR задачах не распространено. Метод классификационных деревьев имеет ряд преимуществ перед широко распространенными регрессионными методами QSAR анализа, такие как быстрый процесс обучения, интуитивно понятная классификационная модель на естественном языке, нелинейность получаемых моделей и возможность построения моделей для случаев, когда используется не чис-

ловая шкала, а классификационная. Однако у моделей классификационных деревьев имеется и ряд недостатков в применении к QSAR задачам, главный из которых трудность определения структурных фрагментов, оказывающих наибольшее влияние на изучаемую активность.

Цели и задачи исследования

Целью настоящей работы явилось решение задачи QSAR для 38 лигандов 5-HT$_{1A}$ рецепторов с помощью метода классификационных деревьев. Для оценки эффективности используемого метода мы сочли целесообразным провести сравнительный анализ его результатов с данными предыдущих аналогичных QSAR исследований [3]. Кроме того, была предпринята попытка разработать и применить для интерпретации полученной QSAR модели подход, позволяющий оценить относительный вклад в активность структурных параметров. На этой основе можно выявить молекулярные фрагменты, которые определяют изучаемую активность. Описываемый ниже подход интерпретации классификационных деревьев применим только к классификационным (ранговым) QSAR моделям, однако это не снижает его значимость, так как большинство исследователей интересуют в первую очередь качественные результаты (ранги или относительные величины активности) QSAR оценок. Указанный подход достаточно универсален, так как применим к различным моделям деревьев независимо от использованного алгоритма при их создании.

Процедура оценки относительного влияния структурных факторов на исследуемую активность

Рассмотрим общий случай. Пусть выборка содержит некоторое число соединений, которое можно разделить на N групп или рангов, с точки зрения величины их активности. При этом активность соединений в каждом последующей группе должна быть больше, чем в предыдущей и отличаться на постоянное число, например 1, т. е. $A_1 < A_2 < A_3 < ... < A_N$ и $A_n = A_{n-1} + 1$, где A — ранг активности, $1 < n < N$. Каждое соединение обладает набором дескрипторов (структурных параметров) S_{ij}, где i — номер соединения в выборке, j — номер дескриптора. На основании этих данных строится модель по методу деревьев классификаций. По полученной модели рассчитывают вклады дескрипторов, анализ которых позволяет сделать вывод об их относительном количественном влиянии на целевую активность.

Расчет вкладов каждого дескриптора производится в каждой вершине дерева только для тех дескрипторов, которые добавились в правило при построении данной вершины. Каждому вкладу дескриптора ставится в соответствие диапазон значений этого дескриптора, в котором данное правило выполняется. Расчет вкладов проводят по формуле (1) аналогичной той, которая используется для тренд-вектора [4, 5]:

$$T_j = \frac{1}{n}\sum_{i=1}^{n}\left[\left(A_i - A_{mean}\right)S_{ij}\right], \qquad (1)$$

где T_j — относительный вклад j-го дескриптора в исследуемую активность, n — число соединений в данной вершине, A_i — ранг активности i-го соединения, A_{mean} — среднее значение ранга активности по всей выборке, S_{ij} — значение j-го дескриптора i-го соединения.

В результате мы получаем вклад дескриптора и диапазон его значений, внутри которого этот вклад справедлив. Таким образом, если значение дескриптора соединения лежит вне этого диапазона, то его вклад в данном случае считается равным нулю, т. е. он не оказывает влияния на активность.

Симплексное представление молекулярной структуры

В качестве структурных параметров в данной работе использовались симплексы — четырёхатомные молекулярные фрагменты фиксированной структуры, хиральности и симметрии. На рис. 2 представлен пример деления молекулы на симплексы.

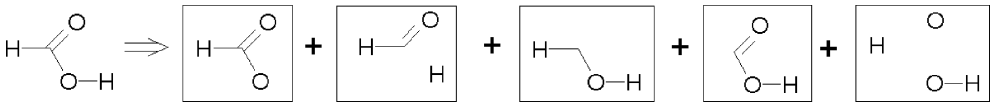


Рис. 2. Представление молекулярной структуры в виде системы симплексов на примере молекулы муравьиной кислоты

Дескриптором в этом случае служит число симплексов конкретного типа [6]. Вершины в симплексах можно дифференцировать по: типу атомов, по липофильности, частичному заряду на атоме, рефракции и способности образовывать водородные связи. Для этого проводят расчет конкретной характеристики для каждого атома и затем делят все атомы на конечное число групп (как правило 3–7) [*]. Мы использовали следующие параметры разделения атомов на группы (римские цифры обозначают тип вершины симплекса):

- для частичных зарядов — $I \le -0.1 < II \le -0.05 < III \le -0.01 < IV \le 0.01 < V \le 0.05 < VI \le 0.1 < VII$;
- для липофильности — $I \le -1.0 < II \le -0.5 < III \le -0.1 < IV \le 0.1 < V \le 0.5 < VI \le 1.0 < VII$;
- для рефракции — $I \le 2 < II \le 3 < III \le 4 < IV \le 6 < V \le 9 < VI \le 12 < VII$;
- донорно-акцепторного взаимодействия I — донор, II — акцептор, III — индифферентный центр.

[*] Количество групп и положение границ между ними являются настроечными параметрами модели.

После определения вкладов структурных параметров (симплексов) проводится их пересчет во вклады атомов, и в результате получаем возможность отразить на структуре полученные данные. Это позволяет выделить фрагменты, влияющие на изучаемую активность (см. ниже).

Решение задачи QSAR для лигандов 5-HT$_{1A}$ рецепторов

Метод классификационных деревьев использован для анализа выборки из 38 соединений (рис. 3), которые являются лигандами 5-HT$_{1A}$ рецепторов. Ранее эта задача решалась другими методами [3], при сравнении с которыми можно проверить эффективность предлагаемого подхода.

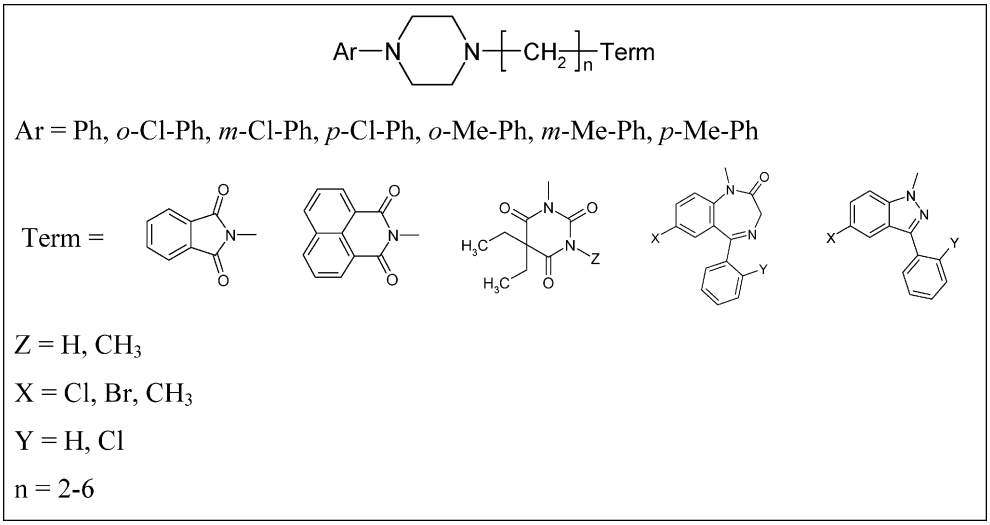


Рис. 3. Структуры соединений обучающей выборки

Поскольку предложенный метод интерпретации применим только к классификационным моделям, на первом этапе было произведено разделение всех соединений выборки на три класса (ранга) активности в зависимости от значения показателя константы ингибирования (pK$_i$), как видно из в таблицы 1.

Таблица 1

Разделения на классы (ранги) активности соединений обучающей выборки

Ранг активности	Диапазон значений pK$_i$	Число соединений данного ранга
1 (A)	< 5.97	13
2 (B)	5.97-7.03	11
3 (C)	> 7.03	14

На втором этапе было рассчитано более 1900 симплексных дескрипторов, количество которых было сокращено до 238, путем отсева взаимнокоррелирующих параметров.

Следующим шагом было построение модели классификационных деревьев. Для этого использовалось программное обеспечение SPSS AnswerTree [7] и алгоритм CART [2] [2], который показал наилучшие результаты при решении поставленной задачи.

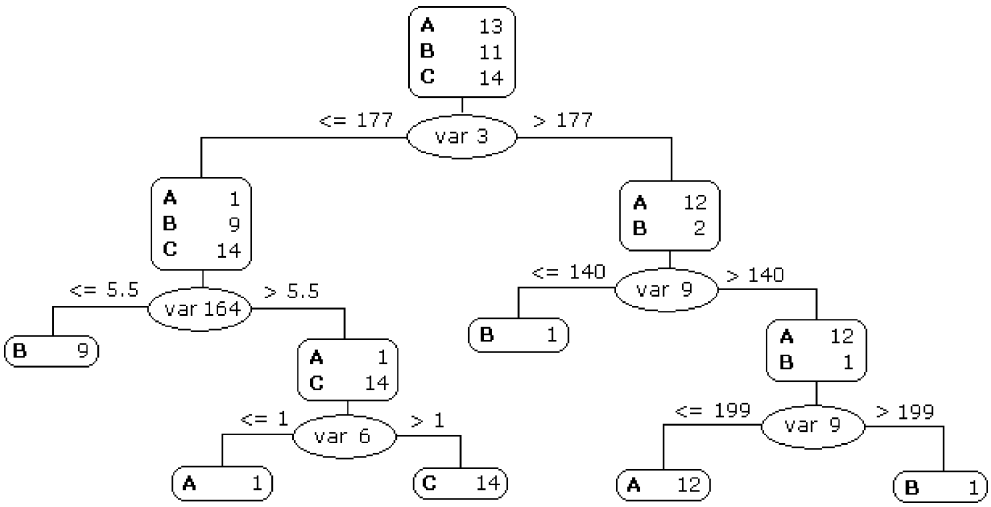


Рис. 4. Модель классификационных деревьев для задачи поиска связи "структура-аффинитет" для лигандов 5-HT_{1A} рецепторов

Полученное дерево (рис. 4) не содержало ошибок классификации (в каждой из шести терминальных вершин находятся соединения только одного класса активности). Проверку надежности этой модели проводили методом скользящего контроля. Для этого из обучающей выборки исключали десятую часть соединений, и для оставшихся строили новую модель, по которой предсказывали ранг активности исключенных соединений. Подобную процедуру повторяли десять раз, каждый раз, исключая другие соединения. Обобщая результаты всех прогнозов, в итоге получаем ошибку классификации в условиях скользящего контроля, которая составила для вышеприведенной модели всего 15%, что можно считать удовлетворительным результатом.

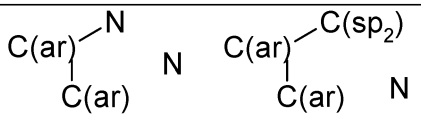
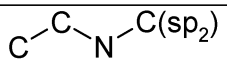
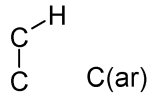
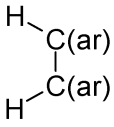
В вершинах дерева (прямоугольниках) в первой колонке содержатся обозначения классов активности, числа второй колонки представляют собой количество соединений, удовлетворяющих правилу данной вершины. В овалах приведены обозначения дескрипторов,

[2] CART — Classification and Regression Trees.

по которым происходит деление соединений соответствующих каждой вершине. Симплексы, использованные в QSAR модели, и их относительные вклады в аффинитет приведены в таблице 2. Поскольку атомы одного элемента могут отличаться друг от друга и иметь разные свойства, то для них в таблице введены такие обозначения: C(sp$_2$) — sp^2-гибридизованный атом углерода, C(ar) — атом углерода, входящий в ароматическую систему. Следует помнить, что вклад каждого симплекса в активность зависит от его количества в соединении. В последней колонке таблицы 2 приведены вклады симплексных дескрипторов, с указанием диапазонов внутри которых эти вклады справедливы.

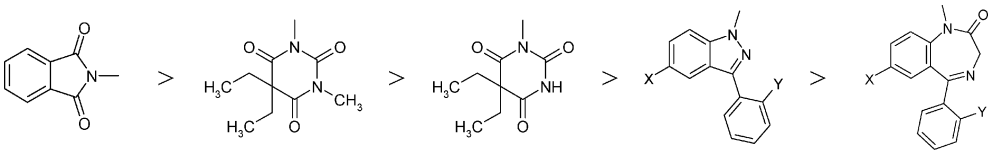
Таблица 2

Типы симплексов и их относительный вклад в аффинитет молекул к 5-HT$_{1A}$ рецепторам

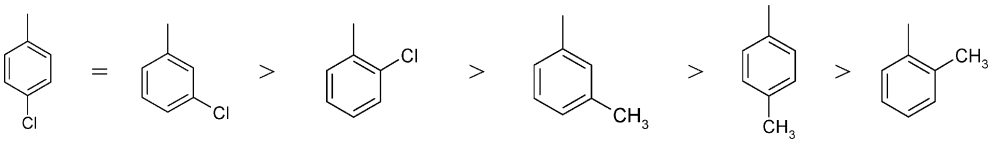
Обозначение дескриптора	Симплексы	Тип дифференциации вершин симплекса	Вклад (диапазон значений дескриптора)
var 3		по рефракции	72 [0; 177] -169 (177; +∞)
var 6		по заряду	0 [0; 1] 2 (1; +∞)
var 9		по липофильности	-88 [0; 140] -438 (140; 190] -280 (190; +∞)
var 164		по типу атома	0 [0; 5.5] 6 (5.5; +∞)

В каждой вершине дерева по описанной выше процедуре проводился расчет вкладов для симплексов, которые использовались при построении правила данной вершины. После пересчета вкладов симплексов во вклады атомов, появилась возможность отразить их значения на молекулярной структуре и выявить, таким образом, фрагменты, определяющие аффинитет исследуемых соединений к 5-HT$_{1A}$ рецепторам. При анализе для всех соединений выборки, удалось выделить типичные фрагменты, влияющие на аффинитет к 5-HT$_{1A}$ рецепторам.

Так для терминальной части молекул лигандов 5-HT$_{1A}$ рецепторов удалось выделить следующий ряд фрагментов в порядке уменьшения положительного эффекта на аффинитет:



Для арильного фрагмента молекул лигандов подобный ряд будет выглядеть так:

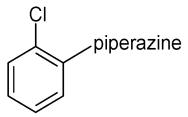
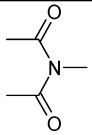
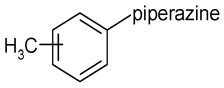
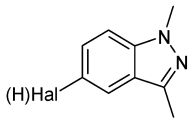
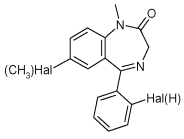


Однозначно оценить влияние длины полиметиленовой цепочки оказалось затруднительно.

Из приведенных данных видно, что наиболее неблагоприятными для аффинитета являются остатки бенздиазепинов, изоксазолов и толила. К фрагментам в наибольшей степени способствующим аффинитету можно отнести остатки фталимида, барбитуровых кислот и хлорфенила.

Таблица 3

Фрагменты, влияющие на аффинитет изучаемых соединений, определенные по результатам работы [3]

Фрагменты, способствующие проявлению аффинитета		
	-$(CH_2)_4$-, -$(CH_2)_5$-	
Фрагменты, препятствующие проявлению аффинитета		
		

Если сопоставить эти результаты с полученными ранее в работе [3] (табл. 3) с применением метода частичных наименьших квадратов, то можно увидеть их хорошее соответствие между собой. Это свидетельствует о том, что предлагаемый подход вполне адекватен, и его можно применять в дальнейшем при решении подобных QSAR задач при поиске зависимости структура-активность. В будущем планируется использовать данный подход для анализа выборки, содержащей более 200 соединений, данные по которым получе-

ны различными авторами и при анализе которых метод деревьев решений с его возможностью построения классификационных моделей является очень перспективным.

Авторы выражают искреннюю признательность академику НАН Украины С. А. Андронати за полезные советы и плодотворное обсуждение.

Литература

1. *Андронати С. А., Макан С. Ю.* Азотсодержащие гетероциклические соединения — лиганды серотониновых рецепторов // Азотистые гетероциклы и алкалоиды, под редакцией д. х. н. В. Г. Карцева и акад. Г. А. Толстикова. — М., 2001. — Т. 1. — С. 20–30. — Р. 368.
2. *Classification* and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. T. Stone. — California, 1984.
3. *Hierarchic* system of QSAR models (1D-4D) on the base of simplex representation of molecular structure / V. E. Kuz'min, A. G. Artemenko, P. G. Polischuk, E. N. Muratov, A. I. Hromov, A. V. Liahovskiy, S. A. Andronati, S. Yu. Makan // J. Mol. Model. — 2005. — Vol. 11. — P. 457–467.
4. *Carhart R. E., Smith D. H., Venkataraghavan R.* Atom pairs as molecular features in structure — activity studies. Definition and application // J. Chem. Inf. Comput. — 1985. — Vol. 25. — N 2. — P. 64.
5. *Витюк Н. В., Кузьмин В. Е.* Механистические модели в хельмометрике для анализа многомерных исследовательских данных. Аналог метода дипольных моментов в анализе связи структура (состав) — свойство // Журнал аналитической химии. — 1994. — Т. 49, № 2. — С. 165–167.
6. *The analysis* of structure-anticancer and antiviral activity relationship for macrocyclic pyridinophanes and their analogues on the basis of 4D QSAR models (simplex representation of molecular structure) / V. E. Kuz'min, A. G. Artemenko, V. P. Lozitsky, E. N. Muratov, A. S. Fedtchouk, N. S. Dyachenko, L. N. Nosach, T. L. Gridina, L. I. Shitikova, L. M. Mudrik, A. K. Mescheriakov, V. A. Chelombitko, A. I. Zheltvay, J.-J. Vanden Eynde // Acta Biochimika Polonica. — 2002. — Vol. 49. — P. 157–168.
7. *Trial* версия программы с http://www.spss.com

П. Г. Поліщук [1], В. Є. Кузьмін [2], А. Г. Артеменко [1], С. Г. Соболєва [2], С. Ю. Макан [1]

[1] Фізико-хімічний інститут ім. О. В. Богатского НАН України

[2] Одеський національний університет ім. І. І. Мечникова

QSAR АНАЛІЗ ЛІГАНДІВ 5-HT1A РЕЦЕПТОРІВ МЕТОДОМ КЛАСИФІКАЦІЙНИХ ДЕРЕВ

Резюме

Авторами запропоновано підхід інтерпретації моделей класифікаційних дерев, що є цілком адекватним для вирішення QSAR задач. Він дозволяє достатньо надійно прогнозувати рівні активності досліджуваних сполук, а також дозволяє виділити структурні фрагменти, що визначають аффінітет сполук до 5-HT$_{1A}$ рецепторів.

Ключові слова: QSAR, ліганди 5-HT$_{1A}$ рецепторів, класифікаційні дерева.

Pavel G. Polischuk [1], Victor E. Kuz'min [2], Anatoly G. Artemenko [1], Svetlana G. Soboleva [2], Svetlana Yu. Makan [1]

[1] O. V. Bogatsky Physico-Chemical Institute of the National Academy of Sciences of Ukraine

[2] I. I. Mechnikov Odessa National University

QSAR ANALYSIS OF LIGANDS OF SEROTONIN RECEPTORS BY CLASSIFICATION TREES METHOD

Summary

Thus we ascertained that the proposed approach was quite adequate for solution of QSAR tasks. This approach could be used for reliable prediction of activity ranks. Also it allowed choosing structure fragments which determine studied affinity.

Keywords: QSAR, ligands of 5-HT1A receptors, classification trees.