УДК 541.69

Локальные молекулярные характеристики в анализе количественной связи «структура—активность»

Е. В. Радченко, В. А. Палюлин, Н. С.Зефиров

ЕВГЕНИЙ ВАЛЕРЬЕВИЧ РАДЧЕНКО — кандидат химических наук, научный сотрудник Химического факультета МГУ им. М.В. Ломоносова. Область научных интересов: исследование соотношений «структура—свойство», «структура—активность» (QSPR/QSAR) органических соединений, молекулярное моделирование, дизайн биологически активных соединений. E-mail genie@qsar.chem.msu.su

ВЛАДИМИР АЛЕКСАНДРОВИЧ ПАЛЮЛИН — кандидат химических наук, ведущий научный сотрудник Химического факультета МГУ им. М.В. Ломоносова. Область научных интересов: органическая химия, исследование соотношений «структура—свойство», «структура—активность» (QSPR/QSAR) органических соединений, молекулярное моделирование. E-mail vap@org.chem.msu.su

НИКОЛАЙ СЕРАФИМОВИЧ ЗЕФИРОВ — академик РАН, профессор, директор Института физиологически активных веществ РАН, заведующий кафедрой органической химии Химического факультета МГУ им. М.В. Ломоносова. Область научных интересов: органическая и медицинская химия, исследование соотношений «структура—свойство», «структура—активность» (QSPR/QSAR) органических соединений.
E-mail zefirov@org.chem.msu.ru

119992 Москва, Ленинские горы, МГУ, Химический факультет, тел. (495)939-35-57, факс (495)939-02-90.

Введение

Среди видов активности, рассматриваемых при поиске эффективных лекарств и других биологически активных соединений, особый интерес представляет специфическая, «рецепторная» активность, определяемая взаимодействием молекулы-лиганда с биологической мишенью. Очевидно, что такая активность должна быть непосредственно связана с локальными молекулярными характеристиками структуры, в частности, со свойствами ее атомов и связей. Следовательно, анализ количественной связи «структура—активность» может быть осуществлен путем корректного сопоставления активности и локальных молекулярных характеристик, причем как в пределах структуры, так и между разными структурами родственных соединений, с последующим построением прогностических моделей, связывающих показатели биологической активности с такими параметрами и позволяющих осуществлять конструирование новых перспективных структур.

Существующие подходы к анализу связи «структура—активность» на основе локальных молекулярных характеристик можно разделить на две большие группы в зависимости от используемого представления структур. Методы первой группы опираются на трехмерную модель структуры, т.е. на данные о пространственном расположении атомов. Другая группа — топологические методы — исходит из структурной формулы вещества, дающей информацию о типах атомов и связей между ними, причем как для атомов, так и для связей дополнительно могут задаваться локальные стереохимические и физико-химические характеристики. Казалось бы, подходы на основе трехмерной модели структуры являются более точными, т.е. лучше согласованными с представлениями о строении вещества. Тем не менее накопленный опыт практического использования методов обеих групп показывает, что рассмотрение трехмерной структуры не всегда целесообразно. Преимущество топологических подходов отчасти связано с тем, что они созвучны привычным представлениям специалистов в области органической химии и облегчают последующее конструирование и планирование синтеза новых перспективных структур. Кроме того, трехмерные методы сопряжены с большими объемами данных о всех деталях строения, конформационного поведения и физико-химических характеристиках исследуемых соединений, которые зачастую начинают играть роль «информационного шума», затрудняя выявление истинных взаимосвязей между их структурой и активностью.

Анализ связи трехмерной структуры с активностью

Наиболее широко применяемым и, можно сказать, классическим подходом к анализу связи трехмерной структуры с активностью (3D QSAR) является опубликованный Р. Крамером с сотр. в 1988 г. [1] метод сравнительного анализа молекулярного поля CoMFA (Comparative Molecular Field Analysis), а также различные усовершенствованные и модифицированные его варианты. Потенциально этот метод позволяет выявить области пространства вокруг молекулы, где ее определенные локальные свойства оказывают положительное или отрицательное влияние на биоактивность.

При разработке метода авторы исходили из того, что взаимодействие органических соединений (лигандов) с биологическими мишенями обычно является нековалентным и существенно зависит от формы молекул, а учет ван-дер-ваальсовых и кулоновских сил, как правило, позволяет адекватно описать нековалентные взаимодействия в рамках методов молекулярной механики. Было высказано предположение, что для понимания наблюдаемого биологического действия лигандов достаточно информации о форме и электростатическом поле их молекул. Ключевым моментом предложенного подхода является сопоставление количественных показателей этих свойств, проводи-

мое после процедуры пространственного совмещения структур родственных соединений. При этом составляется таблица (матрица) рассчитанных для каждого из узлов трехмерной решетки значений энергии ван-дер-ваальсова (стерического) и кулоновского (электростатического) взаимодействия стандартной пробной частицы с молекулами. В зависимости от особенностей задачи в качестве пробной частицы может выбираться, например, протон, sp^3-гибридизованный атом углерода с единичным положительным зарядом и др.

Для построения QSAR-модели по таким дескрипторным матрицам высокой размерности используется метод регрессии частичных наименьших квадратов (PLS — partial least squares) со скользящим (перекрестным) контролем [2]. В отличие от обычной множественной линейной регрессии он позволяет надежно выявлять статистические закономерности, даже когда число независимых переменных многократно превышает число экспериментальных объектов (как в данном случае).

В первом сообщении о методе CoMFA [1] в качестве примера его применения приводилось исследование данных по связыванию серии из 21 стероида с кортикостероид-связывающим глобулином (CBG). Из-за относительной жесткости стероидного каркаса каждое соединение было представлено единственной молекулярной моделью. Была получена модель, содержащая два PLS-фактора. Модель характеризуется достаточно высокими значениями коэффициента корреляции (r = 0,947) и параметра перекрестного контроля (q^2 = 0,662), однако качество предсказания для контрольной выборки из 10 соединений оказалось существенно более низким. Несмотря на такие неоднозначные результаты и ряд погрешностей в исходных данных [3], предложенный подход довольно быстро нашел широкое применение в исследованиях количественной связи «структура—активность». Это связано прежде всего с привлекательностью определяемых по PLS-модели трехмерных карт, которые отражают области благоприятных и неблагоприятных для активности взаимодействий. Немаловажно также то обстоятельство, что, получив патент на метод CoMFA [4], фирма «Tripos» реализовала его в одном из наиболее популярных программных комплексов для молекулярного моделирования Sybyl.

К настоящему времени опубликованы сотни работ по применению метода CoMFA [5] для прогнозирования различных видов биологической активности органических соединений, например, способности органических лигандов к связыванию с дофаминовым рецептором [6], ингибирования протеазы вируса ВИЧ-1 циклическими сульфамидами [7] или антикокцидной активности триазиндионов [8], и даже физико-химических свойств, например, констант диссоциации замещенных имидазолидинов [9] и электронных констант заместителей [10]. При использовании в качестве зависимой переменной логарифма отношения показателей двух видов активности удается получить четкую картину факторов, определяющих селективность действия соединений (так называемые поля селективности) [11]. В ходе развития метода CoMFA было предложено в дополнение к стерическим и электростатическим дескрипторам использовать определяемые в узлах трехмерной решетки значения других параметров, в частности, липофильного потенциала [12, 13] и квантовохимических характеристик, например плотностей орбиталей [14]. Для улучшения устойчивости, предсказательной способности и удобства интерпретации получаемых моделей целесообразно применять процедуры отбора дескрипторов [15—17].

Вместе с тем применение метода CoMFA нередко затрудняется из-за необходимости использования большого числа дескрипторов, а также из-за проблемы совмещения (alignment) трехмерных молекулярных структур [18], особенно в случае конформационно подвижных молекул, когда индуцированное соответствие между лигандом и биомишенью может приводить к конформациям лиганда, значительно отличающимся от оптимальных конформаций для изолированной молекулы. В ряде случаев более высокое качество описания активности и ее прогноза для новых соединений обеспечивают топологические модели (например, основанные на фрагментных дескрипторах) [19].

Наряду с описанными в работе [1] методами совмещения жестких моделей молекул разработаны и другие принципы совмещения структур [20], применяются также альтернативные методы статистического анализа данных [21]. Кроме того, предложен ряд подходов, позволяющих уменьшить чувствительность моделей к совмещению структур или вообще устранить этот этап анализа. Например, в качестве характеристики стерических полей можно использовать объем пересечения ван-дер-ваальсовых объемов пробной частицы и молекулы лиганда, обладающий более гладкой зависимостью от расстояния по сравнению со стандартной функцией Леннард-Джонса [22]. В случае серии родственных соединений для получения полезных результатов часто необходимо и достаточно применение канонических (иначе говоря, формальных) правил совмещения [8, 18, 23]. Дальнейшее развитие этой идеи привело авторов к концепции метода «топомерного» CoMFA [24, 25], который по сути основан на трехмерном анализе структур после топологического их совмещения. Наконец, при рассмотрении векторов пространственной автокорреляции [3] и моментов распределений массы и заряда [26] получаются трехмерные молекулярные дескрипторы, инвариантные к трансляции и вращению молекул.

Таким образом, при достаточно больших возможностях метода CoMFA его применение нередко требует отхода от прямого анализа трехмерной молекулярной модели и частичного возврата к топологическому представлению.

Топологические подходы

Многие методы анализа связи между структурой и биологической активностью так или иначе используют для построения модели локальные молекулярные характеристики (особенно для серий соединений достаточно близкой структуры). Например, это относится к методу Ганча (Hansch) [27—29] (константы заместителей в определенных положениях) и методу Фри—Уилсона [30, 31] (индикаторные переменные, описывающие присутствие в определенных положениях заданных заместителей).

Тем не менее большинство существующих топологических методов анализа связи «структура—активность» опирается на характеристики молекулы как

целого или ее изолированных элементов (например, наличие определенных структурных фрагментов или число их вхождений в анализируемую структуру [32—34]). При этом информация о взаимном расположении структурных элементов молекулы почти теряется, и такие методы нельзя считать эффективными с точки зрения достаточной общности и удобства применения в задачах, связанных с исследованием специфических видов активности. В то же время предложено несколько надструктурных подходов, базирующихся на рассмотрении так называемой суперструктуры — топологической сетки, на которую можно наложить структуры анализируемой серии и построить их однородное описание.

Следует отметить, что понятие суперструктуры или гиперструктуры достаточно широко применяется в химической информатике [35, 36]. Суперструктуры позволяют получить максимально компактное представление структур органических соединений, удобное для хранения в базах данных. Делаются попытки учитывать при построении таких гиперструктур качественные данные об активности соединений [37], однако из-за требований компактности и соответствующих алгоритмов генерации гиперструктур они оказываются практически непригодными для интерпретации. С другой стороны, целый ряд подходов, основанных на построении суперструктур, разработан специально для анализа количественной связи между структурой и активностью органических соединений. Обсудим их более подробно.

Метод DARC/PELCO

Метод DARC/PELCO (Méthode de perturbation d'environnements limités concentriques ordonnés — метод возмущения ограниченного концентрического упорядоченного окружения) [38—42] разработан Ж.-Э. Дюбуа, К. Мерсье и сотр. для целей прогнозирования свойств химических соединений. В рамках этого метода исследуемые структуры рассматриваются как состоящие из ядра — базовая подструктура соединений серии — и окружения, в котором атомы и связи разделяются на последовательные концентрические уровни в зависимости от расстояния до ядра, причем те или иные позиции окружения могут заполняться атомами и связями разных типов. Вводится также понятие следа популяции, объединяющего все структурные элементы, присутствующие хотя бы в одном соединении исследуемой серии. След строится динамически с помощью процедуры пошагового наложения структур [42].

В наиболее простом варианте метода в качестве структурных дескрипторов выступают индикаторные переменные, характеризующие присутствие в каждой позиции окружения конкретных типов атомов или связей («топохроматические сайты»). Активность структуры в целом представляется как сумма некой базовой активности ядра и «возмущений», вносимых всеми занятыми сайтами окружения. Величины этих вкладов определяются с помощью регрессионного анализа. Таким образом, метод DARC/PELCO можно рассматривать как детализацию и обобщение подхода Фри—Уилсона.

В ходе анализа на основе первичных дескрипторов при необходимости могут задаваться производные дескрипторы, описывающие одновременное присутствие двух или нескольких базовых структурных элементов — дескрипторы взаимодействия, присутствие одного базового структурного элемента в отсутствие другого — дескрипторы исключения или эквивалентность определенных элементов — дескрипторы эквивалентности. Такие сложные дескрипторы строятся интерактивно с помощью логических операций. Дескрипторы взаимодействия и исключения служат для уточнения модели и учета отклонений от аддитивности. Дескрипторы эквивалентности отражают сходство во влиянии определенных структурных модификаций и позволяют построить такие удобные для интерпретации переменные, как удлинение цепи, разветвление или прямое замещение в ароматическом ядре.

Метод DARC/PELCO сразу после его публикации стал активно применяться для исследования различных видов биологической активности (в частности, методология формирования сложных дескрипторов была разработана при моделировании метаболизма спиртов [43—45]). В качестве примера приведем модель (r = 0,99), которая была построена для серии 3-тиенилгликолятов, проявляющих антихолинергическую активность [46]. На рис. 1 она представлена в виде «карты активности», или топоинформационной

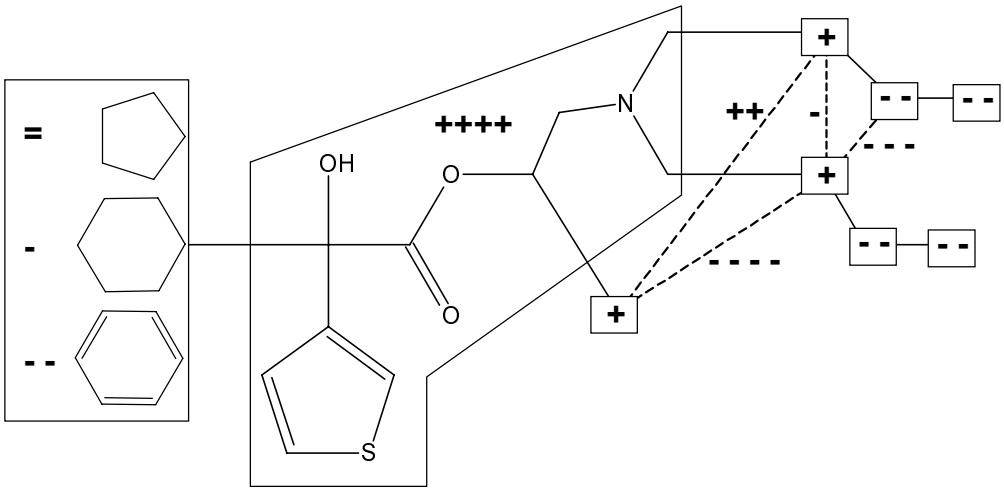


Рис. 1. Представление связи антихолинергической активности 3-тиенилгликолятов со структурой по методу DARC/PELCO

диаграммы. Символами +/−/= обозначен знак влияния (положительное/отрицательное/безразличное) и относительная величина влияния на активность тех или иных особенностей структуры (наличие определенных атомов, связей или фрагментов). Используя эти данные, авторам удалось сконструировать структуру, которая в 10 раз превосходит по активности лучшее из ранее исследованных соединений.

Проводились исследования по развитию методов семейства DARC. Например, универсальная система PATQSAR (Population Analysis by Topologically based QSAR — анализ популяций на основе топологического QSAR) [47, 48] позволяет строить тополого-физико-химические модели, учитывающие наряду с топохроматическими переменными также те или иные «внешние» физико-химические свойства молекулы в целом, прежде всего липофильность и квантовомеханические параметры, связанные с механизмом действия соединений [42, 49]. При этом влияющие на каждый из параметров заместители могут иметь различный вес, что вносит дополнительные элементы неопределенности в построение модели.

Помимо задач QSAR, метод DARC/PELCO успешно применялся для исследования связи структуры с другими локальными свойствами молекул, в том числе химическими сдвигами в спектрах ЯМР [50]. К сожалению, в последние несколько лет развитие этого направления, по-видимому, прекратилось.

Подход Менона и Каммараты

Этот подход был предложен Г. Меноном и А. Каммаратой [51] при разработке способов классификации структур органических соединений по видам проявляемой активности. В качестве основы для описания структуры предлагается использовать совокупность локальных молекулярных дескрипторов (свойств атомов и связей). Обработка их методом анализа главных компонент [51, 52] позволяет выявить скрытые переменные, облегчающие классификацию. Сам процесс формирования наборов локальных дескрипторов опирается на предварительное построение суперструктуры для исходной серии соединений посредством простейшего возможного наложения структур друг на друга «химически согласованным способом», что обеспечивает разумное сопоставление локальных свойств для различных структур.

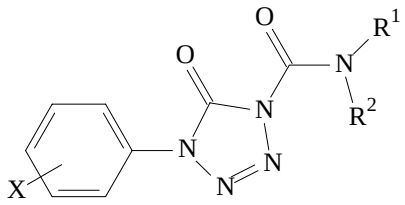
В работе [51] этот подход был использован для классификации соединений четырех фармакологических классов: α- и β-адренергических агентов, холинергических агентов и стимуляторов центральной нервной системы. С учетом изостерического и структурного факторов в качестве локальных дескрипторов выбирались величины относительной молекулярной рефракции групп в различных положениях суперструктуры, характеризующие их стерические требования. Даже при такой небольшой детализации описания структур этот подход обеспечивает почти полную классификацию исследуемых соединений по типу активности на основе значений трех главных компонент. Его недостатками можно считать затруднительность и неоднозначность химически согласованного наложения в случае структурно разнородных соединений и сложность структурной интерпретации результатов. К сожалению, данный метод не получил дальнейшего развития.

Позиционный анализ

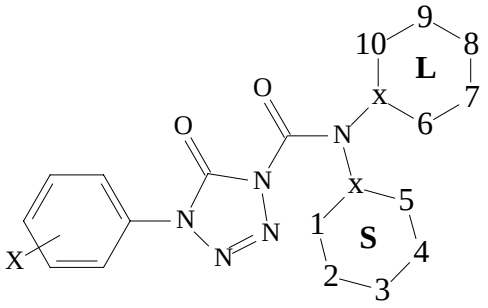
В 80-е гг. прошлого столетия Ф. Маги [53—56] предложил новый метод моделирования биохимических процессов — так называемый позиционный анализ, ориентированный на выявление структурных особенностей, ответственных за взаимодействие между молекулами активного соединения (лиганда) и биологической мишенью.

Процесс моделирования основывается на построении гипермолекулы, которая представляет собой простейшую общую структуру, объединяющую все занятые позиции в рассматриваемом ряду соединений. Каждое положение гипермолекулы можно охарактеризовать набором позиционных дескрипторов, которые необходимо отличать от традиционных дескрипторов заместителей. Автор предлагает использовать набор параметров, описывающих основные типы взаимодействия лиганда с мишенью — гидрофобные взаимодействия, поляризуемость, электростатические и стерические взаимодействия, водородные связи, а также включать индикаторные переменные, характеризующие определенные особенности структуры. Обработка позиционных дескрипторов методом множественной линейной регрессии дает модельное уравнение связи «структура—активность».

Этот подход был успешно применен, в частности при изучении гербицидов на основе замещенных тетразолинонов [54]. Для ряда соединений с общей структурой **1** была построена гипермолекула, имеющая вид **2** (L — бóльшая группа, S — меньшая группа):



1



2

Так, в случае группы NPr_2 в гипермолекуле заняты положения 1, 2, 6, 7, а в случае группы NEt(*i*-Pr) положения 1, 6, 10. Полученное автором регрессионное уравнение (r = 0,833) связывает гербицидную активность с такими параметрами, как сумма констант Гаммета для заместителей в фенильной группе, сумма стерических параметров Чартона для *мета*- и *пара*-замещенной фенильной группы, индикатор присутст-

вия атома в позиции 5 гипермолекулы и вклад заместителя в липофильность.

Существенным недостатком данного подхода является то, что с помощью физико-химических параметров описываются взаимодействия с мишенью лишь для отдельных положений гипермолекулы, хотя структурные различия имеют место и в других частях структуры. Кроме того, формальный протокол совмещения фрагментов молекул на основе размера групп не учитывает другие факторы, которые могут оказывать значительное влияние на активность. Наконец, используемый метод множественной линейной регрессии не позволяет корректно проанализировать связь дескрипторов с активностью для различных положений гипермолекулы.

Метод минимального топологического различия (MTD)

Разработанный З. Симоном с сотр. метод минимального топологического различия [57—59] исходит из предположения, что при исследовании взаимосвязи «структура—активность» для многих рядов соединений и видов активности прежде всего важно рассмотреть стерические соответствия.

Минимальное топологическое различие (MTD — minimal topological difference) молекулы M_i относительно стандарта S определяется как число атомов, которые невозможно совместить при поатомном наложении M_i на S. Атомами водорода при этом пренебрегают. Предполагается, что форма биомишени (рецептора) неизвестна и в ней могут существовать стерически безразличные области, причем наиболее вероятно, что рецепторная полость будет соответствовать по форме тому из возможных стандартов, который приводит к наилучшей корреляции экспериментальных активностей $A_i^{\exp}$ с соответствующими величинами MTD. Таким образом, задача состоит в определении оптимального стандарта для данного множества из N молекул с известными активностями $A_i^{\exp}$.

Процедура анализа исследуемой выборки может быть представлена следующим образом.

1. Необходимо осуществить поатомное наложение всех молекул, в результате чего получается гипермолекула H — атомная сеть с K вершинами.

2. При наложении структуры любого соединения M_i на гипермолекулу H получается набор индикаторных переменных. Следовательно, M_i характеризуется вектором $x_i = \{x_{i1}, \ldots, x_{iK}\}$, где x_{ij} принимает значение 1, если вершина j занята в M_i, и 0 — в противном случае.

Тогда минимальное топологическое различие вычисляется по формуле:

$$\mathrm{MTD}_i = S + \sum_{j=1}^{K} \varepsilon_j x_{ij}$$

где S — число вершин, занятых стандартом при наложении на гипермолекулу M; величина ε_i принимает значение -1 для вершин, занятых стандартом, $+1$ для незанятых вершин и 0 для вершин, которые безразличны для стерического взаимодействия.

Для нахождения оптимального стандарта необходимо минимизировать ошибку корреляции относительно ε_i. В результате получается «карта» области взаимодействия рецептора с N молекулами анализируемой серии. Величины x_{ij} можно использовать также для построения регрессионной модели по аналогии с методом Фри—Уилсона [60]. Главными недостатками данного подхода являются большое число параметров и отсутствие учета других характеристик, помимо стерических. В работе [61] на примере психотомиметической активности замещенных фенилалкиламинов было показано, что сочетание величины MTD с физико-химическими дескрипторами (например, расчетным значением липофильности или величинами энергии низшей свободной молекулярной орбитали и эффективных зарядов на атомах, вычисленными методами квантовой механики) позволяет улучшить качество корреляций и предсказательную способность моделей.

В дальнейшем Симон с сотр. предложил значительно усовершенствованный вариант метода минимального топологического различия, получивший название MTD-PLS [62]. Как ясно из названия, в этом подходе для статистического анализа данных применяется модифицированный вариант регрессии частичных наименьших квадратов [63], что позволяет устранить проблемы, связанные с большим числом дескрипторов. Кроме того, было предложено в дополнение к индикаторным переменным (параметрам занятости) использовать для анализа связи «структура—активность» физико-химические характеристики атомов и структурных фрагментов, в том числе атомную массу, объем фрагмента, частичный атомный заряд, гидрофобность фрагмента [64], поляризуемость [65, 66] и способность выступать в качестве донора и акцептора водородных связей [67]. Это позволило уточнить и улучшить модели по сравнению с «классическим» вариантом метода MTD. Для незанятых положений гипермолекулы предлагается использовать нулевые значения всех дескрипторов [64], что не всегда правомерно (например, с этим нельзя согласиться в случае гидрофобных констант фрагментов).

Данный подход применялся для анализа скорости гидролиза эфиров уксусной кислоты под действием ацетилхолинэстеразы [65, 66], связывания стероидов с эстрогеновым рецептором [67], а также для изучения способности полигалогенированных производных дибензодиоксина, дибензофурана и бифенила к индукции гидроксилазы ароматических углеводородов и связыванию с цитозольным Ah-рецептором печени крыс [64]. В качестве примера рассмотрим ход моделирования активности по отношению к гидроксилазе ароматических углеводородов. Первоначальная модель включала 69 соединений и 105 дескрипторов и обладала следующими статистическими параметрами: $r^2 = 0{,}728$, $q^2 = 0{,}553$. В результате поэтапного отбрасывания некоторых дескрипторов (в частности, параметров занятости) и «выпадающих соединений» была получена модель, основанная на данных для 43 соединений и включающая 52 структурные переменные ($r^2 = 0{,}819$, $q^2 = 0{,}732$). В отличие от других рассматриваемых надструктурных подходов, в данном методе гипермолекула является «квазитрехмерной», так что не всегда достигается полное совмещение соответствующих друг другу атомов различных структур. На основе модели в гипермолекуле можно выделить вершины выгодные, невыгодные и безразличные для активности (с точки зрения стерических факторов).

Сравнение моделей MTD-PLS с результатами других методов анализа связи «структура—активность» и (в ряде случаев) с данными рентгеноструктурного анализа комплексов биомишени с лигандом в целом свидетельствует об их согласованности [64, 66, 67]. В то же время недостаточно убедительным кажется отбрасывание большого числа выпадающих соединений при построении окончательной модели (иногда до 30—50% обучающей выборки). В некоторых случаях это можно объяснить структурными факторами (в частности, в работе [64] оказалось необходимым исключить производные бифенила с существенно неплоской конформацией, хотя предлагаемый метод и не требует полного совмещения структур анализируемых соединений), а в других — авторы опираются на чисто статистические критерии [67]. При этом утверждается, что даже модели с недостаточно высокой предсказательной способностью могут давать полезную информацию о характере взаимодействий лигандов с биомишенью для фрагментов лигандов, отличающихся достаточной структурной вариабельностью [67].

Метод анализа топологии молекулярного поля (MFTA)

Метод анализа топологии молекулярного поля (MFTA, Molecular Field Topology Analysis) был предложен нами [68, 69] как обобщение и расширение ряда существующих подходов к поиску количественной связи «структура—активность». Он развивает характерную для современных исследований QSAR тенденцию использования топологических и квазитопологических подходов при моделировании взаимодействия активных соединений-лигандов с биологическими мишенями и в известном смысле может рассматриваться как топологический аналог метода CoMFA.

Метод опирается на предположение, что во многих случаях рассмотрение топологического, а не пространственного совмещения структур обучающей выборки позволит снять проблемы, характерные для трехмерных методов QSAR, и получить общую методологию прогнозирования биологической активности органических соединений, обусловленной специфическим (рецепторным) механизмом действия. При этом в качестве дескрипторов структуры молекулы применяются ее локальные физико-химические параметры — свойства атомов и связей, допускающие быструю оценку непосредственно по структурной формуле соединения.

Процесс анализа количественной связи «структура—активность» в рамках метода анализа топологии молекулярного поля [70] включает две процедуры — построение модели связи «структура—активность» и прогнозирование активности новых структур. Общая схема анализа по методу MFTA приведена на рис. 2.

Прежде всего для ряда структур с известными из эксперимента значениями активности (обучающая выборка) автоматически строится молекулярный суперграф — простой граф (не обязательно минимальный или уникальный), в виде подграфа которого может быть представлена каждая из структур выборки. Способ построения молекулярного суперграфа иллюстрирует рис. 3. В этой процедуре структуры обучающей выборки обрабатываются последовательно и на

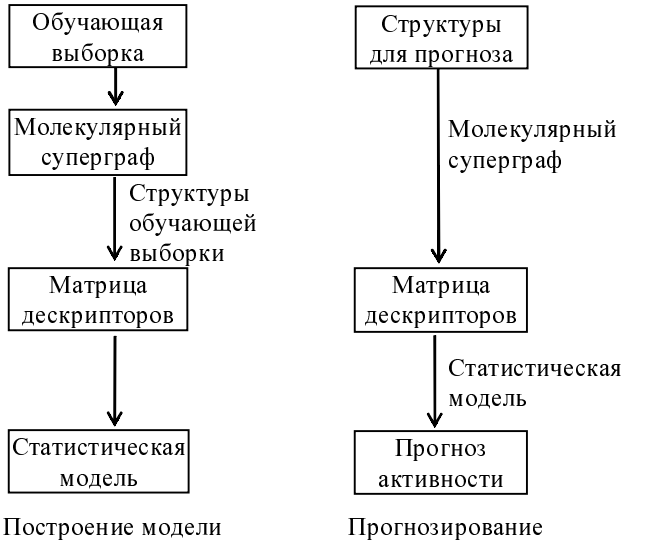


Рис. 2. Общий ход анализа количественной связи «структура—активность» по методу MFTA

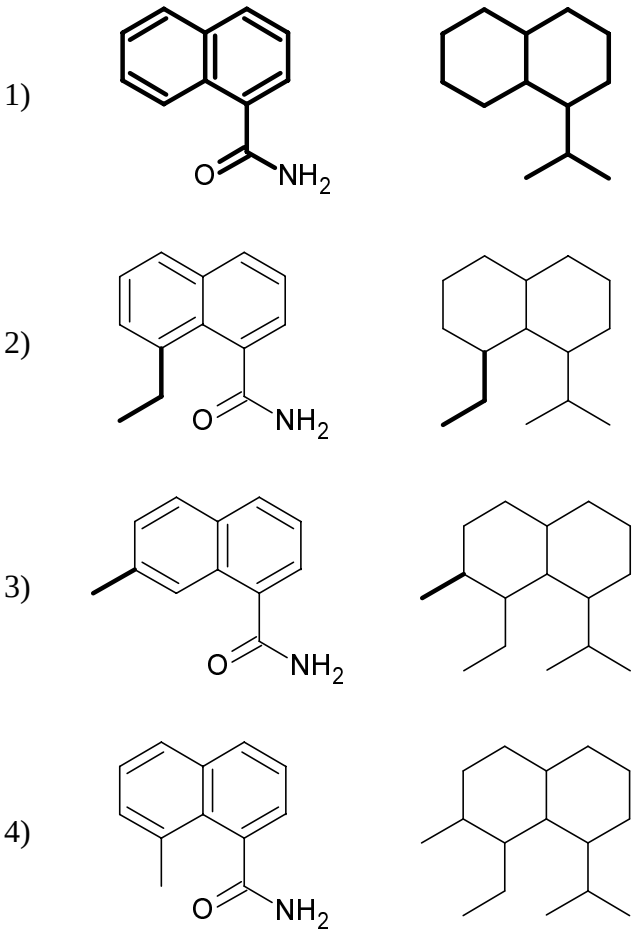


Рис. 3. Построение молекулярного суперграфа для серии структур.

Жирными линиями на каждом шаге выделены фрагменты, которые не вошли в пересечение с полученным к этому моменту молекулярным суперграфом и были дополнительно введены в него

каждом шаге определяется пересечение между построенным к настоящему моменту (первоначально пустым) суперграфом и очередной структурой выборки. Затем суперграф дополняется атомами, которые не вошли в пересечение (например, этильный заместитель в структуре для шага 2 на рис. 3). Существенная особенность применяемого алгоритма поиска пересечений [70] состоит в том, что при наличии нескольких возможных вариантов соответствия атомов структуры и вершин суперграфа предпочтение отдается вершине, обеспечивающей максимальное сходство распределений локальных свойств в их ближайшем окружении. Как правило, данный алгоритм позволяет быстро найти наиболее удачный вариант отображения всей структуры на суперграф благодаря учету наиболее важных атомных параметров, определяющих взаимодействие молекул с биологической мишенью.

Одно из преимуществ предложенного метода связано с тем, что он использует открытый набор дескрипторов, допуская его модификацию и расширение в зависимости от особенностей конкретной задачи моделирования связи «структура—активность». Совокупность разработанных к настоящему времени дескрипторов позволяет адекватно описать основные типы межмолекулярных взаимодействий, значимых для связывания молекул активного лиганда с биологическими мишенями [70]. Это электростатические (например, эффективный заряд на атоме) и стерические дескрипторы (в частности, ван-дер-ваальсовы радиусы атомов и групп), дескрипторы липофильности и способности к образованию водородных связей и др.

Полученный молекулярный суперграф дает возможность построить однородные векторы дескрипторов для всех структур обучающей выборки. При формировании каждого такого вектора вершинам и ребрам суперграфа, которые соответствуют атомам и связям в данной структуре, сопоставляются значения локальных дескрипторов для этих атомов и связей, например значения заряда Q на атоме и его ван-дер-ваальсова радиуса R. Незанятые вершины и ребра суперграфа помечаются «нейтральными» значениями дескрипторов (Q_0, R_0), которые отображают определенную модель свойств в соответствующих незанятых областях пространства рядом с молекулой, так что их нельзя рассматривать как «отсутствующие» значения в статистическом смысле этого термина. Кроме того, при необходимости можно провести оптимизацию этих параметров в конкретных задачах. В то же время проведенные проверки показывают, что величина нейтральных значений в определенном интервале оказывает не очень значительное влияние на качество получаемых моделей связи «структура—активность». Рис. 4 иллюстрирует принцип построения вектора дескрипторов на примере четвертой структуры, показанной на рис. 3. Как обычно принято в топологических подходах к анализу молекулярных структур, атомы водорода не рассматриваются в явном виде, а их свойства при необходимости учитываются в качестве дополнительных дескрипторов для тех атомов структуры, с которыми они связаны.

Полученные таким образом однородные векторы молекулярных дескрипторов для всех структур обучающей выборки образуют матрицу дескрипторов, которая в сочетании с экспериментальными значениями активности соединений представляет собой исходные данные для построения статистической модели связи «структура—активность». При прогнозировании активности новых структур точно так же определяется их наложение на молекулярный суперграф и строятся однородные векторы дескрипторов, а затем по модели рассчитываются величины активности.

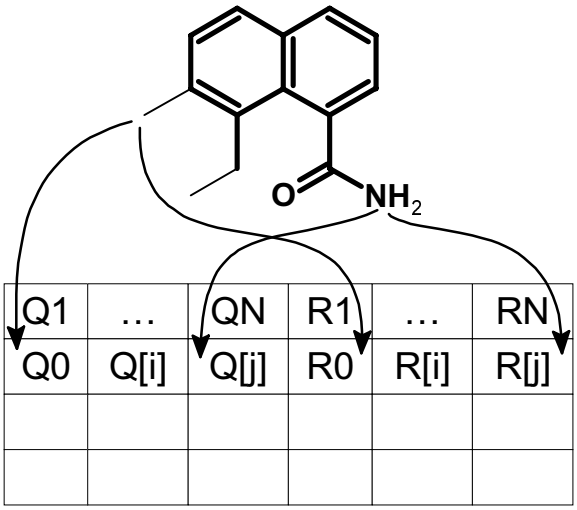


Рис. 4. Формирование однородных векторов дескрипторов для структуры при наложении ее на молекулярный суперграф.

Q[j] и R[j] — значения заряда и ван-дер-ваальсова радиуса для j-го атома структуры, соответствующего N-му положению молекулярного суперграфа; Q0 и R0 — нейтральные значения заряда и ван-дер-ваальсова радиуса

Модель MFTA содержит также информацию о влиянии на активность локальных дескрипторов, соответствующих различным положениям молекулярной структуры, что может помочь при изучении механизма действия исследуемых соединений, направленном конструировании новых перспективных структур [71] и определении реперных точек для корректного пространственного совмещения структур (в случае, если требуется трехмерный анализ).

В большинстве случаев для построения MFTA моделей связи «структура—активность» применяется метод регрессии частичных наименьших квадратов PLS (Partial Least Squares). Как уже отмечалось, этот относительно новый статистический метод [2] позволяет получать устойчивые и корректные модели, когда число независимых переменных (дескрипторов) сопоставимо или превосходит число объектов (соединений) и/или действуют иные факторы, вызывающие корреляцию между независимыми переменными. Метод PLS основан на переходе от большого числа исходных переменных к небольшому числу скрытых переменных (факторов), которые являются их линейными комбинациями и взаимно ортогональны (независимы друг от друга) [72]. Для повышения устойчивости и предсказательной способности PLS-модели может производиться отбор дескрипторов [73]. Иногда при построении модели связи «структура—активность» в методе MFTA применяется методология искусственных нейронных сетей, которая позволяет выявлять нелинейные зависимости связи между дескрипторами и активностью [74].

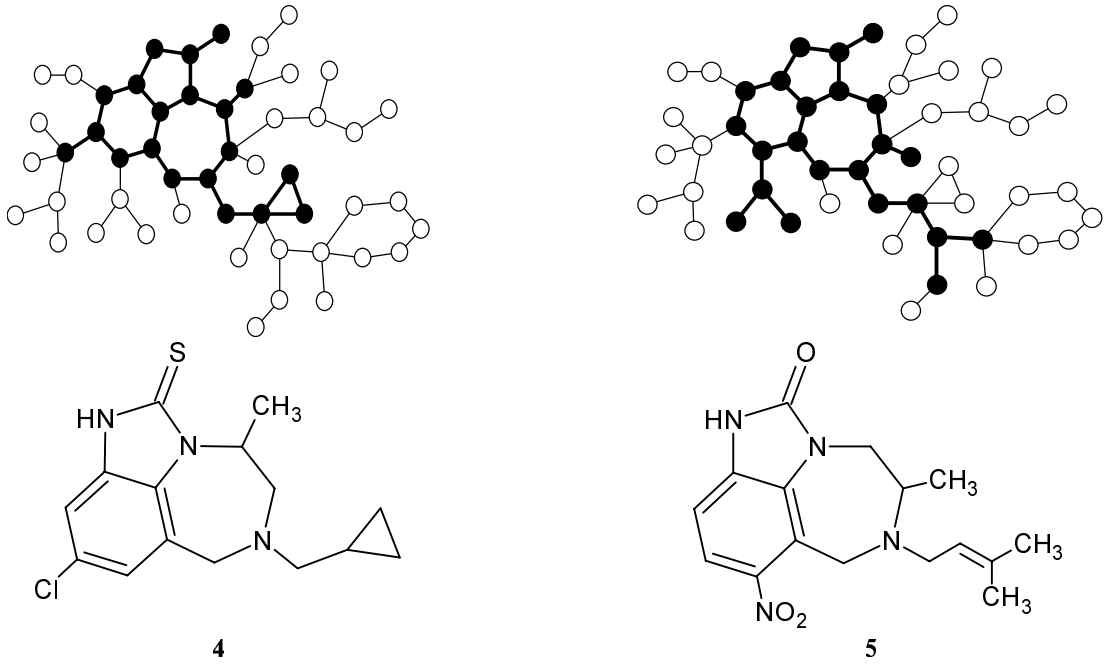
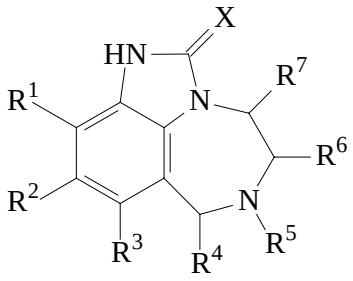


Рис. 5. Структура молекулярного суперграфа и примеры наложения структур обучающей выборки производных тетрагидроимидазобензодиазепинона

Метод анализа топологии молекулярного поля реализован в виде компьютерной программы и нашел применение при решении целого ряда задач, связанных с исследованием соотношений «структура—биологическая активность» и молекулярным дизайном потенциально активных органических соединений различных классов [70, 75]. В качестве примера кратко рассмотрим результаты, полученные при изучении связи структуры и анти-ВИЧ-активности производных тетрагидроимидазобензодиазепинона (TIBO), действующих по механизму ингибирования обратной транскриптазы ВИЧ-1 [75, 76].

Обучающая выборка содержала 73 соединения общей формулы **3**,



3

активность которых была представлена в виде величины $\log(1/IC_{50})$. Оптимальная по статистическим характеристикам модель (число PLS-факторов $N_F = 5$, коэффициент корреляции $r = 0{,}942$, параметр перекрестного контроля $q^2 = 0{,}686$) включала в качестве дескрипторов величины атомных зарядов Q, ван-дер-ваальсовых радиусов R и групповой липофильности L_g (сумма вкладов в липофильность данного атома и связанных с ним атомов водорода). На рис. 5 представлена структура молекулярного суперграфа с примером наложения на нее двух структур обучающей выборки.

Имеющиеся результаты рентгеноструктурного анализа комплексов обратной транскриптазы ВИЧ-1 с некоторыми производными TIBO позволили провести сопоставление данных MFTA-модели по влиянию особенностей структуры лиганда на его активность с реальным строением сайта связывания лиганда. Была обнаружена высокая степень их согласованности. В частности, фрагменты лиганда, в которых увеличение заряда ведет к возрастанию активности, взаимодействуют с участками фермента с относительно более отрицательным электростатическим потенциалом, а фрагменты лиганда, где выгодны более отрицательные значения заряда, взаимодействуют с участками фермента, имеющими относительно более положительный потенциал. Аналогично фрагменты лиганда с положительным вкладом липофильности в активность связываются с более гидрофобными, а с отрицательным вкладом — с более гидрофильными областями белка. Наконец, рис. 6 иллюстрирует соответствие между вкладами стерических дескрипторов (ван-дер-ваальсовы радиусы) и стерическими особенностями сайта связывания. Как видно, участки лиганда, где желательно введение дополнительных атомов или увеличение размера имеющихся атомов, находятся в полости биомишени, а участки с противоположным знаком влияния стерических дескрипторов — вблизи стенок «кармана». Помимо проведения структурной интерпретации предложенную модель можно также использовать для компьютерного конструирования потенциально более активных структур данного класса [76].

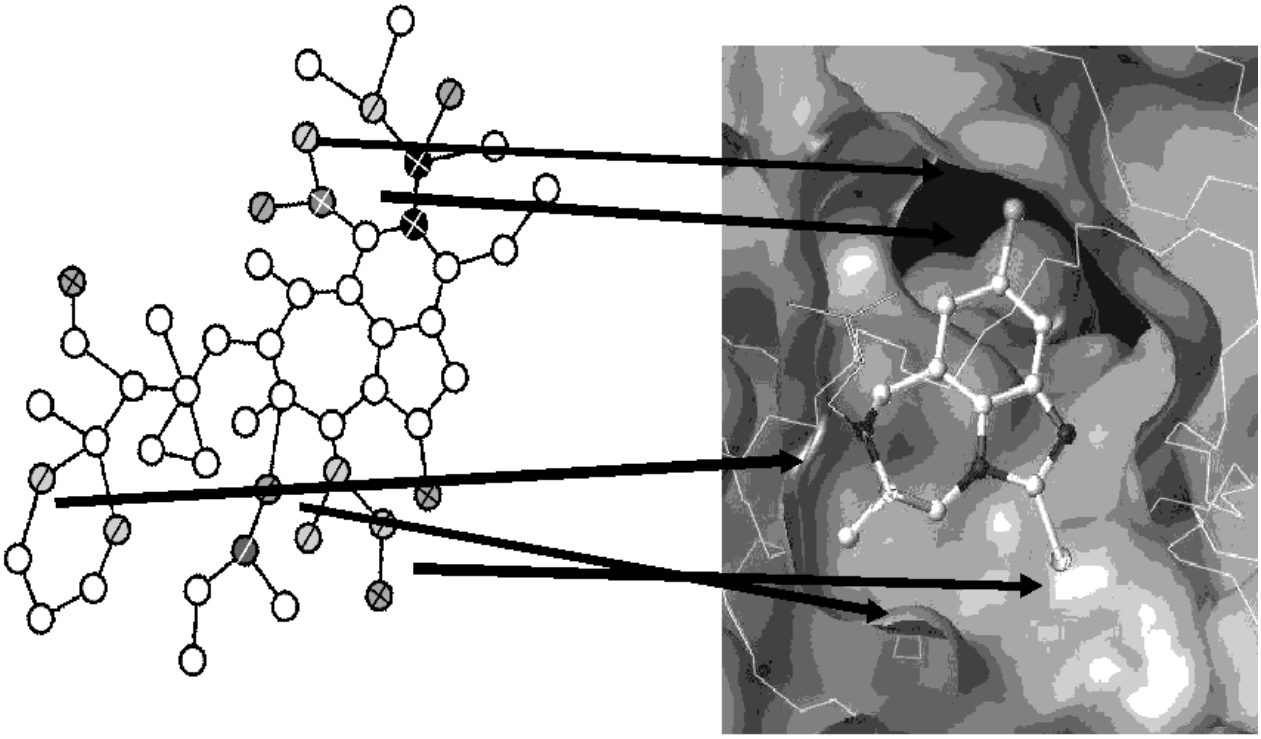


Рис. 6. Соответствие структуры биомишени — обратной транскриптазы ВИЧ-1 и характера влияния стерических дескрипторов на активность производных тетрагидроимидазобензодиазепинона

Заключение

Сравнивая рассмотренные надструктурные методы анализа количественной связи «структура—активность» с другими подходами (включая широко применяемые классические методы QSAR и анализ трехмерной структуры), можно отметить, что они часто обеспечивают более высокое качество описания и более высокую прогнозирующую способность моделей. Применение этих методов наиболее целесообразно по отношению к рядам родственных органических соединений, активность которых непосредственно связана со специфическим (рецепторным) взаимодействием лиганд-биомишень, а прямой учет трехмерной структуры соединений представляет определенные трудности или является излишним.

В большинстве задач анализа «структура—активность» недостаточно учитывать только дескрипторы занятости положения (наличия атомов). Необходимо включать в анализ локальные физико-химические параметры, отражающие характер взаимодействия лиганда с мишенью. Весьма важен и используемый метод наложения структур, который должен учитывать распределение этих параметров и особенности структуры соединений.

Наибольшие возможности для анализа связи «структура—активность» предоставляют такие интенсивно развивающиеся в последние годы методы, как анализ топологии молекулярного поля MFTA и предложенный недавно усовершенствованный метод минимального топологического различия MTD-PLS.

ЛИТЕРАТУРА

1. *Cramer R.D., Patterson D.E., Bunce J.D.* J. Am. Chem. Soc., 1988, v. 110, № 18, p. 5959—5967.
2. *Geladi P., Kowalski B.R.* Anal. Chim. Acta, 1986, v. 185, № 1, p. 1—17.
3. *Wagener M., Sadowski J., Gasteiger J.* J. Am. Chem. Soc., 1995, v. 117, № 29, p. 7769—7775.
4. US Patent № 5025388, 1991.
5. *Thibaut U.* In: 3D QSAR in Drug Design. Theory, Methods and Applications. Ed. H. Kubinyi. Leiden: ESCOM, 1993, p. 661—696.
6. *Wilcox R.E., Huang W.-H., Brusniak M.-Y.K., Wilcox D.M., Pearlman R.S., Teeter M.M., DuRand C.J., Wiens B.L., Neve K.A.* J. Med. Chem., 2000, v. 43, № 16, p. 3005—3019.
7. *Schaal W., Karlsson A., Ahlsén G., Lindberg J., Andersson H.O., Danielson U.H., Classon B., Unge T., Samuelsson B., Hulten J., Hallberg A., Karlen A.* Ibid., 2001, v. 44, № 2, p. 155—169.
8. *McFarland J.W.* Ibid., 1992, v. 35, № 14, p. 2543—2550.
9. *Kim K.H., Martin Y.C.* Ibid., 1991, v. 34, № 7, p. 2056—2060.
10. *Kim K.H., Martin Y.C.* J. Org. Chem., 1991, v. 56, № 8, p. 2723—2729.
11. *Baskin I.I., Tikhonova I.G., Palyulin V.A., Zefirov N.S.* J. Med. Chem., 2003, v. 46, № 19, p. 4063—4069.
12. *Norinder U.* J. Comp.-Aided Mol. Design., 1990, v. 4, № 4, p. 381—389.
13. *Carrupt P.-A., Gaillard P., Billois F., Weber P., Testa B., Meyer C., Pérez S.* In: Lipophilicity in Drug Action and Toxicology. Ed. R. Mannhold Weinheim etc.: VCH, 1996, p. 195—217.
14. *Poso A., Navajas C., Gynter J.* 11th Eur. Symp. on Quantitative Structure-Activity Relationships, September 1—6, 1996, Lausanne, Switzerland, p. P-40D.
15. *Baroni M., Costantino G., Cruciani G., Riganelli D., Valigi R., Clementi S.* Quant. Struct.-Act. Relat., 1993, v. 12, № 1, p. 9—20.

16. *Pastor M., Cruciani G., Clementi S.* J. Med. Chem., 1997, v. 40, № 10, p. 1455—1464.
17. *Cruciani G., Watson K.A.* Ibid., 1994, v. 37, № 16, p. 2589—2601.
18. *Cramer R.D., Clark R.D., Patterson D.E., Ferguson A.M.* Ibid., 1996, v. 39, № 16, p. 3060—3069.
19. *Зефиров Н.С., Петелин Д.Е., Палюлин В.А., Макфарланд Дж.У.* Докл. АН, 1992, т. 327, № 4—6, с. 508—512.
20. *Lipkin M., Salt D., Wynn W.* Computer-Assisted Lead Finding and Optimization. Eds H. Waterbeemd, B. Testa, G. Folkers. Basel: VHCA; Weinheim etc.: Wiley-VCH, 1997, p. 433—442.
21. *Tominaga Y., Fujiwara I.* J. Chem. Inf. Comput. Sci., 1997, v. 37, № 6, p. 1152—1157.
22. *Sulea T., Oprea T.I., Muresan S., Chan S.L.* Ibid., 1997, v. 37, № 6, p. 1162—1170.
23. *Patterson D.E., Cramer R.D., Ferguson A.M., Clark R.D., Weinberger L.E.* J. Med. Chem., 1996, v. 39, № 16, p. 3049—3059.
24. *Cramer R.D.* Ibid., 2003, 46, № 3, p. 374—388.
25. *Cramer R.D., Jilek R.J., Guessregen S, Clark S.J., Wendt B., Clark R.D.* Ibid., 2004, v. 47, № 27, p. 6777—6791.
26. *Silverman B.D., Platt D.E.* Ibid., 1996, v. 39, № 11, p. 2129—2140.
27. *Hansch C., Maloney P.P, Fujita T., Muir M.* Nature, 1962, v. 194, p. 178—180.
28. *Hansch C., Fujita T.* J. Am. Chem. Soc., 1964, v. 86, № 8, p. 1616—1626.
29. *Hansch C.* Acc. Chem. Res., 1969, v. 2, № 8, p. 232—239.
30. *Free S.M., Wilson J.M.* J. Med. Chem., 1964, v. 7, № 7, p. 395—399.
31. *Fujita T., Ban T.* Ibid., 1971, v. 14, № 2, p. 148—152.
32. *Артеменко Н.В., Баскин И.И., Палюлин В.А., Зефиров Н.С.* Докл. АН, 2001, т. 381, № 2, с. 203—206.
33. *Lagunin A., Stepanchikova A., Filimonov D., Poroikov V.* Bioinformatics, 2000, v. 16, № 8, p. 747—748.
34. *Khamin A.S., Palyulin V.A., Tkachenko S.E., Zefirov N.S.* 11th Eur. Symp. on Quantitative Structure-Activity Relationships, September 1—6, 1996, Lausanne, Switzerland, p. P22A.
35. *Brown R.D., Downs G.M., Willett P.* J. Chem. Inf. Comput. Sci., 1992, v. 32, № 5, p. 522—531.
36. *Brown R.D., Jones G., Willett P., Glen R.C.* Ibid., 1994, v. 34, № 1, p. 63—70.
37. *Downs G.M., Gill G.S., Willett P., Walsh P.T.* SAR QSAR Environ. Res., 1995, v. 3, № 4, p. 253—264.
38. *Dubois J.-E., Laurent D., Aranda A.* J. Chim. Phys., 1973, v. 70, № 11—12, p. 1616—1624.
39. *Mercier C., Mekenyan O., Dubois J.-E., Bonchev D.* Eur. J. Med. Chem., 1991, v. 26, № 6, p. 575—592.
40. *Dubois J.-E., Sobel Y.* J. Chem. Inf. Comput. Sci., 1985, v. 25, № 3, p. 326—333.
41. *Dubois J.-E., Carrier G., Panaye A.* Ibid., 1991, v. 31, № 4, p. 574—578.
42. *Sobel Y., Vizet P., Chemtob S., Barbieux F., Mercier C.* SAR QSAR Environ. Res., 1998, v. 9, № 1—2, p. 83—109.
43. *Dubois J.-E., Laurent D., Aranda A.* J. Chim. Phys., 1973, v. 70, № 11—12, p. 1608—1615.
44. *Dubois J.-E., Mercier C., Panaye A.* Acta Pharm. Jugosl., 1986, v. 36, № 1, p. 135—141.
45. *Mercier C., Fabart V., Sobel Y., Dubois J.-E.* J. Med. Chem., 1991, v. 34, № 3, p. 934—942.
46. *Mercier C., Trouiller G., Dubois J.-E.* Quant. Struct.-Act. Relat., 1990, v. 9, № 2, p. 88—93.
47. *Vizet P., Sobel Y., Mercier C.* 11th Eur. Symp. on Quantitative Structure-Activity Relationships, September 1—6, 1996, Lausanne, Switzerland, p. P38A.
48. *Mercier C., Barbieux F., Sobel Y.* 11th Eur. Symp. on Quantitative Structure-Activity Relationships, September 1—6, 1996, Lausanne, Switzerland, p. P11C.
49. *Mercier C., Chemtob S., Vizet P., Sobel Y.* 9th Int. Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences, September 2000, Bourgas, Bulgaria, p. I.15.
50. *Carabedian M., Dubois J.-E.* J. Chem. Inf. Comput. Sci., 1998, v. 38, № 2, p. 100—107.
51. *Menon G.K., Cammarata A.* J. Pharm. Sci., 1977, v. 66, № 3, p. 304—314.
52. *Cammarata A., Menon G.K.* J. Med. Chem., 1976, v. 19, № 6, p. 739—747.
53. *Magee P.S.* Quant. Struct.-Act. Relat., 1990, v. 9, № 4, p. 202—215.
54. *Bell A.R., Covey R.A., Relyea D.I.* Proc. British Crop Protection Conference — Weeds, 1987, v. 1, p. 249—255.
55. *Magee P.S.* In: QSAR: Rational Approaches to the Design of Bioactive Compounds. Eds C. Silipo, A. Vittoria. Amsterdam etc.: Elsevier, 1991, p. 549—552.
56. *Magee P.S.* In: Rational Approaches to Structure, Activity, and Ecotoxicology of Agrochemicals. Eds W. Draber, T. Fujita. Boca Raton etc.: CRC Press, 1992, p. 79—101.
57. *Simon Z., Badilescu I., Racovitan T.* J. Theor. Biol., 1977, v. 66, № 3, p. 485—495.
58. *Balaban A.T., Chiriac A., Motoc I., Simon Z.* Steric fit in QSAR. Lecture notes in chemistry, v. 15. Berlin etc.: Springer, 1980.
59. *Simon Z., Holban S., Motoc I.* Rev. Roum. Biochim., 1979, v. 16, № 2, p. 141—145.
60. *Simon Z.* 11th Eur. Symp. on Quantitative Structure-Activity Relationships, September 1—6, 1996, Lausanne, Switzerland, p. P-60A.
61. *Mracec M., Mracec M., Kurunczi L., Nusser T., Simon Z., Náray-Szabó G.* J. Mol. Struct. (THEOCHEM), 1996, v. 367, p. 139—149.
62. *Olah M., Kurunczi L., Simon Z.* Annals West Univ. Timisoara, Ser. Chem., 1998, v. 7, № 1, p 101—118.
63. *Timofei S., Schmidt W., Kurunczi L., Simon Z.* Dyes and Pigments, 2000, v. 47, № 1, p. 5—16.
64. *Oprea T.I., Kurunczi L., Olah M., Simon Z.* SAR QSAR Environ. Res., 2001, v. 12, № 1—2, p. 75—92.
65. *Olah M., Kurunczi L., Bologa C., Oprea T.I., Simon Z.* Annals West Univ. Timisoara, Ser. Chem., 2001, v. 10, № 2, p. 863—870.
66. *Kurunczi L., Olah M., Oprea T.I., Bologa C., Simon Z.* J. Chem. Inf. Comput. Sci., 2002, v. 42, № 4, p. 841—846.
67. *Kurunczi L., Seclaman E., Oprea T.I., Crisan L., Simon Z.* J. Chem. Inf. Model., 2005, v. 45, № 5, p. 1275—1281.
68. *Radchenko E.V., Palyulin V.A., Zefirov N.S.* 11th Eur. Symp. on Quantitative Structure-Activity Relationships, September 1—6, 1996, Lausanne, Switzerland, P. P-21A.
69. *Зефиров Н.С., Палюлин В.А., Радченко Е.В.* Докл. АН, 1997, т. 352, № 5, с. 630—633.
70. *Palyulin V.A., Radchenko E.V., Zefirov N.S.* J. Chem. Inf. Comp. Sci., 2000, v. 40, № 3, p. 659—667.
71. *Radchenko E.V., Palyulin V.A., Zefirov N.S.* In: Molecular Modeling and Prediction of Bioactivity. Eds K. Gundertofte, F.S. Jшrgensen. New York etc.: Kluwer Plenum, 2000, p. 460—461.
72. *Martens H., Naes T.* Multivariate Calibration. Chichester etc.: Wiley, 1989.
73. *Palyulin V.A, Radchenko E.V., Baranova O.D., Oliferenko A.A., Zefirov N.S.* In: Designing Drugs and Crop Protectants: processes, problems and solutions. Eds M. Ford, D. Livingstone, J. Dearden, H. Waterbeemd. Malden etc.: Blackwell, 2003, p. 188—190.
74. *Radchenko E.V., Baranova O.D, Palyulin V.A, Zefirov N.S.* In: Designing Drugs and Crop Protectants: processes, problems and solutions. Eds M. Ford, D. Livingstone, J. Dearden, H. Waterbeemd. Malden etc.: Blackwell, 2003, p. 317—318.
75. *Radchenko E.V., Belenikin M.S., Sokolov A.A., Palyulin V.A., Zefirov N.S.* 15th Eur. Symp. on Quantitative Structure-Activity Relationships & Molecular Modelling. Istanbul, Turkey, September 5—10, 2004, p. 171.
76. *Радченко Е.В., Соколов А.А., Беленикин М.С., Мельников А.А., Палюлин В.А., Зефиров Н.С.* Докл. АН, 2006, в печати.