

Концепция молекулярного подобия и ее использование для прогнозирования свойств химических соединений

М.И.Скворцова, И.В.Станкевич, В.А.Палюлин, Н.С.Зефирова

Институт органической химии им. Н.Д.Зелинского Российской академии наук

119991 Москва, Ленинский просп., 47, факс (495) 135–5328

Институт элементоорганических соединений им. А.Н.Несмеянова Российской академии наук

119991 Москва, ул. Вавилова, 28, факс (495) 135–5085

Московский государственный университет им. М.В.Ломоносова, Химический факультет

119992 Москва, Ленинские горы, факс (495) 939–0290

Дан обзор основных методов определения степени подобия молекул, используемых в современной теоретической химии, а также методов прогнозирования свойств химических соединений, основанных на постулате «близкие структуры имеют близкие свойства». Приведены примеры применения указанных подходов для моделирования связи между структурой и свойствами органических соединений.

Библиография — 254 ссылки.

Оглавление

I. Введение	1074
II. Интерпретация термина «молекулярное подобие» в химии	1075
III. Теоретико-графовое представление молекул и основанные на нем описания молекулярной структуры	1076
IV. Количественные меры молекулярного подобия	1080
V. Методы прогнозирования свойств химических соединений, основанные на концепции молекулярного подобия	1083
VI. Формализация постулата «близкие структуры имеют близкие свойства» и его анализ	1084
VII. Примеры применения концепции молекулярного подобия	1085
VIII. Заключение	1090

I. Введение

Концепция подобия играет важную роль в химии. Суть ее можно сформулировать следующим образом: подобные объекты имеют подобные свойства. Понятие «подобие» используется при сравнении различных химических объектов (атомов, молекул и др.), а также реакций, баз химических данных, молекулярных свойств и т.д. Популярность термина

«подобие» в химической литературе постоянно растет. Так, за 25 лет число рефератов в журнале *Chemical Abstracts*, в которых использован термин «подобие» (similarity), увеличилось с 233 (1967–1971 гг.) до 1697 (с 1987 до середины 1992 г.).¹ Проходят международные конференции и семинары, посвященные различным аспектам концепции подобия в химии (см., например,^{2–3}); опубликован ряд монографий и обзоров по этой тематике.^{2,4–16} Разработаны коммерческие компьютерные программы, основанные на методах анализа сходства.^{17–19}

Концепция подобия в химии имеет давнюю историю и множество вариантов толкования. Она относится к таким научным понятиям, которые воспринимаются на уровне интуиции и трудно поддаются четкому определению и формализации.^{20,21} Однако широкое применение в химических исследованиях ЭВМ, накопление огромного фактического материала по структурам, свойствам и химическим реакциям различных соединений, бурное развитие таких научных областей, как математическая химия, компьютерная химия, комбинаторная химия и медицинская химия, привело к необходимости формализации ряда основных химических понятий и разработки методов их количественного описания. Понятия «молекулярная структура», «молекулярное подобие», «подобие химических реакций» входят в их число.

Концепция подобия находит применение при решении следующих задач в химии и родственными дисциплинами:

М.И.Скворцова. Кандидат физико-математических наук, старший научный сотрудник лаборатории математической химии и компьютерного синтеза ИОХ РАН. Телефон: (495)135–9089, e-mail: stan@ineos.ac.ru

И.В.Станкевич. Кандидат физико-математических наук, ведущий научный сотрудник ИНЭОС РАН. Телефон: (495)135–7998, e-mail: stan@ineos.ac.ru

Область научных интересов авторов: квантовая и математическая химия, прогнозирование свойств химических соединений.

В.А.Палюлин. Кандидат химических наук, ведущий научный сотрудник кафедры органической химии Химического факультета МГУ. Телефон: (495)939–3969, e-mail: vap@org.chem.msu.ru

Н.С.Зефирова. Академик, заведующий той же кафедрой. Телефон: (495)939–1620, e-mail: zefirov@org.chem.msu.ru

Область научных интересов авторов: органическая химия и компьютерная химия.

Дата поступления 11 ноября 2005 г.

— прогнозирование физико-химических свойств и биологической активности химических соединений;^{22–33}

— классификация и упорядочение данных в информационно-поисковых системах по химическим структурам (типа *Beilstein's Handbook of Organic Chemistry*);³⁴

— планирование химических экспериментов и теория реакционной способности;^{35–40}

— поиск биологически активных аналогов в базах химических данных с целью оптимизации структур уже известных активных соединений;^{41–51}

— кластеризация баз данных и создание представительной выборки соединений для скрининга;^{16, 52–59}

— создание баз данных с максимальным структурным многообразием для дальнейшего поиска новых базовых соединений.^{18, 60, 61}

В данном обзоре рассматривается понятие молекулярного подобия, а также основанные на нем методы прогнозирования свойств химических соединений.

Наиболее распространенным методом установления количественных соотношений между структурой и активностью/свойством соединений (QSAR/QSPR-анализ) является так называемый статистический подход,^{62–64} ряд ограничений которого и привел к развитию подхода, основанного на концепции подобия. Для реализации статистического метода необходимо иметь обучающую выборку соединений с известными значениями исследуемого свойства. Структуры рассматриваемых соединений описываются при помощи некоторого набора молекулярных дескрипторов $x_1, \dots, x_N$ (топологических, геометрических, электронных); затем строится уравнение вида

$$y = f(x_1, \dots, x_N),$$

где f — некоторая функция N переменных, y — значение свойства. Обычно вид функции f задается заранее (например, линейная, квадратичная и др.) и зависит от ряда подгоночных параметров, которые определяются по обучающей выборке соединений. Чаще всего в качестве f используется линейная функция. Получаемое уравнение называют линейной регрессионной моделью. Для расчета свойств новых соединений на ее основе сначала вычисляют значения дескрипторов $x_1, \dots, x_N$ для этих соединений, а затем и само значение свойства y по формуле $y = f(x_1, \dots, x_N)$.

Однако возможен и другой подход к решению задачи прогнозирования свойств соединений. Он основан на следующей гипотезе: «близкие» структуры имеют «близкие» свойства.^{13, 29, 65} Для оценки свойства некоторого соединения находят в выборке другое соединение, «близкое» к нему (в определенном смысле), с известным значением свойства и полагают, что рассматриваемые свойства этих соединений одинаковы (или мало отличаются). Возможно использование также нескольких «ближайших соседей» с последующим усреднением значений их свойств. Для количественной оценки степени близости двух молекул M_1 и M_2 им сопоставляют некоторые математические объекты m_1 и m_2 (например, векторы, последовательности, графы и т.д.). Затем вводят скалярную функцию $d(m_1, m_2)$, определенную на множестве введенных объектов. Вычисляя значения этой функции для фиксированного соединения M_1 и всех остальных соединений выборки, определяют соединение, наиболее «близкое» к M_1 по мере $d(m_1, m_2)$. Существует много разных способов описания структуры молекул в математических терминах. В свою очередь для каждого из них можно определить много различных мер подобия объектов. Таким образом, будут или не будут близки (подобны) две молекулы, зависит от способа их математического описания и от выбора меры подобия. Следует отметить, что нет универсальной меры подобия, одинаково «хорошей» для разных свойств и классов соединений. Как способ описания соединений, так и мера подобия

должны специально подбираться для конкретной задачи. Результат поиска «похожей» структуры зависит также и от той выборки соединений, внутри которой происходит поиск.

Сравнивая два описанных выше подхода к моделированию связи структура–свойство, можно отметить, что: 1) во втором имеются более широкие возможности для описания молекулярной структуры, так как молекулы можно описывать не только векторами из молекулярных дескрипторов $(x_1, \dots, x_N)$, но и графами, последовательностями символов, множествами и т.д.; кроме того, при использовании векторов $(x_1, \dots, x_N)$ в первом подходе обычно N не превышает числа соединений в выборке, а во втором случае число N может быть любым; 2) при построении модели в первом подходе используются свойства всех соединений обучающей выборки, поэтому результат оценки свойства нового соединения зависит от совокупности значений; во втором случае результат зависит только от свойства «близкого» соединения. Оба подхода имеют и очевидные недостатки: в каждом случае возникает проблема выбора способа описания молекулярной структуры, а также функций f и $d(m_1, m_2)$.

Многочисленные публикации по тематике, в основе которой лежит понятие молекулярного сходства, позволяют говорить о соответствующем направлении в теоретической химии. Из анализа таких публикаций видно, что они в основном отличаются друг от друга по одному или нескольким из следующих пунктов: 1) рассматриваемыми выборками соединений; 2) исследуемыми свойствами; 3) способами математического описания химических структур; 4) количественными мерами молекулярного подобия; 5) методами оценки свойств соединений на основе выбранной меры подобия. С методологической точки зрения наибольший интерес представляют пункты 3–5.

Цель данного обзора — рассмотреть наиболее распространенные методы описания и сравнения структур молекул, встречающиеся в литературе, а также методы прогнозирования свойств химических соединений на основе концепции молекулярного подобия. Определенное внимание уделено также анализу справедливости постулата «близкие структуры имеют близкие свойства» и возможным способам его формализации. Кроме того, приведен ряд примеров, иллюстрирующих достоинства и недостатки данного метода, и отмечены соответствующие проблемы.

II. Интерпретация термина «молекулярное подобие» в химии

Термин «молекулярное подобие» в химии используется в весьма широком смысле, так как молекулы характеризуются целым комплексом структурных и физико-химических свойств и могут проявлять различную биологическую активность. В работе⁶⁶ приведен перечень различных областей химии, где в той или иной форме использовалось понятие молекулярного (или атомного) подобия. В частности, определения таких фундаментальных понятий химии, как изомерия, стереоизомерия, гомологичность, периодичность в изменении свойств химических элементов и т.д., основаны на подобии определенных свойств молекул (или атомов).

Приведем несколько примеров современного использования термина «молекулярное подобие».

1. Классификация органических соединений основана в значительной степени на классической теории химического строения. Близость химического строения молекул означает определенную степень подобия в соответствующих структурных формулах и приводит обычно к относительно малым изменениям некоторых физико-химических свойств таких систем и к аналогичным методам их синтеза. Однако очевидно, что проблема оценки степени подобия молекул только

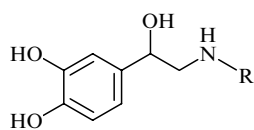
по их структурным формулам не может быть решена однозначно.

2. Молекулы, имеющие одинаковые остовы, но отличающиеся гетероатомами (например, пиррол и тиофен, бензол и пиридин), могут считаться структурно-родственными соединениями.

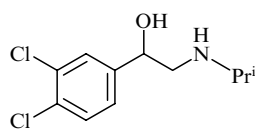
3. Молекулы гомологических рядов (например, $\text{CH}_3-(\text{CH}_2)_n-\text{CH}_3$, $\text{R}^1-\text{CH}=\text{CH}-\text{R}^2$) также считают структурно-родственными. Изменение физико-химических свойств с увеличением n в таких рядах происходит обычно по линейному закону.

4. Принцип изоlobalной аналогии, предложенный в работе⁶⁷, характеризует способность атомных группировок к химическому связыванию. Валентные возможности химических групп определяются их граничными молекулярными орбиталями (ГМО), а именно их числом, симметрией, формой и энергией. Две химические группы называются изоlobalными, если число и симметрия их ГМО совпадают, а их формы и энергии «достаточно близки». Принцип изоlobalной аналогии состоит во взаимозаменяемости в химических соединениях изоlobalных групп независимо от природы образующих их атомов. Этот принцип связывает валентные представления классической органической химии и химии элементоорганических соединений и дает возможность с единых позиций объяснить структурное многообразие химических соединений и близость их свойств.

Приведенные выше примеры показывают, что концепция химического подобия (химического родства, химического сходства) относится к фундаментальным понятиям химии. Однако само определение термина «подобие» неоднозначно и основано, как правило, на интуитивных соображениях.^{13, 68} Одна из важнейших областей применения этого понятия — прогнозирование физико-химических свойств или биологической активности соединений на основе постулата «подобные структуры имеют подобные свойства». Однако если оценивать сходство структур на основе визуального сравнения структурных формул, то этот постулат может оказаться неверным. Например, в ряду близких аналогов (норадреналина, адреналина, изопротеренола) α -адренергическая активность уменьшается, а β -адренергическая — увеличивается. А замена фенольных гидроксильных групп изопротеренола на атомы хлора превращает этот агонист β -адренергических рецепторов в антагонист — дихлоризопротеренол.⁶⁹

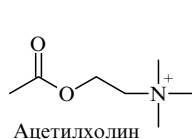


R = H (норадреналин),
Me (адреналин),
Prⁱ (изопротеренол).

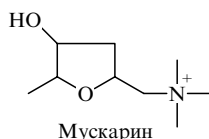


Дихлоризопротеренол

В качестве другого примера можно рассмотреть яд мускарина, выделенный из мухомора, и ацетилхолин, вырабатываемый в организме человека и животных и участвующий в передаче нервного возбуждения.⁷⁰



Ацетилхолин

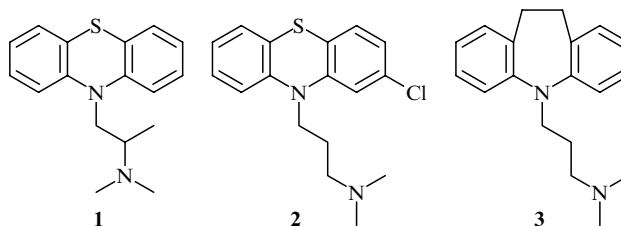


Мускарин

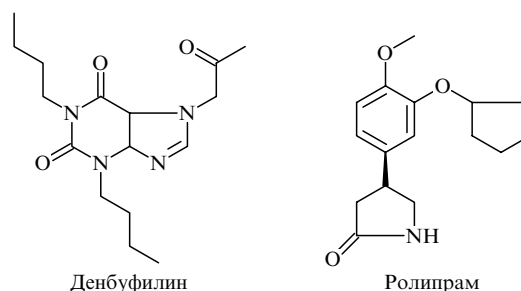
Структуры этих соединений похожи,⁷⁰ и оба они взаимодействуют с холинергическими системами человека.

Иногда сходные по структуре соединения могут проявлять различные виды биологической активности. Так,

соединение **1** проявляет антигистаминную активность, соединение **2** — нейрорептическую (антагонист дофаминовых рецепторов), а соединение **3** — антидепрессантную.⁷¹



В то же время различные по структуре соединения могут обладать одинаковой биологической активностью, как денбуфин и ролипрам, являющиеся ингибиторами фосфоэстеразы.



Денбуфин

Ролипрам

Для описания степени подобия этих соединений можно использовать такую характеристику структуры, как электростатический потенциал молекулы.^{69, 71}

Как видно из приведенных выше примеров, помимо визуального структурного подобия необходимо использовать другие способы описания молекулярной структуры и разрабатывать новые подходы к определению степени сходства молекул.

III. Теоретико-графовое представление молекул и основанные на нем описания молекулярной структуры

1. Молекулярные графы

Согласно современным представлениям, структуру молекулы можно описать с разной степенью детализации. Наиболее простой способ описания — при помощи брутто-формулы. Следующий шаг на пути усложнения этого описания — использование классической структурной формулы. Дальнейшие этапы — построение трехмерных моделей молекул и вычисление квантово-химических характеристик.

Однако на задачу математического описания структуры молекулы можно посмотреть и с другой точки зрения. Молекуле можно поставить в соответствие меченый молекулярный граф, вершины и ребра которого соответствуют атомам и связям молекулы, а метки вершин и ребер (числа или символы) кодируют атомы и связи различной химической природы. По сути классическая структурная формула является частным случаем меченого молекулярного графа, где в качестве метки вершин выбран химический символ соответствующего атома, а метки ребра — кратность соответствующей связи. В общем случае в качестве меток вершин могут быть использованы любые характеристики атомов (заряд ядра, поляризуемость, атомный вес, атомный радиус и др.), а в качестве меток ребер — любые характеристики соответствующих связей (длина; порядок, полученный из квантово-химических расчетов, и т.д.^{29, 72}). Например, были предложены⁷³ так называемые *be*-матрицы (bond and electron

matrices). Их недиагональные элементы равны порядкам связей, а диагональные — числам валентных электронов соответствующих атомов, не участвующих в связывании. Брутто-формуле можно формально сопоставить несвязный граф, т.е. граф, состоящий только из изолированных вершин (с метками) и не имеющий ребер.

Можно построить и полные молекулярные графы, в которых каждая пара вершин соединена ребром. В этом случае метка (или вес) приписывается и каждой паре несвязанных атомов. В качестве такой метки можно взять топологическое (или геометрическое) расстояние между вершинами. В последнем случае граф будет соответствовать хорошо известной матрице геометрических расстояний.

Предположим, что известны координаты атомов молекулы в трехмерном пространстве. В этом случае можно построить граф, состоящий из изолированных вершин, метки которых представляют собой последовательности трех чисел — координат атомов. В качестве метки вершины графа можно также использовать какую-либо физико-химическую характеристику соответствующего атома.

Таким образом, по способу выбора метки молекулярного графа можно разделить на три основных типа: рассчитанные на основе структуры самого графа; взятые из справочников (например, стандартные физико-химические характеристики атомов и связей); полученные из эксперимента (например, определенные методом рентгеноструктурного анализа). Возможна и комбинация этих способов.

В ряде случаев используются так называемые «простые» молекулярные графы. Вершины этих графов изображают точками, а ребра — отрезками прямых. Хотя у простого графа и нет явных меток, но считается, что метки вершин равны нулю. Если атомы связаны химической связью, то им сопоставляется ребро с меткой «1», в противном случае метка ребра равна нулю. Такие графы используются, как правило, для представления структуры углеводородных молекул.

2. Описания молекулярных структур, основанные на их теоретико-графовых представлениях

Меченый граф задается своей топологической матрицей

$$A = (a_{ij}),$$

где a_{ii} — метка i -ой вершины, a_{ij} — метка пары вершин (i, j) . (В случае простого графа матрица A называется обычно матрицей смежности.) Очевидно, что при изменении нумерации вершин графа получается, вообще говоря, другая матрица. Это является существенным недостатком описания химических структур в терминах матриц. Данная проблема привела к понятию «инвариант графа». Инвариант графа — это такое число (или функция, если метки графа — символы), вычисляемое по матрице графа A , которое не зависит от способа нумерации вершин. В теоретической химии инварианты простых графов обычно называют топологическими индексами (ТИ). Часто это определение переносится и на инварианты меченых графов, которые могут отражать не только топологию молекулы, но также элементы электронного и пространственного строения. Аналогичным образом можно определить и инвариант любого подграфа графа. Они называются локальными ТИ. Топологические индексы широко используются в теоретической органической химии для количественной характеристики молекулярной структуры при поиске соотношений структура — свойство, в компьютерном синтезе, при построении и анализе баз данных по структурам и свойствам. Наиболее известные ТИ и их приложения в химии описаны в обзорах^{63, 64, 74–76}. Популярность данного подхода к описанию молекулярной структуры связана с простотой вычисления ТИ, возможностью учитывать при их построении элементы электронного и про-

странственного строения, а также с наличием огромного количества удачных корреляций вида свойство — ТИ. Вместе с тем этот подход имеет и очевидный недостаток: он не позволяет различать разные конфигурации молекул (например, стереоизомеры) и не учитывает их конформационные особенности.

Приведем примеры некоторых наиболее популярных «именных» ТИ.^{63, 64}

Индекс Рандича χ определяется по формуле

$$\chi = \sum (v_i v_j)^{-1/2},$$

где v_i и v_j — степени i -ой и j -ой вершин графа, суммирование проводится по всем ребрам графа. Индекс χ обобщен Киром и Холлом, которые ввели определение так называемого индекса связности m -го порядка

$$^m\chi^v = \sum (\delta_i \delta_j \delta_k \dots)^{-1/2}, \quad m \geq 1, \quad \delta_i = \frac{Z_i^v - h_i}{Z_i - Z_i^v - 1},$$

где Z_i — общее число электронов, Z_i^v — число валентных электронов i -го атома (соответствующего i -ой вершине); h_i — число атомов водорода, связанных с i -ым атомом; суммирование проводится по всем цепочкам, состоящим из m ребер графа; i, j, k — номера вершин, образующих соответствующую цепочку.

Индекс Винера W определяется как сумма по всем парам вершин

$$W = 0.5 \sum d_{ij},$$

где d_{ij} — кратчайшее расстояние между i -ой и j -ой вершинами. Матрица $D = (d_{ij})$ называется матрицей топологических расстояний графа.

Индекс Хосойя Z имеет вид

$$Z = \sum p_k,$$

где p_k — число способов выбрать в графе k ребер ($k \geq 1$) так, что никакие два из них не являются смежными.

Индекс Балабана B (для ациклических графов (деревьев)) определяется по формуле

$$B = \sum r_i^2,$$

где r_i — число вершин дерева, отсекаемых на i -ом шаге процедуры обрезки деревьев. Эта процедура состоит из последовательного удаления вершин степени 1 исходного дерева и инцидентных им ребер.

При построении информационных топологических индексов⁷⁷ выбирается некоторое множество фрагментов (например, множество вершин графа), и это множество разбивается на классы эквивалентности по какому-либо признаку (например, по картинкам окружения первого порядка). Пусть n — число вершин графа, n_i — число вершин в i -ом классе эквивалентности, $i = 1, \dots, m$. Основным информационный ТИ определяется как

$$IC = - \sum \frac{n_i}{n} \log_2 \frac{n_i}{n}.$$

Он характеризует степень неоднородности множества вершин графа и принимает свои экстремальные значения для наименее однородного ($m = n$) и наиболее однородного ($m = 1$) множества вершин. Вместо множества вершин можно рассматривать множество других фрагментов.

Другой большой класс ТИ — это ТИ спектрального типа.⁶⁴ Обозначим через λ_i и $C^{(j)} = (C_1^{(j)}, \dots, C_n^{(j)})$ собствен-

ные числа и собственные векторы матрицы графа с n вершинами. По ним можно построить разнообразные ТИ, например такие как

$$J_1 = \max(\lambda_i),$$

$$J_2 = \lambda_r - \lambda_m,$$

где λ_r — наименьшее положительное, а λ_m — наибольшее отрицательное собственные числа, или

$$q_i = \sum C_i^{(k)} C_i^{(k)},$$

$$p_{ij} = \sum C_i^{(k)} C_j^{(k)}$$

(суммирование в последних двух формулах распространено на те собственные векторы, которым соответствуют положительные собственные числа). Топологические индексы спектрального типа возникают естественным образом в рамках ряда квантово-химических методов и обычно имеют достаточно ясную физико-химическую интерпретацию.

Следует отметить, что даже для одного фиксированного способа расстановки меток на молекулярном графе существует бесконечно много ТИ. Например, произведение или сумма двух ТИ приводит к новым ТИ. Поэтому в принципе невозможно описать все ТИ, которые можно построить и использовать в QSAR/QSPR-анализе. При построении каждой конкретной модели связи структура — свойство возникает проблема мотивированного выбора небольшого числа ТИ из бесконечного числа вариантов. Эта проблема связана с тем, что часто трудно дать достаточно ясную физико-химическую или структурную интерпретацию ТИ и связать их с рассматриваемым свойством. Кроме того, некоторые ТИ связаны друг с другом строгими или приближенными математическими соотношениями. Эти факты, а также здравый смысл подсказывают, что для произвольного конечного набора взвешенных графов должен существовать конечный набор ТИ $J_1, J_2, \dots, J_N$, называемый базисным. Тогда любой ТИ представляется в виде

$$\text{ТИ} = F(J_1, J_2, \dots, J_N),$$

где $F(J_1, J_2, \dots, J_N)$ — некоторая функция от N переменных, вид которой зависит от рассматриваемого топологического индекса. Строго доказано,⁷⁸ что такой базис всегда существует, причем в качестве базисных инвариантов можно взять числа вхождений в граф определенных структурных фрагментов, а в качестве F — линейную функцию. В работе⁷⁹ приведены примеры разложения по базису различных инвариантов молекулярных графов. Результаты этих работ можно рассматривать как математическое обоснование хорошо известного метода аддитивных схем, используемого для оценок физико-химических свойств соединений.^{80,81}

Предложен⁸² иной подход к проблеме поиска базисных ТИ, когда они строятся на основе специально выбранных подграфов и зависят от параметров, которые можно варьировать. В этой же работе дано определение базиса, более удобное для задач QSAR/QSPR. Разработанный подход успешно применен для оценок различных физико-химических свойств углеводородов разных классов.

Числа вхождения в структуру молекулы определенных структурных фрагментов являются частным случаем ТИ. Простейшие структурные фрагменты издавна используются в теоретической химии для оценок физико-химических свойств в рамках аддитивных схем.^{80,81} В QSAR/QSPR-исследованиях используются и более сложные фрагменты.^{62, 83–85}

Как легко заметить, практически все ТИ связаны тем или иным способом с подграфами. Фрагментные дескрипторы

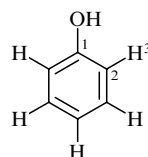
Таблица 1. Атомные пары для молекулы ацетона.

Атом	Атомные пары	Число вхождений фрагмента в структуру
1	CX ₁ –1–C.X ₃	2
2	C.X ₃ –1–O.X ₁	1
3	CX ₁ –2–O.X ₁	2
4	CX ₁ –2–CX ₁	1

играют большую роль в методах QSAR/QSPR. Рассмотрим ряд таких дескрипторов.

Пусть в некоторой молекуле цепочка из $L-1$ связей (т.е. содержащая L атомов) «соединяет» атомы i и j с кодами Y_i и Y_j . Такой цепочечный фрагмент молекулы обычно называют «атомной парой (i, j) » и кодируют символом $Y_i-(L-1)-Y_j$.⁸⁶ Код Y_i (или Y_j) содержит информацию о типе химического элемента, количестве соседних атомов (без учета атомов водорода) и числе π -электронов. В качестве примера в табл. 1 приведены «атомные пары» и числа их вхождений в структурную формулу молекулы ацетона.⁶⁹ Символы C, O (см. табл. 1) соответствуют атомам углерода и кислорода; символ X_n ($n = 1, 2, 3$) показывает, что данный атом имеет в качестве соседей n атомов (без учета атомов водорода); символ «.» означает, что атом содержит один π -электрон. Число $L-1$ в этом примере равно 1 или 2. Предложен⁸⁷ и иной метод описания атомных пар, основанный на использовании физико-химических свойств атомов.

Приведем примеры дескрипторов, обозначаемых MNA/ k , которые связаны с фрагментами молекул, являющимися окрестностями k -го порядка атомов различных типов ($k = 0, 1, \dots$).^{69,88} При построении этих дескрипторов атому приписывается символ «—», если он не входит в цикл; фрагмент нулевого порядка — это сам атом, фрагмент первого порядка — это атом с указанием (в скобках) его соседей первого порядка и т.д. В качестве примера в табл. 2 приведены фрагменты MNA/0 и MNA/1 структурной формулы фенола для трех атомов в молекуле фенола, отмеченных номерами 1, 2 и 3.⁶⁹

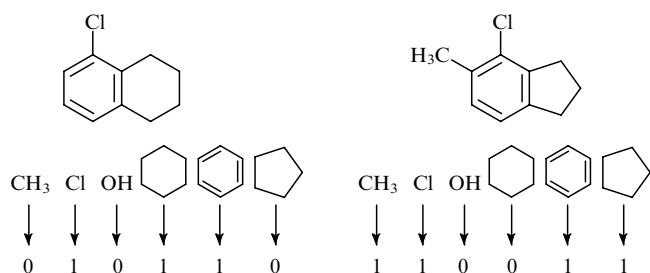


Атомно-центрированные фрагменты различного порядка неоднократно использовались и ранее.^{65,77,80,81} Они рассматриваются также в работах^{89,90}.

Обычно на основе выбранного набора фрагментов составляют векторы $X = (X_1, X_2, \dots)$, где $X_i = 0$, если i -ый фрагмент отсутствует в данной структуре, $X_i = 1$, если i -ый фрагмент присутствует в структуре. Можно взять также в качестве X_i число вхождений i -го фрагмента (например, цепочки атомов) в структуру.⁹¹ (Такой подход впервые был предложен для алканов в работе⁹².) Подобные (0,1)-векторы $X = (X_1, X_2, \dots)$ в зарубежной литературе называют «bit string». Далее приведены соответствующие примеры.⁶⁹

Таблица 2. Фрагменты MNA/0 и MNA/1 для атомов 1, 2 и 3 в молекуле фенола.

Атом	MNA/0	MNA/1
1	C	C(CC—O)
2	C	C(CC—H)
3	—H	—H(C)



Различные ТИ X_i могут быть использованы для построения векторов $X = (X_1, X_2, \dots)$. В частности, если компоненты вектора X неотрицательны, а $\sum X_i \neq 0$, то можно построить другой вектор $P = (P_1, \dots, P_k)$, где $P_i = X_i / \sum X_i$.⁹

Примеры построения векторов даны в работах^{69,93}. Для структуры с N атомами вводятся две матрицы P и C : P — прямоугольная матрица размера $N \times 3$, где i -ая строка имеет вид $(1, p'_i, q'_i)$ (p'_i, q'_i — нормированные специальным образом значения электроотрицательности атома i и его эффективного заряда), C — $(0,1)$ -матрица смежности. Затем определяется семейство квадратных матриц TM_n размера 3×3 , названных электротопологическими матрицами:

$$TM_n = \frac{P^* C^n P}{N}, \quad n = 1, 2, \dots, n_{\max},$$

где P^* — матрица, транспонированная к P ; C^n — n -ая степень матрицы C ; $n_{\max}$ — максимальная длина пути в соответствующем графе. Эти матрицы не зависят от способа нумерации вершин молекулярного графа, так что их элементы можно использовать в качестве дескрипторов, называемых электротопологическими индексами. На их основе строятся векторы, состоящие из элементов этих матриц при $n = 1, 2, \dots, n_{\max}$.

В работах^{94,95} предлагается строить векторы, характеризующие молекулярную структуру, используя для этого определенные специальным образом прямоугольные матрицы, зависящие от фрагментных дескрипторов.

Векторы можно построить также из квантово-химических⁹⁶ и любых других дескрипторов,⁹⁷ а также на основе физико-химических свойств соединений.⁹⁸

Для характеристики молекулярной структуры можно использовать молекулярный граф с указанием различных множеств фрагментов,⁹ для описания которых используется последовательность символов. Примерами могут служить линейные номенклатуры. В такой номенклатуре структурная формула кодируется строкой символов, каждый из которых, в свою очередь, определяет структурные фрагменты. Начальный этап развития линейных номенклатур связывают с именем Висвессера.⁹⁹ Например, для соединений $\text{CH}_3-\text{CH}_2-\text{O}-\text{CH}_2-\text{CH}_2-\text{CH}_3$ и $\text{H}_2\text{N}-\text{CH}_2-\text{CH}_2-\text{C}(\text{O})-\text{OH}$ коды Висвессера имеют вид (2O3) и (Z2VQ) соответственно.⁶² Здесь «O» обозначает атом кислорода, не связанный с атомами H; «Q» — гидроксильную группу; «V» — карбонильную группу; «Z» — атом азота в группе NH_2 . Числом 2 (или 3) обозначена неразветвленная цепь, состоящая из двух (или трех) групп CH_2 .

Другая линейная номенклатура носит название SMILES.^{100,101} Она используется, в частности, для поиска структурного сходства соединений. Примеры линейных кодов, получаемых в рамках языка ФКСП (фрагментарного кода суперпозиции подструктур), можно найти в работе⁸⁵. Код SLN (SYBYL Line Notation) описан в работе¹⁰²; он использован¹⁰³ для построения меры подобия белков.

Закодировать простые графы можно и при помощи наименьшего бинарного кода. Для его построения конструи-

руются все возможные ($n!$) матрицы смежности (n — число вершин графа), затем каждая из них записывается в одну строку, состоящую из 0 и 1. Такую строку можно рассматривать как запись числа в двоичной системе исчисления. В качестве кода выбирается лексикографически минимальная последовательность, которой соответствует наименьшее десятичное число.⁶⁴

Существует еще один способ построения последовательностей. Пусть задано множество, состоящее из N молекул $\{M_i\}$, $i = 1, \dots, N$, и для каждой пары M_1 и M_2 из этого множества задана некоторая симметричная функция $d(M_1, M_2) = d(M_2, M_1)$. Эту функцию рассматривают как «расстояние» между M_1 и M_2 . Тогда для каждой молекулы M_k из множества $\{M_i\}$ можно построить последовательность $\{d(M_k, M_i)\}$, $i = 1, \dots, N$, $i \neq k$ (см.^{9,29,104}), а также последовательности, состоящие из локальных инвариантов графов, например из степеней вершин графа.¹⁰⁵ В отличие от векторов, которые применяются при описании химических структур, последовательности могут иметь разные длины.

Для описания структуры молекул используют также функции. Пусть молекула задана простым графом G , d_{ij} — топологическое расстояние между i -ой и j -ой вершинами. В работах^{106,107} рассмотрена функция, называемая топологической трансформацией графа G :

$$t(s) = \frac{k(G) \sum \sum Z_i Z_j (\sin(sd_{ij}))}{sd_{ij}},$$

где s — вещественный параметр, $k(G)$ — нормировочная константа, такая, что $\int t(s)^2 ds = 1$, Z_i и Z_j — заряды ядер соответствующих атомов. Аналогичная формула была предложена¹⁰⁸ и для трехмерных структур, когда d_{ij} — геометрическое расстояние между i -ым и j -ым атомами. В последнем случае используются также функции нескольких переменных, например электронные плотности, электростатические потенциалы и др.^{13,109}

Предложен^{110–115} алгоритм для описания и анализа формы молекул на основе анализа электронных плотностей или электростатических потенциалов, определяемых в результате квантово-химических расчетов. Такая процедура начинается с исследования молекулярной поверхности. Для этого используется электронная плотность, для которой определяют поверхности уровней. Семейство поверхностей уровней рассматривается в качестве топологических объектов и характеризуется группами гомологий, называемыми группами формы и используемыми для характеристики формы молекулы и анализа степени их подобия. Такой подход удачно комбинируется с описанием стереоизомерии в терминах теории узлов и может быть легко реализован на ЭВМ, что дает возможность проводить автоматизированный не визуальный анализ степени молекулярного подобия.

На основе структурных формул, используя методы молекулярной механики или квантовой химии,⁶² можно построить трехмерные геометрические модели молекул, которые способны служить источником новых молекулярных дескрипторов. Так, например, если известны декартовы координаты (X_{i1}, X_{i2}, X_{i3}) молекулы, состоящей из N атомов с массами m_i ($i = 1, \dots, N$), то формулы

$$U_j = \frac{1}{M} \sum_{i=1}^N m_i X_{ij},$$

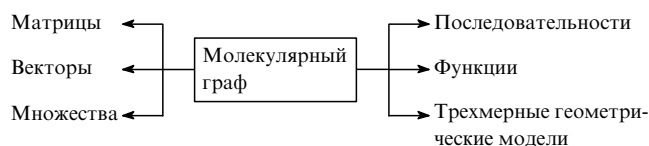
где $j = 1, 2, 3$; $M = \sum_{i=1}^N m_i$ — масса молекулы,

определяют координаты центра масс. Обозначим через $R = (r_{jk})$ — тензор инерции

$$r_{jk} = \sum_{i=1}^N m_i (X_{ij} - U_j)(X_{ik} - U_k),$$

а через X , Y , Z — собственные числа матрицы R . Эти собственные числа, а также некоторые функции от них, такие как X/Y , X/Z и др., рассматривались⁶² в качестве молекулярных геометрических дескрипторов. Некоторые авторы при описании трехмерной структуры молекулы используют такие величины, как ван-дер-ваальсов объем или площадь поверхности молекулы. Детальное описание способов вычисления этих дескрипторов приведено в работе⁶².

Способы описания структуры молекул обычно подразделяют на 2D- или 3D-методы. В первом случае не учитываются значения валентных углов и евклидовы расстояния между атомами. Поэтому методы, основанные на использовании матриц топологических расстояний молекулярного графа, относятся к 2D-методам, а найденные с их помощью инварианты — к 2D-дескрипторам. Методы, оперирующие матрицами евклидовых расстояний, являются 3D-методами, а найденные с их помощью инварианты — 3D-дескрипторами. Меры молекулярного сходства, которые рассматриваются далее, зависят, как правило, только от математического способа представления молекул (в виде графов, множеств, матриц, векторов, последовательностей и т.д.) и не зависят от той информации, которую кодируют данные математические модели. Поэтому мы не рассматриваем отдельно 2D- и 3D-методы анализа молекулярного сходства. Способы представления структуры молекул, среди которых теоретико-графовое описание является основным, а остальные способы — производными от него, можно обобщить в виде следующей схемы.



IV. Количественные меры молекулярного подобия

Количественная мера подобия двух молекул M_1 и M_2 — это скалярная функция $d(S_1, S_2)$ от двух переменных S_1 и S_2 . Ее вид зависит от того, каким образом описаны структуры S_1 и S_2 молекул M_1 и M_2 . Значение $d(S_1, S_2)$ для конкретной пары структур характеризует количественно степень их близости (или сходства). Предположим, что функция $d(S_1, S_2)$ удовлетворяет следующим условиям:

$$d(S_1, S_2) \geq 0;$$

$$d(S_1, S_2) = 0 \text{ тогда и только тогда, когда } S_1 = S_2;$$

$$d(S_1, S_2) = d(S_2, S_1) \text{ (симметрия);}$$

$$d(S_1, S_2) \leq d(S_1, S_3) + d(S_2, S_3) \text{ (неравенство треугольника).}$$

В этом случае $d(S_1, S_2)$ называется метрикой. Большинство мер, используемых для характеристики молекулярного сходства, являются метриками. В данном разделе мы опишем ряд мер подобия, разработанных для различных математических способов представления молекул, которые перечислены на схеме в конце раздела III.⁹

1. Пусть структуры S_1 и S_2 представлены n -мерными векторами $X = (x_1, x_2, \dots, x_n)$ и $Y = (y_1, y_2, \dots, y_n)$, компонентами которых служат какие-либо инварианты молекулярных графов. В литературе используются следующие меры подобия:

$$d_1(X, Y) = \sum |x_i - y_i| \text{ (} L_1 \text{ — метрика),}$$

$$d_2(X, Y) = \left(\sum |x_i - y_i|^p \right)^{1/p} \text{ (} L_p \text{ — метрика);}$$

при $p = 2$ d_2 — это расстояние между точками X и Y в евклидовом пространстве R^n (евклидово расстояние широко используется при оценках степени подобия молекул^{24, 65, 116–122}).

$$d_3(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left| \sum (x_i - \bar{x}) \sum (y_i - \bar{y}) \right|^{1/2}},$$

$$\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{y} = \frac{1}{n} \sum y_i$$

(коэффициент корреляции);

$$d_4(X, Y) = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i}$$

(коэффициент Танимото).

Коэффициент Танимото является одной из самых популярных мер подобия в исследованиях QSAR/QSPR, а также при анализе баз химических данных.^{2, 53, 122–127} При этом обычно в качестве векторов X и Y рассматриваются (0,1)-векторы, значения компонент которых зависят от наличия (1) или отсутствия (0) определенных фрагментов в структурах S_1 и S_2 . В этом случае коэффициент Танимото принимает вид

$$d_4(X, Y) = \frac{N_{xy}}{N_x + N_y - N_{xy}},$$

где N_x и N_y — количество компонент в каждом из векторов X и Y , равных 1; N_{xy} — число «общих» компонент этих векторов, равных 1.

$$d_5(X, Y) = \frac{\sum \min(x_i, y_i)}{\sum x_i^2 + \sum y_i^2 - \sum \min(x_i, y_i)} \text{ (см.}^{88}\text{),}$$

$$d_6(X, Y) = \sum (x_i - y_i) \log_2 \frac{x_i}{y_i} \quad (x_i, y_i > 0),$$

$$d_7(X, Y) = \frac{\sum x_i y_i}{\left(\sum x_i^2 \right)^{1/2} + \left(\sum y_i^2 \right)^{1/2}}$$

(косинус угла между векторами),^{91, 121, 126, 127}

$$d_8(X, Y) = \frac{\sum \min(x_i, y_i)}{0.5 \left(\sum x_i + \sum y_i \right)}.$$

Метрика $d_8(X, Y)$ была введена для случая, когда компоненты вектора X (или Y) равны числу вхождений в структуры S_1 или S_2 специального фрагмента — атомной пары (см. раздел III).⁸⁶ Метод оценки степени сходства на основе атомных пар иногда называют «методом атомных пар».

В работе⁶⁹ рассматривалась мера

$$d_9(X, Y) = \left[1 + \frac{1}{n} \sum_i (X_i - Y_i)^2 \right]^{-1},$$

которая была определена для векторов, составленных из элементов электротопологических матриц^{70, 92} (см. раздел III).

2. Пусть структуры S_1 и S_2 представлены в виде конечных множеств D_1 и D_2 . Обозначим через $|D|$ число элементов в

множестве D . В литературе используются следующие меры подобия:⁹

$$d_{10}(D_1, D_2) = |D_1| + |D_2| - 2|D_1 \cap D_2| = |D_1 \setminus D_2| + |D_2 \setminus D_1|$$

(метрика Хэмминга¹²⁸),

$$D_{11}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \parallel D_2|}$$

(коэффициент Охайи).

3. Пусть структуры S_1 и S_2 представлены мечеными графами G_1 и G_2 ; E — множество вершин графа, $|G|$ — полное число вершин и ребер графа G ; $|\text{MACS}(G_1, G_2)|$ — полное число вершин и ребер в максимальном общем подграфе G_1 и G_2 . Для графов были введены следующие меры:⁹

$$d_{12}(G_1, G_2) = |E_2 \setminus E_1| + |E_1 \setminus E_2|,$$

$$d_{13}(G_1, G_2) = |G_1| + |G_2| - 2|\text{MACS}(G_1, G_2)|,$$

$$d_{14}(G_1, G_2) = \frac{|\text{MACS}(G_1, G_2)|^2}{|G_1 \parallel G_2|}.$$

Обзор методов поиска максимального общего подграфа имеется в работах^{129, 130}. Отметим, что его можно искать также у реберных графов, порождаемых данными графами.¹³¹ В работе¹³² приведен список восьми мер подобия, основанных на максимальном общем подграфе.

Для определения меры $d_{15}(G_1, G_2)$ выберем некоторый подграф (фрагмент) H . Положим $n(G) = 1$, если H содержится в графе G , и $n(G) = 0$, если H не содержится в G . Введем меру

$$d_{15}(G_1, G_2) = |n(G_1) - n(G_2)|.$$

Очевидно, что $d_{15}(G_1, G_2) = 1$, если в одном графе содержится фрагмент H , а в другом такого фрагмента нет (графы G_1 и G_2 «непохожи») и $d_{15}(G_1, G_2) = 0$, если H либо содержится в обоих графах, либо такого фрагмента нет ни в одном из этих графов (G_1 и G_2 «похожи»). Эта мера позволяет рассортировать структуры выборки на два класса: содержащие и не содержащие фрагмент H . Использование этой меры в задачах QSAR приводит к поиску такого фрагмента H , который присутствовал бы во всех активных соединениях выборки и отсутствовал бы у неактивных соединений.^{62, 85}

4. Пусть структуры представлены в виде последовательностей l_1 и l_2 ; $|l|$ — длина последовательности; $|\text{LCS}(l_1, l_2)|$ — длина наибольшей общей последовательности в l_1 и l_2 . Для последовательностей были введены следующие меры:⁹

$$d_{16}(l_1, l_2) = |l_1| + |l_2| - 2|\text{LCS}(l_1, l_2)|,$$

$$d_{17}(l_1, l_2) = \frac{2|\text{LCS}(l_1, l_2)|^2}{|l_1 \parallel l_2|},$$

$$d_{18}(l_1, l_2) = \frac{1 - d_{17}(l_1, l_2)}{|l_1 \parallel l_2|}.$$

Кроме того, определим элементарные операции «редактирования» последовательностей: удаление элемента x , добавление элемента x , замена x на y в данной последовательности. Мету подобия двух последовательностей l_1 и l_2 можно представить как наименьшее число подобных операций, которые надо выполнить, чтобы преобразовать l_1 в l_2 .¹⁴

5. Пусть структурам S_1 и S_2 поставлены в соответствие функции f_1 и f_2 . На их основе определяются следующие меры:

$$d_{19}(f_1, f_2) = \left(\int |f_1(x) - f_2(x)|^2 dx \right)^{1/2},$$

$$d_{20}(f_1, f_2) = \int |f_1(x) - f_2(x)| dx,$$

$$d_{21}(f_1, f_2) = \frac{\int f_1 f_2 dx}{\int |f_1|^2 dx \int |f_2|^2 dx}.$$

Аналогичные формулы можно ввести и для функций от нескольких переменных.

Имеется ряд «именных» мер сходства, связанных с функциями. Рассмотрим меру,¹³³ заданную формулой

$$d(f_1, f_2) = \iint f_1(r_1) \Omega(r_1, r_2) f_2(r_2) dr_1 dr_2,$$

где f_1 и f_2 — функции электронной плотности, Ω — положительно определенный оператор. Мера $d(f_1, f_2)$ использовалась для определения других мер, например:

$$d_C(f_1, f_2) = \frac{d(f_1, f_2)}{d(f_1, f_1)^{1/2} d(f_2, f_2)^{1/2}}$$

(индекс Карбо^{134–140}),

$$d_{\text{HR}}(f_1, f_2) = \frac{2d(f_1, f_2)}{d(f_1, f_1) + d(f_2, f_2)}$$

(индекс Ходжкина – Рихардса^{109, 134–136, 141}),

$$d_P(f_1, f_2) = \frac{d(f_1, f_2)}{\max(d(f_1, f_1), d(f_2, f_2))}$$

(индекс Петке^{142, 143}).

В работе¹³³ введены также два семейства индексов подобия, зависящих от параметров x и k :

$$D_1(f_1, f_2) = \left[\frac{k(d(f_1, f_1) + d(f_2, f_2))}{2} - xd(f_1, f_2) \right]^{1/2},$$

$$D_2(f_1, f_2) = \frac{(k - x)d(f_1, f_2)}{D_1^2(f_1, f_2)}.$$

При определенных значениях k и x индексы D_1 и D_2 совпадают с индексами Карбо и Ходжкина – Рихардса. Приведенные выше меры сходства обычно называют квантово-химическими, так как функции f_i (электронная плотность, электростатический потенциал и др.) вычисляются методами квантовой химии.^{133, 135, 142, 144–171}

В работе¹⁷² предложено считать подобными молекулы, состоящие из 3 или 4 атомов, если они имеют одинаковое число электронов. Установлено, что в этом случае подобные молекулы имеют близкие значения ряда физико-химических параметров (например, теплоты атомизации). Однако для некоторых свойств (например, для потенциала ионизации) такой зависимости нет.

Рассмотрены^{173–175} различные модификации указанных выше квантово-химических мер подобия, которые приводят к новым, несимметричным мерам d'_{ij} для некоторых выборок соединений. Например, рассматривается так называемое «стохастическое преобразование» d'_{ij} меры подобия d_{ij} :

$$d'_{ij} = d_{ij}(d_{i1} + d_{i2} + \dots + d_{iN})^{-1},$$

где N — число соединений в выборке. Величина d'_{ij} , называемая мерой самоподобия, рассматривается как молекулярный дескриптор и используется для построения моделей в QSAR/QSPR-анализе.¹⁷⁶

6. Пусть структуры S_1 и S_2 представлены матрицами $A = (a_{ij})$ и $B = (b_{ij})$ одного порядка k . В этом случае для

сравнения структур S_1 и S_2 могут быть использованы меры $d_{22}(A, B) - d_{24}(A, B)$.^{9, 177, 178}

Определим меру $d_{22}(A, B)$ следующим образом:

$$d_{22}(A, B) = \frac{1}{\sqrt{k(k-1)}} \left(\sum (a_{ij} - b_{ij})^2 \right)^{1/2}.$$

Очевидно, что $d_{22}(A, B)$ зависит от способа нумерации атомов в структурах S_1 и S_2 . Однако можно определить меру подобия $d_{23}(A, B)$, не зависящую от нумерации. При фиксированной нумерации атомов в S_1 рассматривают все возможные нумерации атомов в S_2 и вычисляют все возможные значения для $d_{22}(A, B)$. Среди полученных значений выбирают наименьшее и считают его мерой сходства $d_{23}(A, B)$ структур S_1 и S_2 .^{9, 177}

При сравнении между собой изомеров, структуры которых описываются в терминах так называемых *be*-матриц,⁷³ часто используют меру $d_{24}(A, B)$, называемую «минимальным химическим расстоянием» и определяемую по формуле¹⁷⁸

$$\text{MCD}(A, B) = d_{24}(A, B) = \min \sum_{i,j} |a_{ij} - b_{p(i)p(j)}|.$$

Здесь суммирование проводится по индексам i и j ; p — всевозможные перестановки n символов, где n — порядок матрицы. При вычислении величины $\text{MCD}(A, B)$ возникают комбинаторные проблемы. В работе¹⁷⁸ предлагаются некоторые способы для их решения.

Предложен^{179, 180} способ оценки степени сходства молекул по их матрицам геометрических расстояний.

7. Рассмотрим меры подобия, определяемые для молекул, которые представлены в виде трехмерных геометрических объектов Γ_1 и Γ_2 .

Для вычисления таких мер выполняется процедура пространственного наложения трехмерных структур молекул с последующей минимизацией какой-либо функции многих переменных, зависящей от их взаимного расположения. Можно использовать, например, функцию

$$R(\Gamma_1, \Gamma_2) = \left[\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^3 \left(x_{ij}^{(1)} - x_{ij}^{(2)} \right)^2 \right]^{1/2},$$

где N — число атомов в структурах Γ_1 и Γ_2 , $x_{ij}^{(1)}$ (или $x_{ij}^{(2)}$) — координаты атомов в Γ_1 (или Γ_2). Минимум $R(\Gamma_1, \Gamma_2)$ при всевозможных вращениях и сдвигах в пространстве R^3 структуры Γ_2 при фиксированном положении Γ_1 может служить мерой сходства $d_{25}(\Gamma_1, \Gamma_2)$ структур Γ_1 и Γ_2 .

Вместо функции $R(\Gamma_1, \Gamma_2)$ для оценки степени сходства структур Γ_1 и Γ_2 можно использовать функцию

$$R'(\Gamma_1, \Gamma_2) = \left[\frac{1}{N} \sum_{i=1}^N m_i \sum_{j=1}^3 \left(x_{ij}^{(1)} - x_{ij}^{(2)} \right)^2 \right]^{1/2},$$

где m_i — масса i -го атома. Как и в предыдущем случае, определяется минимум этой функции, зависящий от относительного положения структур Γ_1 и Γ_2 . Значение этого минимума и принимается за меру сходства $d_{26}(\Gamma_1, \Gamma_2)$ структур Γ_1 и Γ_2 .

Для определения индекса подобия Мейера $S(\Gamma_1, \Gamma_2)$ (см.¹⁸¹) в пространстве R^3 строится трехмерная решетка с достаточно малыми расстояниями между ее узлами (например, 0.02 нм). В качестве модели молекулы рассматривается геометрическая фигура, образованная системой перекрывающихся ван-дер-ваальсовых сфер, центрированных на атомах. Молекулы Γ_1 и Γ_2 вкладываются в трехмерное пространство, оснащенное решеткой, и подсчитывается число узлов $T(\Gamma_1)$, $T(\Gamma_2)$ и $T(\Gamma_1, \Gamma_2)$, входящих в области пространства, ограниченные ван-дер-ваальсовыми поверхностями

этих молекул, а также в области их наложения. Индекс Мейера вычисляется по формуле¹⁸¹

$$0 \leq S(\Gamma_1, \Gamma_2) = \frac{T(\Gamma_1, \Gamma_2)}{(T(\Gamma_1)T(\Gamma_2))^{1/2}} \leq 1.$$

Наложение структур выполняют таким образом, чтобы величина $S(\Gamma_1, \Gamma_2)$ была максимальной.

В качестве меры подобия двух молекул используют также величину $d(\Gamma_1, \Gamma_2)$, определяемую формулой

$$d(\Gamma_1, \Gamma_2) = \frac{V_{\Gamma_1 \Gamma_2}}{V_{\Gamma_2}},$$

где V_{Γ_2} — объем молекулы Γ_2 ; $V_{\Gamma_1 \Gamma_2}$ — объем части пространства, соответствующего перекрыванию Γ_1 и Γ_2 .¹⁸² Отметим, что такую меру целесообразно использовать при фиксированном положении одной из структур.

В работе¹⁸³ при оценке степени подобия структур рассматриваются элементарные операции, при помощи которых можно трансформировать структуру Γ_1 в структуру Γ_2 . В качестве таких операций использованы удаление и добавление атомов, изменение меток атомов и др. Введена мера подобия структур Γ_1 и Γ_2 , зависящая от минимального числа операций, с помощью которых можно преобразовать одну молекулу в другую. Дан обзор методов наложения трехмерных структур. Еще один вариант оптимального совмещения трехмерных молекулярных структур предложен в работе¹⁸⁴, для оценки сходства молекул применены индексы Карбо и Мейера. В исследовании¹⁸⁵ для построения 3D-QSAR-моделей используются меры подобия, основанные на пространственном совмещении молекул.

Биологическая активность молекул тесно связана с их пространственным строением. Поэтому в последнее время большое внимание уделяется анализу сходства трехмерных структур.^{41, 42, 46, 186–202} Основной проблемой для 3D-анализа сходства является проблема выбора биологически активной конформации молекулы. Обычно выбирают конформацию, полученную методом рентгеноструктурного анализа для комплекса лиганда с рецептором либо расчетным путем на основе двумерной структуры с помощью специальных программ-конверторов. Ранее считалось, что конформация молекулы, соответствующая глобальному минимуму ее потенциальной энергии, является биологически активной. Однако установлено, что это не обязательно так.⁶⁹ Поэтому использование только основной конформации для анализа сходства соединений может привести к ошибочным результатам.

В большинстве 3D-баз данных хранится только одна конформация. Поэтому 3D-методы анализа сходства, вероятно, не будут иметь преимуществ перед 2D-методами без дополнительного учета гибкости молекулы. Для этого необходимо либо генерировать и хранить набор конформаций для каждой молекулы, либо проводить конформационные расчеты в процессе работы, что значительно увеличивает время анализа.⁶⁹

Суперпозиция (или наложение) двух молекул для расчета величины сходства является лимитирующим этапом, который ограничивает применение 3D-методов для больших баз данных. Разработаны способы, позволяющие повысить скорость процедуры суперпозиции.^{202, 203} В работе²⁰⁴ рассматривается 4D-мера подобия, являющаяся функцией от конформации. В исследовании²⁰⁵ для определения степени сходства белковых молекул используется их 2D-представление.

Наряду с 2D-, 3D-, 4D-представлениями используют также и 1D-представление.²⁰⁶ В этом случае молекулу изображают в виде графа-цепи с помеченными вершинами, которые соответствуют атомам молекулы, а метки вершин кодируют символ элемента, тип гибридизации и связность

соответствующих атомов. Методы сравнения таких цепочечных графов аналогичны методам сравнения последовательностей, состоящих из символов.

Проведен²⁰⁷ сравнительный анализ 2D- и 3D-методов при классификации соединений на активные и неактивные в четырех различных выборках. Обсуждается¹⁸⁶ возможность применения 2D- и 3D-методов для оценки сходства различных измеряемых и расчетных физико-химических свойств, связанных с такими видами взаимодействия, как гидрофобное, дисперсионное, электрическое, стерическое, донорно-акцепторное. Лучшие предсказания свойств были получены на основе применения 2D-фрагментных дескрипторов.

В исследовании²⁰⁸ использованы различные наборы 2D- и 3D-дескрипторов для классификации соединений в соответствии с их биологической активностью. Оказалось, что только 2D-фрагменты обеспечивают эффективное разделение веществ на активные и неактивные и поэтому предпочтительны для анализа степени сходства химических соединений.

В работе⁵⁰ приведены результаты сравнения 2D-метода атомных пар и его 3D-модификаций, использованных для поиска активных аналогов морфина, циклопана, диазепамы, апоморфина, серотонина, габоксадола, диэтилстильбэстрола в базе данных из 30 000 соединений. Показано, что трехмерный подход не превосходит 2D-метод по количеству найденных биологически активных аналогов. Вместе с тем некоторые активные соединения, «не похожие» на соответствующий образец с точки зрения топологии, были найдены только при использовании 3D-методов. В работе²⁰⁹ проведено также сравнение 2D- и 3D-дескрипторов для классификации соединений на активные и неактивные; оба метода дали одинаковые результаты. При сравнении меры подобия, задаваемой коэффициентом Танимото, и мер, связанных с наложением структур, также было показано,²¹⁰ что 2D- и 3D-методы обладают приблизительно одинаковой эффективностью.

При поиске аналогов в 3D-базах данных метод наложения структур дал⁶⁸ наилучшие результаты в неоднородных базах данных, тогда как для более однородных баз лучшим оказался подход, основанный на поиске максимального общего подграфа.

В работах^{211–216} проведено сравнение нескольких мер подобия в задаче поиска соединений в базах данных; предложен²¹⁶ метод определения молекулярного сходства на основе комбинации 22 мер подобия.

Задача аналитического описания произвольной меры подобия, заданной на некотором конечном множестве меченых графов, рассмотрена в работах^{217, 218}. Получена в аналитическом виде общая формула для такой меры, зависящая от двух матриц A_1 и A_2 . Из результатов этих работ следует, что в задачах прогнозирования свойств химических соединений при помощи методов, основанных на использовании меры подобия, меру следует подбирать оптимальным образом по обучающей выборке соединений, определяя наилучшие матрицы A_1 и A_2 . Такой подход к построению меры является универсальным и может быть автоматизирован (в противоположность обычно используемым подходам, где практикуется «угадывание» меры). Метод был реализован в работе²¹⁹ для прогнозирования липофильности ($\lg P$) соединений, где P — коэффициент распределения в системе n -октанол–вода. Достоинства такого подхода отмечены в работах^{13, 220}.

Таким образом, имеется много разнообразных мер молекулярного подобия. На их основе можно построить новые варианты мер. В работе²²¹ обсуждаются причины появления и использования в теоретической и компьютерной химии такого многообразия мер.

В заключение этого раздела отметим, что количественные меры подобия, рассматриваемые в теоретической химии, обычно подразделяются на меры сходства (similarity measures) и меры различия (dissimilarity measures). В первом случае чем больше степень подобия молекул (с интуитивной точки зрения), тем больше значение соответствующей меры сходства. Во втором случае — наоборот. Обращая меру сходства $d(M_1, M_2)$ молекул M_1 и M_2 или рассматривая величину $1 - d(M_1, M_2)$, можно получить меру различия этих молекул. По любой мере подобия $d(M_1, M_2)$ можно построить новый молекулярный дескриптор для молекулы M : $d(M) = d(M, M_0)$, где M_0 — фиксированная молекула.⁹ Варианты способов определения молекулярного «разнообразия» рассматриваются, например, в работах^{12, 222–234}.

V. Методы прогнозирования свойств химических соединений, основанные на концепции молекулярного подобия

1. Метод k ближайших соседей

Пусть задана выборка из N химических соединений $M_1, \dots, M_N$ с известными значениями некоторого свойства $y_1, \dots, y_N$. Тогда, согласно методу k ближайших соседей (k Nearest Neighbours, kNN),^{24, 69, 235, 236} свойство y_0 некоторого соединения M_0 вычисляется по формуле

$$y_0 = \frac{1}{k} \sum W(M_j) y_j,$$

где $W(M_j) = 1$, если M_j входит в число k ближайших соседей, и $W(M_j) = 0$ в противном случае. В более общем подходе величины $W(M_j)$ можно заменить на другие весовые функции; в частности, в качестве $W(M_j)$ можно взять какую-либо меру подобия $d(M, M_j)$. В работах^{235, 236} обсуждаются достоинства и недостатки метода kNN , а также рассматриваются некоторые его модификации. Одна из основных проблем в этом методе — выбор числа k . Например, при прогнозировании липофильности²³ оптимальное k выбрано из значений 1, 2, 3, 4, 5; для оценок температуры кипения и мутагенности использовано^{25, 69} значение $k = 1$, в работе²⁴ — $k = 1, \dots, 10; 15, 20, 25$. Результат применения метода kNN зависит также от меры, с помощью которой определяются ближайшие соседи. Большое значение имеет и состав выборки (может оказаться, что «хороших» соседей в данной выборке нет).

2. Оценка свойства на основе системы уравнений

В работе²¹⁹ предложен подход, основанный на оптимальном выборе меры подобия $d(M_i, M_j)$ с помощью системы соотношений типа

$$|y_i - y_j| \approx d(M_i, M_j).$$

Чтобы устранить проблему выбора числа соседей k , возникающую в методе kNN , предложено определять свойство y_0 новой структуры M_0 на основе системы уравнений вида

$$|y_0 - y_i| \approx d(M_0, M_i),$$

где $\{M_i\}$ — обучающее множество соединений. По существу такой способ основан на гипотезе, что полученная модель «сходство структур — близость свойств» справедлива и для нового соединения M_0 .

3. Методы теории распознавания образов

Исходные предпосылки теории распознавания образов (ТРО)^{62, 76, 85, 237–239} состоят в следующем. Имеется некоторое множество объектов, каждый из которых принадлежит одному из классов (образов). Основная задача ТРО состоит в выработке правил, обычно записанных в виде неравенств, позволяющих проводить разбиение исследуемых объектов на соответствующие классы. Если заранее известно, к какому классу принадлежит объект, то возникает задача обучения распознаванию образов, если нет — то встает проблема самообучения. Качество решающего правила определяется путем подсчета ошибок при классификации объектов контрольной выборки.

Обсудим геометрическую интерпретацию ТРО. Во многих задачах описание химических объектов задается при помощи векторов $X = (x_1, \dots, x_n)$, т.е. любому объекту сопоставляется точка в евклидовом пространстве R^n . Мерой близости этих точек может служить евклидово расстояние $d(X_1, X_2) = (\sum (x_{1i} - x_{2i})^2)^{1/2}$. При удачном выборе описания объектов точки, отвечающие данному классу, занимают некоторую компактную область в R^n (образуют так называемый кластер) и области, соответствующие разным классам, не пересекаются. Понятие «кластер» является основным для задачи самообучения. Его можно определить как такое множество точек $\{A\}$, в котором каждая точка ближе к другой точке этого множества, чем к какой-либо точке, не принадлежащей $\{A\}$. В простейшем случае решающее правило может быть следующим: объект, описываемый вектором X , относится к тому классу, расстояние до которого минимально.

Основная задача ТРО имеет и вероятностную интерпретацию. В этом случае решающее правило формулируется на языке теории вероятности.

Теория распознавания образов находит широкое применение при прогнозировании биологической активности соединений. В простейшем случае вещества делятся на два класса: активные и неактивные; для описания соединений используются векторы, состоящие из дескрипторов.

Метод линейной регрессии, используемый в случае количественного описания свойств, можно рассматривать как частный случай ТРО. Действительно, интервал допустимых значений свойства можно разбить на непересекающиеся интервалы, а химические соединения — на соответствующие классы. Уравнение регрессии задает решающее правило, позволяющее классифицировать соединения контрольной выборки. Однако обычно, когда говорят о методе молекулярного подобия для прогнозирования свойств химических соединений, не имеют в виду метод линейной регрессии.

4. Методы установления количественных соотношений между сходством соединений и их биологической активностью или свойством (QSiAR/QSiPR-анализ)

В данных подходах устанавливаются соотношения вида

$$y_j = \sum a_i S(i, j) + a_0,$$

где $S(i, j)$ — коэффициент сходства между i -ым и j -ым соединениями, a_i , a_0 — коэффициенты регрессии. Для выделения значимых переменных в корреляционном уравнении (т.е. выбора значений индекса i) используют метод анализа главных компонент или метод частичных наименьших квадратов. Примеры применения QSiAR/QSiPR-анализа можно найти в работах^{29, 144, 240–243}.

VI. Формализация постулата «близкие структуры имеют близкие свойства» и его анализ

Как отмечалось выше, применение концепции молекулярного подобия для прогнозирования свойств химических соединений базируется на постулате «близкие структуры имеют близкие свойства» (БСИБС), принимаемом в качестве аксиомы без доказательства.

Естественно, возникают вопросы: как корректно интерпретировать этот постулат, в какой степени он справедлив, как можно его формализовать?

Гипотеза о том, что похожие структуры имеют похожие (близкие) свойства, возникла на основе анализа многочисленных экспериментальных соотношений, связывающих свойства химических соединений с наличием и числом определенных структурных фрагментов в молекуле. Две структуры называют близкими, если они имеют ряд общих фрагментов, вклад которых в величину свойства наиболее существен, и отличаются фрагментами, которые не влияют на величину свойства. В таком случае можно ожидать, что действительно «близкие» структуры будут иметь «близкие» свойства.

В настоящее время для оценки степени молекулярного подобия часто используют довольно абстрактные и сложные меры подобия, не связанные непосредственно с рассматриваемым свойством (раздел IV). Несмотря на это, постулат БСИБС по-прежнему применяется, однако его интерпретация и степень обоснованности в этих случаях не совсем ясны.

В работе²³⁵ сообщается об исследованиях эмпирическим путем справедливости постулата БСИБС для различных частных случаев. Рассматривались 16 различных свойств (как физико-химических, так и разных видов биологической активности), 16 баз данных, 72 количественные меры подобия, основанные на векторном описании структуры молекул. Расчет свойств проводили методом k NN. Установлено, что меры, основанные на коэффициентах корреляции, наиболее удачные. Для описания структуры молекулы наиболее предпочтительно использовать вектор, компоненты которого равны числам вхождений в нее специальных структурных фрагментов.

Таким образом, справедливость постулата БСИБС существенно зависит от выбора химических дескрипторов, меры подобия, класса химических соединений, рассматриваемого свойства, метода оценки свойств.²³⁵ Поскольку применение этого постулата для решения конкретных задач теоретической органической химии может оказаться некорректным, то успехи в прогнозировании свойств химических соединений указанными выше методами носят, вообще говоря, случайный характер, и требуется теоретическое обоснование постулата БСИБС.

Для анализа этого постулата рассмотрим самый общий случай, когда химические структуры $\{M_i\}$, $i = 1, \dots, N$, представлены в виде некоторых математических объектов $\{S_i\}$, $i = 1, \dots, N$, и на парах этих объектов S_i и S_j задана некоторая симметричная функция $d(S_i, S_j) \geq 0$ (такая, что $d(S_i, S_j) = 0$ только тогда, когда $S_i = S_j$) — мера подобия соответствующих структур M_i и M_j . Пусть, кроме того, известны числа $\{y_i\}$ — значения некоторого свойства структур $\{M_i\}$, $i = 1, \dots, N$. Естественно считать мерой близости свойства величину $|y_i - y_j|$. Предположим, что заранее указаны численные критерии подобия свойств и структур, т.е. такие числа $\varepsilon \geq 0$ и $\delta \geq 0$, что если $d(S_i, S_j) \leq \delta$, то структуры

считаются близкими, и если $|y_i - y_j| \leq \varepsilon$, то свойства считаются близкими. Очевидно, что число ε всегда задается исследователем и зависит от конкретной задачи, а варьируемыми характеристиками являются $d(S_i, S_j)$ и δ . Теперь постулат БСИБС можно сформулировать так: если для любых структур M_i и M_j $d(S_i, S_j) \leq \delta$, то $|y_i - y_j| \leq \varepsilon$.

Для анализа справедливости постулата сформулируем следующее простое утверждение: для любой выборки структур $\{M_i\}$ ($i = 1, \dots, N$), любого свойства y и любой меры подобия $d(S_i, S_j) \geq 0$ ($d(S_i, S_j) = 0$ только в том случае, когда $S_i = S_j$) верны неравенства

$$ad(S_i, S_j) \leq |y_i - y_j| \leq bd(S_i, S_j),$$

где a и b — константы, зависящие от меры, свойств и выборки структур, а именно

$$a = \min_{i,j, i \neq j} \frac{|y_i - y_j|}{d(S_i, S_j)} \geq 0,$$

$$b = \max_{i,j, i \neq j} \frac{|y_i - y_j|}{d(S_i, S_j)} \geq 0.$$

Из этого утверждения следует ряд выводов.

1. Постулат будет всегда справедлив (в формулировке, указанной выше), если выбрать $\delta = \varepsilon/b$.

2. Предположим, что для данной выборки не все значения свойств близки, т.е. найдется такая пара S_i и S_j , что $|y_i - y_j| > \varepsilon$. Тогда если взять

$$\delta = \max_{i,j} d(S_i, S_j),$$

то постулат не будет справедлив на данной выборке.

3. Ранее мы считали, что ε задано. Однако если выбрать ε очень большим, то постулат будет справедлив при любых δ и $d(S_i, S_j)$.

4. Из полученных двусторонних неравенств следует, что чем меньше величина $d(S_i, S_j)$, тем меньше величина $|y_i - y_j|$, так что для «очень близких структур» их свойства также «очень близки». Этот качественный вывод, следующий из строгих математических рассуждений, по сути есть утверждение неформализованного постулата, обычно используемого в теоретической химии.

Итак, мы показали, что справедливость постулата при фиксированной мере $d(S_i, S_j)$ зависит от выбора ε и δ . Однако практически нигде при использовании концепции молекулярного подобия не вводится критерий близости свойств. Поэтому трудно судить, имеют ли структуры близкие свойства, если не введен этот критерий. Обычно в работах по анализу QSAR/QSPR приводят экспериментальные и расчетные значения свойства, но не указывают, какая ошибка прогноза считается допустимой. Поэтому читатель должен сам судить о качестве используемых методов.

В работе²¹⁹ использован другой подход к выбору меры $d(S_i, S_j)$: она подбирается по выборке структур, исходя из условий $d(S_i, S_j) \approx |y_i - y_j|$ при всех i и j . В данном случае мера представляется специальным выражением, зависящим от варьируемой квадратной симметричной матрицы A . Это выражение имеет вид скалярного произведения векторов AX и X , где X — вектор, составленный из определенных дескрипторов. Такой подход позволяет обеспечить справедливость постулата БСИБС на данной выборке при любом ε и $\delta = \varepsilon$ и представляется более эффективным, чем «угадывание» меры.

Обсудим вопрос о возможности применения постулата БСИБС для предсказания свойств соединений, не принадлежащих исходной выборке. Если этот постулат был справедлив для исходной выборки с некоторым числом δ , то для

другого множества он может не быть справедливым с этим же δ . Таким образом, очевидно, что применение его для оценки свойств других соединений может оказаться некорректным.

Для оценки δ всегда требуется некоторое «обучающее множество» соединений с известными значениями свойств. Однако при наличии такого множества можно построить стандартное регрессионное уравнение и решить проблему предсказания свойств новых соединений другими методами. Оба метода (регрессионный анализ и метод подобия) требуют выбора способа математического описания структуры молекул, молекулярных дескрипторов, вида функциональной зависимости дескрипторы — свойство или меры подобия, а также обучающей выборки соединений. Достоинством метода подобия является возможность анализировать случаи, когда зависимость структура — свойство не обязательно линейная, и вообще не использовать ее явный вид (в регрессионном анализе, как правило, используются линейные модели). Кроме того, метод подобия выглядит более привлекательным (по сравнению с регрессионным анализом), так как создается иллюзия, что он не требует проведения сложных расчетов и обучающей выборки соединений. Возможны варианты мер, когда происходит сравнение структур целиком, без вычисления каких-либо дескрипторов. Например, находят максимальный общий подграф и его характеристики (число вершин, ребер и др.) Однако при таком упрощенном подходе прогнозирование свойств соединений по методу подобия носит случайный характер.

VII. Примеры применения концепции молекулярного подобия

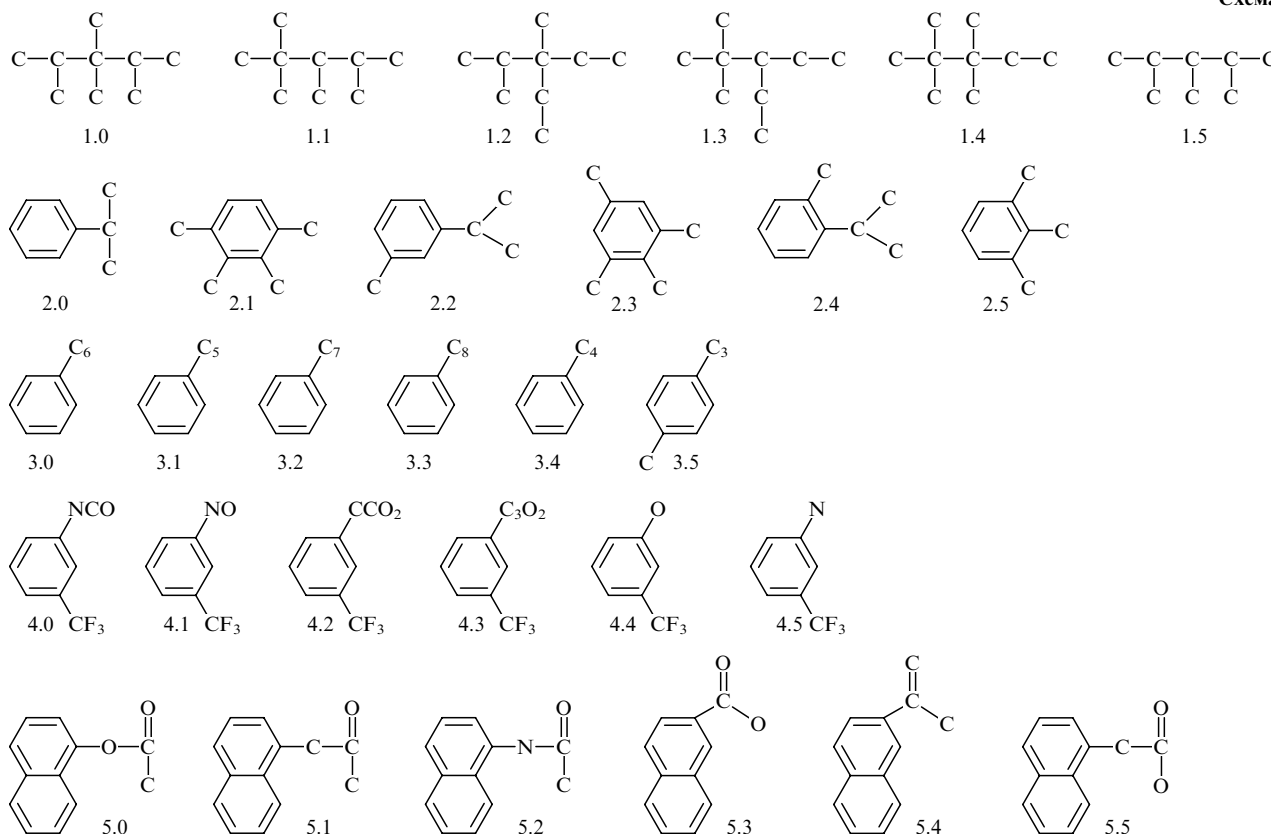
1. Определение структурного подобия в терминах топологических индексов

В работе²⁴³ для 3692 органических соединений различных классов рассчитаны значения 90 топологических индексов (в том числе индексов Рандича, Винера, ряда информационных индексов, индексов связности и др.). В терминах этих ТИ для каждой из рассмотренных структур построены 90-мерные векторы. Так как некоторые из использованных индексов были связаны между собой линейными соотношениями, при помощи метода анализа главных компонент были построены 10 новых ТИ $X_1, \dots, X_{10}$, являющихся линейными комбинациями исходных. Пусть $X^{(i)} = (X_1^{(i)}, \dots, X_{10}^{(i)})$ ($i = 1, 2$) — два таких вектора, соответствующие паре структур S_1 и S_2 . В работе²⁴⁴ определена мера подобия d структур S_1 и S_2 как евклидово расстояние между двумя точками в пространстве R^{10}

$$d(X^{(1)}, X^{(2)}) = \left(\sum (X_1^{(1)} - X_1^{(2)})^2 \right)^{1/2}$$

и исследована возможность использования этой меры для нахождения структурных аналогов некоторых соединений в указанной базе данных. Были выбраны случайным образом пять соединений и для каждого из них определены пять ближайших (по мере $d(X^{(1)}, X^{(2)})$) соседей. На схеме 1 показаны результаты такого выбора; индексы 1.0, 2.0, 3.0, 4.0, 5.0 соответствуют выбранным образцам, а индексы вида i, j ($i, j = 1, \dots, 5$) — найденным их ближайшим «соседям»; соединения приведены в порядке возрастания соответствующих значений меры $d(X^{(1)}, X^{(2)})$. Найденные соединения соответствуют интуитивным представлениям химика о подобии структур молекул. Для того чтобы найти в базе «хорошие» соединения, необходимо, по крайней мере, чтобы они там были.²⁴⁴

Схема 1



2. Зависимость результата кластеризации выборки соединений от способа описания химической структуры

В работе²⁴⁵ исследовали зависимость токсичного действия на мышей от структуры для 25 простых эфиров R^1-O-R^2 ($R^1 = \text{Me, Et, Pr}^n, \text{Pr}^i, \text{CH}_2=\text{CH}$; $R^2 = \text{Me, Et, Pr}^n, \text{Pr}^i, \text{Bu}$ и

др.). Для классификации соединений использовали дескриптор электроотпий ϵ (см.²⁴⁵) и валентно-взвешенный индекс связности ${}^m\chi^v$ при $m = 1$.

На рис. 1 показаны кластеры на плоскостях дескриптор-свойство для ϵ и χ^v . Видно, что качество кластеризации зависит от выбора дескриптора и для ϵ оно выше, чем для χ^v .

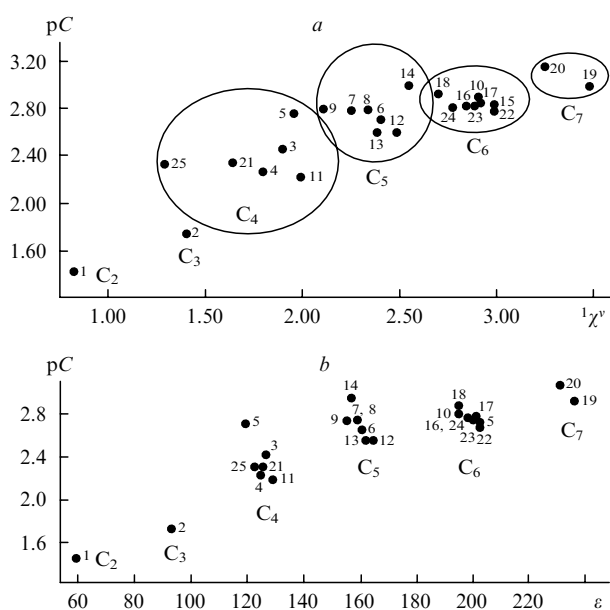
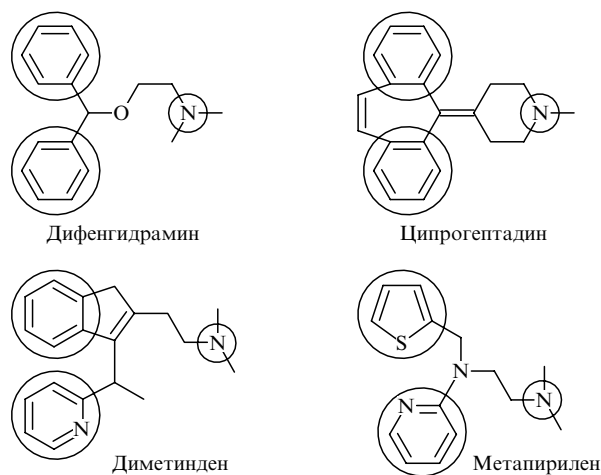


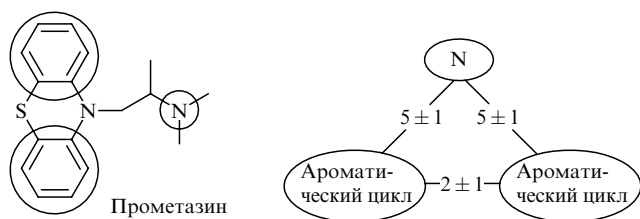
Рис. 1. Кластеризация выборки из 25 эфиров с числом атомов углерода C_2-C_7 на плоскостях дескриптор ${}^1\chi^v$ -свойство (a) и дескриптор ϵ -свойство (b).²⁴⁵

pC — степень токсичности.

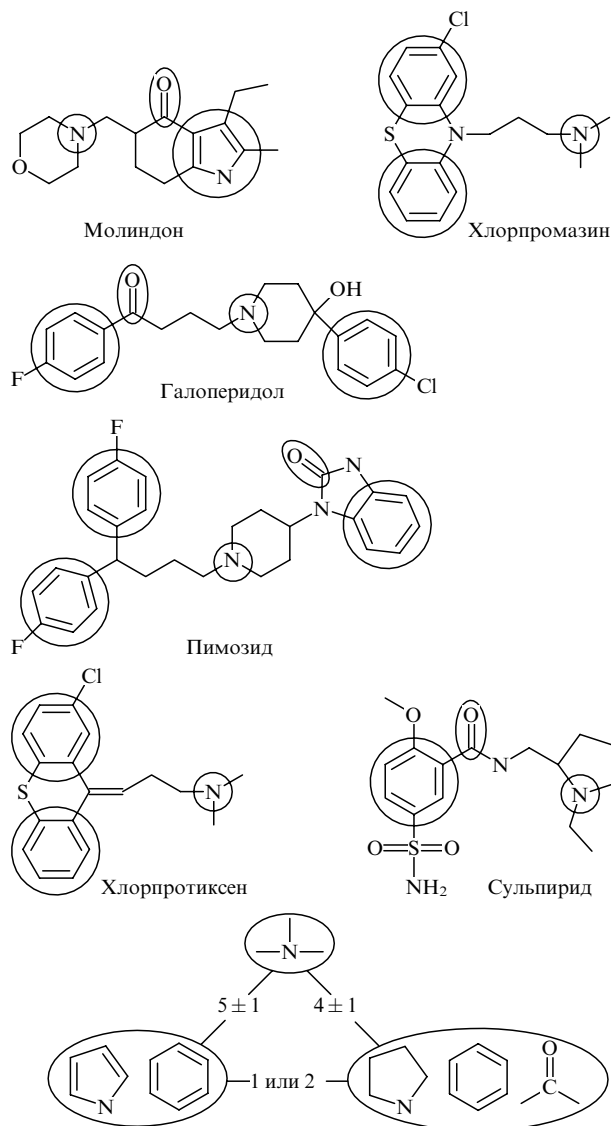
3. Поиск общих структурных фрагментов для соединений с одинаковыми свойствами

Оказалось,²⁴⁶ что структурное подобие антигистаминных препаратов (дифенгидрамина, метопирилена, ципрогептадина, диметиндена, прометазина) может быть выражено при помощи топологического треугольника, основанного на двух ароматических кольцах и третичном атоме азота.





Для антипсихотических агентов (молиндон, хлорпромазин, галоперидол, пимозид, хлорпротиксен, сульпирид) также обнаружен общий «треугольник», вершинами которого служат некоторые функциональные группы.



4. Зависимость результата прогнозирования свойств по методу *k*NN от выбора меры подобия и числа *k* ближайших соседей

В работе²³ для оценки липофильности (*lgP*) 76 соединений различных классов рассматривались две меры молекулярного подобия. Одна из них определялась по формуле

$$d' = \frac{2C}{T_1 + T_2},$$

Таблица 3. Коэффициенты корреляции между экспериментальными и теоретическими значениями *lgP* при десяти способах расчета.²³

<i>k</i>	Коэффициент корреляции (<i>r</i>)	
	для <i>d'</i>	для <i>d''</i>
1	0.774	0.788
2	0.854	0.821
3	0.869	0.814
4	0.874	0.846
5	0.852	0.845

где *T*₁, *T*₂ — число фрагментов, являющихся атомными парами в первой и второй структурах соответственно; *C* — число атомных пар, общих для двух данных молекул (см. разделы III и IV). Вторая мера подобия определялась по формуле

$$d'' = \left(\sum (X_i^{(1)} - X_i^{(2)})^2 \right)^{1/2},$$

где $X^{(j)} = (X_1^{(j)}, \dots, X_9^{(j)})$, *j* = 1, 2 — некоторые векторы из евклидова пространства *R*⁹, полученные методом анализа главных компонент из 90 различных ТИ (см. пункт 1 данного раздела).

Для каждого соединения выборки определяли пять ближайших соседей по мерам *d'* и *d''* и рассчитывали значение *lgP* по методу *k*NN пятью различными способами (при *k* = 1, 2, 3, 4, 5). Таким образом получено 10 различных вариантов оценки *lgP*. В табл. 3 приведены коэффициенты корреляции *r* между экспериментальными и расчетными значениями *lgP*. Наилучший результат наблюдается при использовании меры *d'* с *k* = 4 (*r* = 0.874). Авторы работы²³ утверждают, что оба метода дают хорошие результаты.

Для тех же данных был применен²¹⁹ другой метод анализа сходства, основанный на оптимальном подборе меры подобия *d*. Для описания структур был выбран некоторый набор структурных фрагментов. Установлено, что по ряду критериев получаемая модель является более точной, чем используемая в работе²³. Меняя множество фрагментов, можно получить более точную модель, однако в обсуждаемой работе такие исследования не проводились.

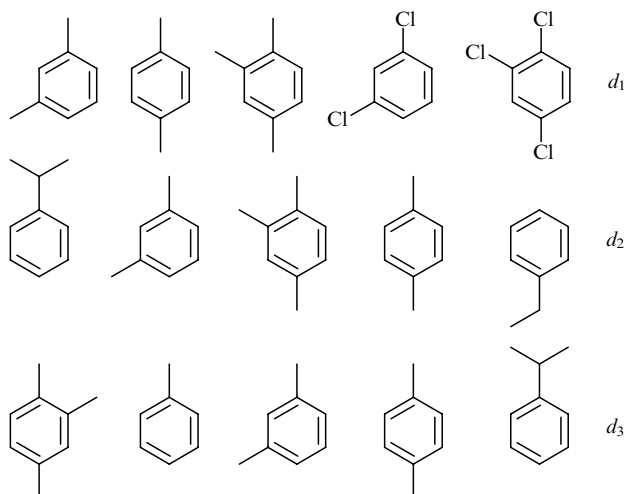
5. Выбор структурных аналогов при помощи разных мер молекулярного подобия

В работе⁴³ рассмотрены три меры молекулярного подобия *d*₁, *d*₂ и *d*₃ и в соответствии с ними найдены 5 структурных аналогов *o*-ксилола. При построении меры *d*₁ для описания молекул использовали семь физико-химических свойств (в том числе молярный объем, *lgP*, температуру кипения и др.). На основе анализа главных компонент из этих семи параметров получены пять новых, являющихся их линейными комбинациями. Кодирова структуры соответствующими пятимерными векторами, определили меру *d*₁ как обычную евклидову метрику в пространстве *R*⁵.

Мера *d*₂ построена аналогично мере *d*₁ на основании рассмотрения 81 ТИ, из которых методом анализа главных компонент построено 10 новых дескрипторов. Таким образом, *d*₂ определена как евклидова метрика в пространстве *R*¹⁰.

Мера *d*₃ основана на атомных парах и рассматривалась нами ранее.

Ниже приведены структурные формулы аналогов *o*-ксилола, выбранных с использованием указанных мер подобия (структуры расположены по убыванию численных значений мер). Результат построения «подобных» структур зависит от выбора меры. Однако имеются и примеры совпадения структур при использовании разных мер подобия.



6. Выявление сходства соединений, обладающих мускусным запахом, методом теории распознавания образов

В работе⁶² рассмотрена выборка из 300 соединений, содержащая 60 мускусных одорантов и 240 соединений, обладающих другим запахом (камфорным, цветочным, эфирным, мятным и др.). При этом мускусные одоранты относились к различным структурным типам: 23 макроциклических соединения, 19 полинитробензолов, 11 стероидов, 5 γ -бутиролактонов и два соединения других типов. Первоначально использовали 68 дескрипторов — геометрических и фрагментных; далее из них при помощи ТРО были выделены 13 наиболее существенных дескрипторов для классификации этих соединений на мускусные и немускусные одоранты (табл. 4). Для проверки прогнозирующей способности полученных дескрипторов была построена контрольная выборка, состоящая из не использовавшихся ранее соединений: 121 мускусного и 120 немускусных одорантов. В результате

Таблица 4. Основные дескрипторы, использованные для классификации соединений на мускусные и немускусные одоранты.⁶²

Дескриптор	Доля правильных классификаций
Число атомов O	84.3
Число двойных связей	82.3
Число ароматических связей	80.0
Наибольшее собственное значение тензора инерции	86.7
Наименьшее собственное значение тензора инерции	80.0
Число фрагментов	
	80.0
	80.0
	86.0
	90.3
	80.7
	80.0
	80.0
	83.0

^a — — ароматическая связь.

109 из 121 мускусных одорантов (90.1%) и 117 из 120 немускусных одорантов (97.5%) были распознаны правильно.

Проблема связи структуры соединений и их электронного строения с амбровым запахом рассматривалась в работах^{247, 248}. Авторы этих исследований использовали комбинированный подход, основанный на сочетании методов распознавания образов и квантово-химических расчетов с детальным электронно-топологическим анализом их потенциально активных участков, что позволило связать появление запаха амбры с наличием в структуре молекул так называемых амбровых треугольников.

7. Сравнение точности различных методов анализа сходства при прогнозе свойств соединений

Проведено^{69, 88} сравнение точности оценок различных свойств шестью методами анализа степени сходства свойств химических соединений, сокращенно обозначенных как PCs, TIu, TIs, AP, TEI, MNA. В качестве исходных данных рассмотрены две выборки, состоящие из 139 углеводородов с известными значениями температуры кипения и 15 нитрозаминов с известными значениями мутагенности. В трех методах сходства (PCs, TIu, TIs) степень близости определяется в терминах евклидова расстояния. В этих случаях использованы различные наборы топологических индексов, полученные из 96 ТИ путем различных процедур, приводящих к уменьшению размерности данных.²⁵ Четвертый метод (AP) основан на атомных парах и соответствующей мере подобия (см. разделы III и IV). В остальных методах (TEI и MNA) использовались векторы, составленные из электротопологических индексов и MNA-дескрипторы, основанные на атомно-центрированных фрагментах (см. меры $d_9(X, Y)$ и $d_5(X, Y)$ в разделе IV, а также раздел III). Для оценки свойств использовался метод kNN при $k = 1$.

Из данных, приведенных в табл. 5 и 6, следует, что в первом случае наилучшим является метод TIs, в котором используются топологические индексы, отобранные с помощью процедуры VARCLUS.²⁴ В набор отобранных десяти ТИ входят информационные индексы и индексы связности Кира и Холла различных порядков (см. раздел III). Этот результат можно объяснить⁶⁹ тем, что температура кипения зависит в основном от формы и степени разветвленности молекул, которые хорошо описываются топологическими индексами. В случае прогноза мутагенности (см. табл. 6) наилучший результат дают методы MNA и AP, основанные на фрагментах, а метод TEI занимает третье место. Метод TEI в данном случае превосходит методы, основанные на ТИ, так как прогноз мутагенности требует дополнительной информации о физико-химических характеристиках молекул, и именно такая информация содержится в электротопологических индексах.⁶⁹

Таблица 5. Сравнение эффективности шести методов анализа степени сходства для предсказания температуры кипения углеводородов.⁶⁹

Метод	r	s
TIs	0.993	19.34
AP	0.986	26.26
MNA	0.985	27.89
TIu	0.983	29.36
PCs	0.966	41.29
TEI	0.959	45.37

Примечание. Здесь и в табл. 6 r — коэффициент корреляции, s — среднее квадратичное отклонение.

Таблица 6. Сравнение эффективности шести методов анализа степени сходства для предсказания мутагенности нитрозаминов.⁶⁹

Метод	<i>r</i>	<i>s</i>
MNA	0.945	1.25
AP	0.944	1.26
TEI	0.931	1.43
TIu	0.923	1.47
PCs	0.830	2.33
TIs	0.740	2.67

8. Построение QSiAR-моделей

В работе¹⁰⁴ построена QSiAR-модель для сродства к σ -рецептору (pIC_{50}) четырнадцати *N*-замещенных 3-(3-гидроксифенил)пиперидинов (табл. 7). В качестве меры сходства структур А и В использовался индекс Мейера S_{AB} (см. раздел IV), где А — это исходное (незамещенное) соединение, В — все остальные соединения выборки. При этом рассматривались *R*- и *S*-энантиомеры соответствующих соединений. Построены две модели, каждая по набору семи соединений:

для *R*-изомеров ($r^2 = 0.750$)

$$pIC_{50} = -0.04 S_{AB} + 2.29,$$

для *S*-изомеров ($r^2 = 0.974$)

$$pIC_{50} = -10.21 S_{AB} + 9.30.$$

9. Исследование эффективности методов анализа сходства при прогнозе биологической активности

Для оценки эффективности применения методов анализа степени подобия при прогнозе биологической активности использована⁶⁹ процедура скользящего контроля. На каждом ее шаге из заданной базы данных поочередно исключалось одно из соединений (с номером *k*) и оценивалась вероятность $P(k)$ его отнесения к классу активных веществ:

$$P(k) = \frac{A_k}{A_k + B_k},$$

$$A_k = (n-1)^{-1} \sum S_i^1,$$

Таблица 7. Значения сродства к σ -рецептору (pIC_{50}) и индекса Мейера (S_{AB}) для ряда *N*-замещенных 3-(3-гидроксифенил)пиперидинов.¹⁰⁴

Соединение	S_{AB}	pIC_{50} (см. ^a)
заместитель	конфигурация	
H	<i>R</i>	1.000
	<i>S</i>	1.000
Me	<i>R</i>	0.952
	<i>S</i>	0.950
Et	<i>R</i>	0.917
	<i>S</i>	0.912
Pr ⁿ	<i>R</i>	0.881
	<i>S</i>	0.877
Pr ⁱ	<i>R</i>	0.884
	<i>S</i>	0.877
Bu ⁿ	<i>R</i>	0.849
	<i>S</i>	0.845
EtC ₆ H ₄	<i>R</i>	0.792
	<i>S</i>	0.786

^a IC_{50} выражена в мкмоль $\cdot$ л⁻¹.

$$B_k = (m-1)^{-1} \sum S_j^0,$$

где S_i^1 и S_j^0 — значения меры сходства *k*-го соединения с *i*-ым активным и *j*-ым неактивным соединениями соответственно; *n* и *m* — число активных и неактивных соединений в обучающей выборке соответственно. В выражении для A_k суммирование происходит по всем значениям *i* от 1 до *n* (кроме *i* = *k*). Аналогично в формуле для B_k суммирование происходит по всем значениям *j* от 1 до *m* (кроме *k* = *j*).

Пусть для каждой пары соединений M_i и M_j , где M_i — активное, M_j — неактивное соединение (*i* = 1, ..., *n*; *j* = 1, ..., *m*), вычислены значения вероятностей $P(i)$ и $P(j)$, а *N* — число пар, для которых $P(i) > P(j)$. Для оценки качества прогноза использовалась величина $IAP = N/mn$ ($0 \leq IAP \leq 1$). Для тестирования методов анализа степени подобия были отобраны 17 124 соединения из базы данных MDDR 96.2 (MDL Information System) с известной биологической активностью 16 видов. Исследованы два варианта метода подобия (с использованием в качестве меры сходства коэффициента Танимото), основанные на электротопологических индексах (TEI) и на подструктурных дескрипторах (MNA). Точность предсказания в первом случае в среднем составляет 86%, а во втором — 96%.

10. Выбор оптимального числа *k* в методе *k*NN

Эффективность кластеризации молекул в различных базах химических данных, основанная на использовании метода *k* ближайших соседей, зависит от выбора параметра *k*.

В работе²⁴⁹ исследована температура кипения для 267 галогеналканов, содержащих от 1 до 4 атомов углерода. Молекулярное подобие (различие) оценивалось в терминах евклидовой метрики в восьмимерном пространстве топологических дескрипторов, построенных методом выделения главных компонент из множества 59 ТИ. Коэффициенты корреляции между экспериментальными и расчетными значениями лежат в интервале от 0.85 до 0.94. Значения *k* выбраны равными 1–10, 15, 20, 25. Показано, что наилучшее стандартное отклонение соответствует *k* = 5 и составляет 23.7°C. Аналогичные результаты получены в работах^{23, 250}.

11. Сравнение эффективности 2D- и 3D-дескрипторов для оценки свойств соединений

На основе базы данных Medchem 94b созданы¹⁸⁶ три новые базы данных: 1) 8651 соединение с известными значениями $\lg P_{\text{октанол/вода}}$; 2) 762 соединения с известными значениями $\lg P_{\text{циклогексан/вода}}$; 3) 8416 соединений с известными значениями pK_a . Для кодирования структур использовали как 2D-, так и 3D-дескрипторы (четыре варианта). В частности, рассматривались векторы, компоненты которых равны числам вхождения в структуры фиксированных небольших фрагментов (в одном случае 153 фрагмента, известных под названием MACCS, в другом случае — 960 фрагментов, известных под названием SSKEYS). Для оценки свойств соединений предварительно проводилась кластеризация соответствующих баз разными методами с использованием мер подобия. Свойство соединения рассчитывали как среднее значение свойств всех соединений кластера, в который оно попало. Для оценки качества прогноза проводили процедуру скользящего контроля. Для этого поочередно из базы данных исключали соединение; для него рассчитывали величину свойства; затем экспериментальные и рассчитанные значения свойств сравнивали, вычисляя коэффициент их корреляции и среднеквадратичное отклонение в соответствующей регрессии. Установлено, что наилучшими во всех случаях являются 2D-дескрипторы MACCS. Проведено также сравнение

результатов расчета $\lg P$ для 7054 структур (из 8651) указанными выше методами и при помощи специальной программы CLOGP. Для методов молекулярного подобия $r = 0.95$, среднеквадратичное отклонение $rmse = 0.60$, что приближается по точности к результатам расчетов по известной программе CLOGP ($r = 0.96$, $rmse = 0.49$).

VIII. Заключение

Концепция молекулярного подобия («подобные соединения имеют подобные свойства») очень популярна в теоретической и экспериментальной химии. Одна из областей ее применения, рассмотренная в данном обзоре, — это прогнозирование свойств химических соединений на основании понятия «молекулярное подобие». В связи с широким внедрением компьютеров в химические исследования возникает необходимость его формализации и разработки методов количественного описания степени молекулярного сходства. В настоящее время имеется множество различных мер молекулярного подобия, которые, как правило, оперируют только со структурами и не связаны с какими-либо физико-химическими свойствами молекул или с биологической активностью. Однако меры можно подбирать оптимальным образом в зависимости от конкретной задачи.

Важным вопросом является формализация и обоснование постулата «близкие структуры имеют близкие свойства». Как показывают исследования, справедливость этого утверждения зависит от класса соединений, рассматриваемого свойства, выбранной меры подобия и т.д. В данном обзоре впервые обсуждается вопрос о его формализации. Показано, что при определенных условиях он всегда будет справедлив, и приведено качественное обоснование неформализованного постулата.

На основании анализа литературных данных можно сделать вывод о том, что метод молекулярного подобия, используемый для прогноза свойств химических соединений, достаточно часто дает удовлетворительные результаты. К его достоинствам относится, в частности, то, что не требуется знания вида функциональной зависимости структура — свойство. Кроме того, для определения некоторых мер подобия используется структурная формула (молекулярный граф) целиком и не нужно вычислять какие-либо молекулярные дескрипторы. В качестве примера, иллюстрирующего полезность концепции подобия в химии, можно привести историю открытия фуллеренов: эти необычные формы углерода были впервые получены в 1985 г.,²⁵¹ в то время как их существование было предсказано приблизительно на 15 лет раньше^{252, 253} на основании предположения о подобию электронных свойств плоских поликонденсированных бензоидных углеводородов и сферических полиэдрических полициклических сопряженных молекул. Это позволило эффективно использовать для фуллеренов математические модели, основанные на теории резонанса и методе Хюккеля.

В то же время следует отметить, что при реализации метода подобия возникает ряд проблем. К ним относятся выбор способа математического описания структуры молекулы, выбор меры молекулярного подобия и выбор способа оценки свойств соединений на основе меры подобия. Анализ литературных данных показывает, что для решения этих проблем нет четких, обоснованных методов и выбор происходит субъективно. Вероятно, здесь уместно вспомнить замечание, сделанное Р.Фейнманом в Нобелевской лекции, что эмпирический способ открытия закономерностей, основанный на интуиции, имеет право на существование.²⁵⁴

Литература

1. S.R.Heller. *J. Chem. Inf. Comput. Sci.*, **32**, 578 (1992)
2. D.H.Rouvray. *J. Chem. Inf. Comput. Sci.*, **34**, 446 (1994)
3. C.Jochum, M.G.Hicks, J.Sunkel. *J. Chem. Inf. Comput. Sci.*, **32**, 577 (1992)
4. *Advances in Molecular Similarity. Vol. 2.* (Eds R.Carbo-Dorca, P.G.Mezey). JAI Press, London, 1998
5. *Molecular Similarity and Reactivity: from Chemical to Phenomenological Approaches.* (Ed. R.Carbo). Kluwer Academic, Dordrecht, 1995
6. *Molecular Similarity in Drug Design.* (Ed. P.M.Dean). Blackie Academic and Professional, London, 1995
7. M.A.Johnson, M.Maggiora, M.S.Lajiness, J.B.Moon, J.D.Petke, D.C.Roher. In *Advanced Computer-Assisted Techniques in Drug Discovery.* (Ed. H.van de Waterbeemd). VCH, New York, 1994. P. 89
8. *Computer-Assisted Lead Finding and Optimization.* (Eds H.van de Waterbeemd, B.Testa, G.Folkers). VHCA, Basel; Wiley-VCH, Weinheim, 1997
9. M.A.Johnson. *J. Math. Chem.*, **3**, 117 (1989)
10. *Concepts and Applications of Molecular Similarity.* (Eds M.A.Johnson, G.M.Maggiora). Wiley, New York, 1990
11. V.Monev. *MATCH — Commun. Math. Comput. Chem.*, (51), 7 (2004)
12. Y.C.Martin. *J. Combinat. Chem.*, **3**, 231 (2001)
13. N.Nikolova, J.Jaworska. *QSAR Combinat. Sci.*, **22**, 1006 (2003)
14. M.Randić. In *Mathematical Methods in Contemporary Chemistry.* (Ed. S.I.Kuchanov). Gordon and Breach, Amsterdam, 1996. P. 1
15. M.Johnson, S.Basak, G.Maggiora. *Math. Modell. Sci. Techn.*, **11**, 630 (1988)
16. P.Willett. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press, Letchworth, 1987
17. G.Grethe, T.E.Moock. *J. Chem. Inf. Comput. Sci.*, **30**, 511 (1990)
18. W.A.Warr. *Persp. Drug Discov. Design*, **7–8**, 115 (1997)
19. O.A.Raevsky, S.V.Trepalin, H.P.Trepalina, V.A.Gerasimenko, O.E.Raevskaja. *J. Chem. Inf. Comput. Sci.*, **42**, 540 (2002)
20. D.Horvath, C.Jeandenans. *J. Chem. Inf. Comput. Sci.*, **43**, 680 (2003)
21. D.Horvath, C.Jeandenans. *J. Chem. Inf. Comput. Sci.*, **43**, 691 (2003)
22. S.C.Basak, S.Bertelsen, G.D.Grunwald. *J. Chem. Inf. Comput. Sci.*, **34**, 270 (1994)
23. S.C.Basak, G.D.Grunwald. *New J. Chem.*, **19**, 231 (1995)
24. S.C.Basak, G.D.Grunwald. *J. Chem. Inf. Comput. Sci.*, **35**, 366 (1995)
25. S.C.Basak, G.D.Grunwald. *SAR QSAR Environ. Res.*, **2**, 289 (1994)
26. A.C.Good, S.J.Peterson, W.G.Richards. *J. Med. Chem.*, **36**, 2929 (1993)
27. A.C.Good. In *Molecular Similarity in Drug Design.* (Ed. P.M.Dean). Blackie Academic and Professional, London, 1995. P. 24
28. H.Kubinyi. In *Computer-Assisted Lead Finding and Optimization.* (Eds H.van de Waterbeemd, B.Testa, G.Folkers). VHCA, Basel; Wiley-VCH, Weinheim, 1997. P. 7
29. O.Ivanciuc. *Rev. Roum. Chim.*, **46**, 787 (2001)
30. B.Bertoša, B.Kojić-Prodić, R.C.Wade, M.Ramek, S.Piperaki, A.Tsantili-Kakoulidou, S.Tomić. *J. Chem. Inf. Comput. Sci.*, **43**, 1532 (2003)
31. R.D.Cramer, R.J.Jilek, S.Guessregen, S.J.Clark, B.Wendt, R.D.Clark. *J. Med. Chem.*, **47**, 6777 (2004)
32. R.P.Sheridan, B.P.Feuston, V.N.Maiorov, S.K.Kearsley. *J. Chem. Inf. Comput. Sci.*, **44**, 1912 (2004)
33. D.Pan, M.Iyer, J.Z.Liu, Y.Li, A.J.Hopfinger. *J. Chem. Inf. Comput. Sci.*, **44**, 2083 (2004)
34. M.G.Hicks. *J. Chem. Inf. Comput. Sci.*, **32**, 631 (1992)
35. J.Gasteiger, W.D.Ihlenfeldt, R.Fick, J.R.Rose. *J. Chem. Inf. Comput. Sci.*, **32**, 700 (1992)
36. A.J.Lawson. *J. Chem. Inf. Comput. Sci.*, **32**, 675 (1992)
37. R.Ponec. *J. Chem. Inf. Comput. Sci.*, **36**, 554 (1996)
38. R.Ponec, M.Strnad. *J. Chem. Inf. Comput. Sci.*, **32**, 693 (1992)

39. R.Ponec, M.Strnad. *Collect. Czech. Chem. Commun.*, **55**, 2583 (1990)
40. G.Sello. *J. Chem. Inf. Comput. Sci.*, **32**, 713 (1992)
41. P.J.Artymiuk, P.A.Bath, H.M.Grindley, C.A.Pepperrell, A.R.Poirrette, D.W.Rice, D.A.Thorner, D.J.Wild, P.Willett, F.H.Allen, R.Taylor. *J. Chem. Inf. Comput. Sci.*, **32**, 617 (1992)
42. P.A.Bath, A.R.Poirrette, P.Willett, F.H.Allen. *J. Chem. Inf. Comput. Sci.*, **34**, 141 (1994)
43. S.C.Basak, G.D.Grunwald. *Math. Modell. Sci. Comput.*, **2**, 735 (1993)
44. G.M.Downs, P.Willett. *Rev. Comput. Chem.*, **7**, 1 (1996)
45. W.Fisanick, A.H.Lipcus, A.Rusino III. *J. Chem. Inf. Comput. Sci.*, **34**, 130 (1994)
46. C.M.R.Ginn, D.B.Turner, P.Willett, A.M.Ferguson, T.W.Heritage. *J. Chem. Inf. Comput. Sci.*, **37**, 23 (1997)
47. M.S.Lajiness. In *Structure – Property Correlations in Drug Research*. (Ed. H.van de Waterbeemd). Academic Press, New York, 1996. P. 179
48. C.A.Pepperrell, A.R.Poirrette, P.Willett, R.Taylor. *Pestic. Sci.*, **33**, 97 (1991)
49. N.C.Perry, V.J.van Geerestein. *J. Chem. Inf. Comput. Sci.*, **32**, 607 (1992)
50. R.P.Sheridan, M.D.Miller, D.J.Underwood, S.K.Kearsley. *J. Chem. Inf. Comput. Sci.*, **36**, 128 (1996)
51. D.B.Turner, P.Willett, A.M.Ferguson, T.W.Heritage. *SAR QSAR Environ. Res.*, **3**, 101 (1995)
52. J.M.Barnard, G.M.Downs. *J. Chem. Inf. Comput. Sci.*, **32**, 644 (1992)
53. R.D.Clark. *J. Chem. Inf. Comput. Sci.*, **37**, 1181 (1997)
54. G.M.Downs, P.Willett. In *Applied Multivariate Analysis in SAR and Environmental Studies*. (Eds J.Devillers, W.Karcher). Kluwer Academic, Dordrecht, 1991. P. 247
55. J.D.Holliday, S.S.Ranadhe, P.Willett. *Quantum Struct. – Act. Relat.*, **14**, 501 (1995)
56. M.Johnson, M.Lajiness, G.Maggior. In *QSAR: Quantitative Structure – Activity Relationships in Drug Design*. (Ed. J.L.Fauchere). Alan R. Liss Inc., New York, 1989. P. 167
57. J.W.Godden, J.R.Furr, L.Xue, F.L.Stahura, J.Bajorath. *J. Chem. Inf. Comput. Sci.*, **44**, 21 (2004)
58. L.Xue, J.W.Godden, F.L.Stahura, J.Bajorath. *J. Chem. Inf. Comput. Sci.*, **44**, 1275 (2004)
59. L.Xue, F.L.Stahura, J.Bajorath. *J. Chem. Inf. Comput. Sci.*, **44**, 2032 (2004)
60. V.J.Gillet, P.Willett, J.Bradshaw. *J. Chem. Inf. Comput. Sci.*, **37**, 731 (1997)
61. P.Willett. *Persp. Drug Discov. Design*, **7–8**, 1 (1997)
62. Э.Стьюпер, В.Брюггер, П.Джурс. *Машинный анализ связи химической структуры и биологической активности*. Мир, Москва, 1982
63. О.А.Раевский. *Успехи химии*, **68**, 555 (1999)
64. М.И.Станкевич, И.В.Станкевич, Н.С.Зефирова. *Успехи химии*, **57**, 337 (1988)
65. L.H.Hall, L.B.Kier, B.B.Brown. *J. Chem. Inf. Comput. Sci.*, **35**, 1074 (1995)
66. D.H.Rouvray. *J. Chem. Inf. Comput. Sci.*, **32**, 580 (1992)
67. R.Hoffmann. *Angew. Chem., Int. Ed. Engl.*, **21**, 711 (1982)
68. C.Pepperrell. In *Computers and Chemical Structure Information Series. Vol. 3*. (Ed. D.Bawden). Research Studies Press Ltd., Taunton; Somerset, England, 1994
69. Ю.В.Бородин. Дис. канд. биол. наук. НИИ Биомедицинской химии РАМН, Москва, 1998
70. С.Н.Голиков. *Яды и противоядия*. Знание, Москва, 1968
71. F.E.Blaney, P.Finn, R.Phippen, M.Wyatt. *J. Mol. Graphics*, **11**, 98 (1993)
72. O.Ivanciu, S.L.Taraviras, D.Cabrol-Bass. *J. Chem. Inf. Comput. Sci.*, **40**, 126 (2000)
73. I.Ugi, M.Wochner, E.Fontain, J.Bauer, B.Gruber, R.Karl. In *Concepts and Applications of Molecular Similarity*. (Eds M.A.Johnson, G.M.Maggiora). Wiley, New York, 1990. P. 239
74. Д.Рувре. В кн. *Химические приложения топологии и теории графов*. (Под ред. Р.Кинга). Мир, Москва, 1987. С. 181
75. *Topological Indices and Related Descriptors in QSAR and QSPR*. (Eds J.Devillers, A.T.Balaban). Gordon and Breach, Amsterdam, 1999. P. 811
76. П.Джурс, Т.Айзенауэр. *Распознавание образов в химии*. Мир, Москва, 1977
77. В.Магнусон, Д.Харрис, С.Бейсак. В кн. *Химические приложения топологии и теории графов*. (Под ред. Р.Кинга). Мир, Москва, 1987. С. 206
78. I.I.Baskin, M.I.Skvortsova, I.V.Stankevich, N.S.Zefirov. *J. Chem. Inf. Comput. Sci.*, **35**, 527 (1995)
79. M.I.Skvortsova, I.I.Baskin, L.A.Skvortsov, V.A.Palyulin, N.S.Zefirov, I.V.Stankevich. *J. Mol. Struct. (THEOCHEM)*, **466**, 211 (1999)
80. Н.Ф.Степанов, М.Е.Ерлыкина, Т.Г.Филиппов. *Методы линейной алгебры в физической химии*. Химия, Москва, 1976
81. Yu.G.Papulov, V.M.Smoljakov. In *Mathematical Methods in Contemporary Chemistry*. (Ed. S.I.Kuchanov). Gordon and Breach, Amsterdam, 1996. P. 143
82. М.И.Скворцова, К.С.Федяев, В.А.Палюлин, Н.С.Зефирова. *Изв. АН. Сер. хим.*, 1527 (2004)
83. P.Willett. *Curr. Opin. Biotechnol.*, **11**, 8588 (2000)
84. N.S.Zefirov, V.A.Palyulin. *J. Chem. Inf. Comput. Sci.*, **42**, 1112 (2002)
85. Н.Ф.Розенблит, В.Е.Голендер. *Логико-комбинаторные методы в конструировании лекарств*. Зинатне, Рига, 1983
86. R.E.Carhart, D.H.Smith, R.Venkataraghavan. *J. Chem. Inf. Comput. Sci.*, **25**, 64 (1985)
87. S.K.Kearsley, S.Sallamack, E.M.Fluder, J.D.Andose, R.T.Mosley, R.P.Sheridan. *J. Chem. Inf. Comput. Sci.*, **36**, 118 (1996)
88. D.Filimonov, V.Poroikov, Yu.Borodina, T.Gloriozova. *J. Chem. Inf. Comput. Sci.*, **39**, 666 (1999)
89. A.Bender, H.Y.Mussa, R.C.Glen, S.Reiling. *J. Chem. Inf. Comput. Sci.*, **44**, 1708 (2004)
90. A.Bender, H.Y.Mussa, R.C.Glen, S.Reiling. *J. Chem. Inf. Comput. Sci.*, **44**, 170 (2004)
91. P.Willett, V.Winterman. *J. Chem. Inf. Comput. Sci.*, **26**, 36 (1986)
92. Е.А.Смоленский. *Журн. физ. химии*, **38**, 1288 (1964)
93. Yu.V.Borodina, D.A.Filimonov, V.V.Poroikov. *Quantum Struct. – Act. Relat.*, **17**, 459 (1998)
94. R.D.Hull, S.B.Singh, R.B.Nachbar, R.P.Sheridan, S.K.Kearsley, E.M.Fluder. *J. Med. Chem.*, **44**, 1177 (2001)
95. R.D.Hull, S.B.Singh, R.B.Nachbar, R.P.Sheridan, S.K.Kearsley, E.M.Fluder. *J. Med. Chem.*, **44**, 1185 (2001)
96. M.Karelson, V.S.Lobanov, A.R.Katritzky. *Chem. Rev.*, **96**, 1027 (1996)
97. L.Xue, J.W.Godden, F.L.Stahura, J.Bajorath. *J. Chem. Inf. Comput. Sci.*, **43**, 1218 (2003)
98. G.M.Downs, P.Willett, W.Fisanick. *J. Chem. Inf. Comput. Sci.*, **34**, 1094 (1994)
99. W.J.Wisswesser. *The Wisswesser Line-Formula Chemical Notation*. McGraw-Hill, New York, 1968
100. A.Drefahl, M.Reinhard. *J. Chem. Inf. Comput. Sci.*, **33**, 886 (1993)
101. D.Weiniger. *J. Chem. Inf. Comput. Sci.*, **30**, 237 (1990)
102. S.Ash, M.A.Cline, R.W.Homer, T.Hurst, G.B.Smith. *J. Chem. Inf. Comput. Sci.*, **37**, 71 (1997)
103. A.Schuffenhauer, P.Floersheim, P.Acklin, E.Jacoby. *J. Chem. Inf. Comput. Sci.*, **43**, 391 (2003)
104. A.Sery-Levy, S.West, W.G.Richards. *J. Med. Chem.*, **37**, 1727 (1994)
105. Ф.Харари. *Теория графов*. Мир, Москва, 1973
106. Z.Gabányi, P.Surján, G.Náray-Szabó. *Eur. J. Med. Chem.*, **17**, 307 (1982)
107. L.J.Soltzberg, C.L.Wilkins. *J. Am. Chem. Soc.*, **99**, 439 (1977)
108. J.H.Schuur, P.Selzer, J.Gasteiger. *J. Chem. Inf. Comput. Sci.*, **36**, 334 (1996)
109. E.E.Hodgkin, W.G.Richards. *J. Chem. Soc., Chem. Commun.*, 1342 (1986)
110. P.G.Mezey. *J. Chem. Inf. Comput. Sci.*, **32**, 650 (1992)

111. P.G.Mezey. *J. Chem. Inf. Comput. Sci.*, **34**, 244 (1994)
112. P.G.Mezey. *Shape in Chemistry*. VCH, New York, 1993
113. P.G.Mezey, Z.Zimpel, P.Warburton, P.D.Walker, D.G.Irvine, S.L.Dixon, B.Greenberg. *J. Chem. Inf. Comput. Sci.*, **36**, 602 (1996)
114. P.G.Mezey. In *Encyclopedia of Computational Chemistry. Vol. 4*. (Eds P.R.Schleyer, N.L.Allinger, T.Clark, J.Gastaiger, P.R.Kollman, H.F.Schaefer III, P.R.Schreiner). Wiley, Chichester, 1999. P. 2582
115. P.D.Walker, P.G.Mezey. *J. Am. Chem. Soc.*, **116**, 1202 (1994)
116. A.Golbraikh. *J. Chem. Inf. Comput. Sci.*, **40**, 414 (2000)
117. M.Randić. *J. Chem. Inf. Comput. Sci.*, **36**, 1092(1996)
118. M.Randić. *J. Chem. Inf. Comput. Sci.*, **32**, 686 (1992)
119. M.Randić. *J. Chem. Inf. Comput. Sci.*, **35**, 373 (1995)
120. W.Zheng, A.Tropsha. *J. Chem. Inf. Comput. Sci.*, **40**, 185 (2000)
121. I.Gutman, L.Türker, I.Bálint. *Indian J. Chem., Sect. A*, **34**, 1 (1995)
122. X.Chen, C.H.Reynolds. *J. Chem. Inf. Comput. Sci.*, **42**, 1407 (2002)
123. D.M.Bayada, H.Hamersma, V.J.van Geerestein. *J. Chem. Inf. Comput. Sci.*, **39**, 1 (1999)
124. D.R.Flower. *J. Chem. Inf. Comput. Sci.*, **38**, 379 (1998)
125. J.W.Godden, L.Xul, J.Bajorath. *J. Chem. Inf. Comput. Sci.*, **40**, 163 (2000)
126. E.Abrahamian, P.C.Fox, L.Narum, I.T.Christensen, H.Thogersen, R.D.Clark. *J. Chem. Inf. Comput. Sci.*, **43**, 458 (2003)
127. J.D.Holliday, S.P.Jelfs, P.Willett, P.Gedeck. *J. Chem. Inf. Comput. Sci.*, **43**, 406 (2003)
128. M.Randić, M.Razinger. *J. Chem. Inf. Comput. Sci.*, **35**, 594 (1995)
129. J.W.Raymond, P.Willett. *J. Comput.-Aid. Mol. Design*, **16**, 521 (2002)
130. D.M.Bayada, R.W.Simpson, A.P.Johnson. *J. Chem. Inf. Comput. Sci.*, **32**, 680 (1992)
131. J.W.Raymond, E.J.Gardiner, P.Willett. *J. Chem. Inf. Comput. Sci.*, **42**, 305 (2002)
132. J.W.Raymond, P.Willett. *J. Comput.-Aid. Mol. Design*, **16**, 59 (2002)
133. D.Robert, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **38**, 469 (1998)
134. R.Benigni, M.Cotta-Ramusino, F.Giorgi, G.Gallo. *J. Med. Chem.*, **38**, 629 (1995)
135. G.Boon, W.Langenaeker, F.De Proft, H.De Winter, J.P.Tollenaere, P.Geerlings. *J. Phys. Chem. A*, **105**, 8805 (2001)
136. P.E.Bowen-Jenkins, W.G.Richards. *Int. J. Quantum Chem.*, **30**, 763 (1986)
137. R.Carbo, B.Calabuig. *J. Mol. Struct. (THEOCHEM)*, **254**, 517 (1992)
138. R.Carbo, L.Leyda, M.Arnau. *Int. J. Quantum Chem.*, **17**, 1185 (1980)
139. R.Carbo, B.Calabuig, L.Vera, E.Besalu. *Adv. Quantum Chem.*, **25**, 253 (1994)
140. F.Manaut, F.Sanz, J.Jose, M.Milesi. *J. Comput.-Aid. Mol. Design*, **5**, 1 (1991)
141. E.E.Hodgkin, W.G.Richards. *Int. J. Quantum Chem., Quantum Biol. Symp.*, **14**, 105 (1987)
142. C.Podlipnic, J.Koller. *Acta Chim. Slov.*, **48**, 325 (2001)
143. J.D.Petke. *J. Comput. Chem.*, **14**, 928 (1993)
144. L.Amat, D.Robert, E.Besalu, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **38**, 624 (1998)
145. P.Bultinck, R.Carbo-Dorca, C.Van Alsenoy. *J. Chem. Inf. Comput. Sci.*, **43**, 1208 (2003)
146. A.Gallegos, X.Girones. *J. Chem. Inf. Model.*, **45**, 321 (2005)
147. A.G.Saliner, L.Amat, R.Carbo-Dorca, T.W.Schultz, M.T.D.Cronin. *J. Chem. Inf. Comput. Sci.*, **43**, 1166 (2003)
148. P.Bultinck, T.Kuppens, X.Girones, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **43**, 1143 (2003)
149. X.Girones, R.Carbo-Dorca, R.Ponec. *J. Chem. Inf. Comput. Sci.*, **43**, 2033 (2003)
150. E.Besalu, X.Girones, L.Amat, R.Carbo-Dorca. *Acc. Chem. Res.*, **35**, 289 (2002)
151. U.A.Chaudry, P.L.A.Popelier. *J. Phys. Chem. A*, **107**, 4578 (2003)
152. U.A.Chaudry, P.L.A.Popelier. *J. Org. Chem.*, **69**, 233 (2004)
153. G.Boon, C.Van Alsenoy, F.De Proft, P.Bultinck, P.Geerlings. *J. Phys. Chem. A*, **107**, 11120 (2003)
154. C.Amovilli, R.McWeeny. *J. Mol. Struct. (THEOCHEM)*, **227**, 1 (1991)
155. N.L.Allan, D.L.Cooper. *J. Chem. Inf. Comput. Sci.*, **32**, 587 (1992)
156. S.E.O'Brien, P.L.A.Popelier. *J. Chem. Soc., Perkin Trans. 2*, 478 (2002)
157. G.Boon, F.De Proft, W.Langenaeker, P.Geerlings. *Chem. Phys. Lett.*, **295**, 122 (1998)
158. S.E.O'Brien, P.L.A.Popelier. *J. Chem. Inf. Comput. Sci.*, **41**, 764 (2001)
159. S.E.O'Brien, P.L.A.Popelier. *Can. J. Chem.*, **77**, 28 (1999)
160. L.Cooper, N.L.Allan. *J. Am. Chem. Soc.*, **114**, 4773 (1992)
161. R.Carbo, E.Besalu, L.Amat, X.Fradera. *J. Math. Chem.*, **19**, 47 (1996)
162. R.Carbo, B.Calabuig. *J. Chem. Inf. Comput. Sci.*, **32**, 600 (1992)
163. R.Carbo, B.Calabuig. *Int. J. Quantum Chem.*, **42**, 1681 (1992)
164. J.Cioslowski, E.D.Fleischmann. *J. Am. Chem. Soc.*, **113**, 64 (1991)
165. X.Fradera, L.Amat, E.Besalu, R.Carbo-Dorca. *Quantum Struct. – Act. Relat.*, **16**, 25 (1997)
166. A.C.Good, S.-S.So, W.G.Richards. *J. Med. Chem.*, **36**, 433 (1993)
167. P.L.A.Popelier, U.A.Chaudry, P.J.Smith. *J. Chem. Soc., Perkin Trans. 2*, 1231 (2002)
168. P.L.Popelier. *J. Phys. Chem. A*, **103**, 2883 (1999)
169. R.Ponec, M.Strnad. *Croat. Chim. Acta*, **66**, 123 (1993)
170. D.Robert, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **38**, 620 (1998)
171. M.Sola, J.Mestres, J.M.Oliva, M.Duran, R.Carbo. *Int. J. Quantum Chem.*, **58**, 361 (1996)
172. R.Hefferlin, M.T.Matus. *J. Chem. Inf. Comput. Sci.*, **41**, 484 (2001)
173. X.Girones, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **42**, 317 (2002)
174. X.Girones, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **42**, 1185 (2002)
175. P.Bultinck, R.Carbo-Dorca. *J. Chem. Inf. Comput. Sci.*, **43**, 170 (2003)
176. L.Amat, E.Besalu, R.Carbo-Dorca, R.Ponec. *J. Chem. Inf. Comput. Sci.*, **41**, 978 (2001)
177. J.Danzinger, P.M.Dean. *J. Theor. Biol.*, **116**, 215 (1985)
178. E.Fontain. *J. Chem. Inf. Comput. Sci.*, **32**, 748 (1992)
179. M.G.Bures, Y.C.Martin, P.Willett. *Top. Stereochem.*, **21**, 467 (1994)
180. M.J.Sippl, H.Stegbuchner. *Comput. Chem.*, **15**, 73 (1991)
181. A.Y.Meyer, W.G.Richards. *J. Comput.-Aid. Mol. Design*, **5**, 427 (1991)
182. D.E.Walters, A.J.Hopfinger. *J. Mol. Struct. (THEOCHEM)*, **134**, 317 (1986)
183. X.Wang, J.T.L.Wang. *J. Chem. Inf. Comput. Sci.*, **40**, 442 (2000)
184. T.Kotani, K.Higashiura. *J. Chem. Inf. Comput. Sci.*, **42**, 58 (2002)
185. G.Bringmann, C.Rummey. *J. Chem. Inf. Comput. Sci.*, **43**, 304 (2003)
186. R.D.Brown, Y.C.Martin. *J. Chem. Inf. Comput. Sci.*, **37**, 1 (1997)
187. J.W.Raymond, P.Willett. *J. Chem. Inf. Comput. Sci.*, **43**, 908 (2003)
188. N.Rhodes, P.Willett, A.Calvet, J.B.Dunbar, C.Humblet. *J. Chem. Inf. Comput. Sci.*, **43**, 443 (2003)
189. F.Boeckler, H.Lanig, P.Gmeiner. *J. Med. Chem.*, **48**, 694 (2005)
190. V.Consonni, R.Todeschini, M.Pavan. *J. Chem. Inf. Comput. Sci.*, **42**, 682 (2002)
191. V.Consonni, R.Todeschini, M.Pavan, P.Gramatica. *J. Chem. Inf. Comput. Sci.*, **42**, 693 (2002)
192. R.D.Clark, A.M.Ferguson, R.D.Crame. *Persp. Drug Discov. Design*, **9–11**, 213 (1998)
193. A.C.Good, W.G.Richard. *Persp. Drug Discov. Design*, **9–11**, 321 (1988)
194. B.B.Goldman, W.T.Wipke. *J. Chem. Inf. Comput. Sci.*, **40**, 644 (2000)
195. S.Handschuh, M.Wagener, J.Gastaiger. *J. Chem. Inf. Comput. Sci.*, **38**, 220 (1998)
196. T.W.Heritage, A.M.Ferguson, D.B.Turner, P.Willett. *Persp. Drug Discov. Design*, **9–11**, 381 (1998)
197. H.Kubinyi. *Persp. Drug Discov. Design*, **9–11**, 225 (1998)
198. R.S.Pearlman, K.M.Smith. *Persp. Drug Discov. Design*, **9–11**, 339 (1998)
199. B.D.Silverman, D.E.Platt. *J. Med. Chem.*, **39**, 2129 (1996)
200. D.A.Thorner, D.J.Wild, P.Willett, P.M.Wright. *Persp. Drug Discov. Design*, **9–11**, 301 (1998)
201. R.Todeschini, P.Gramatica. *Persp. Drug Discov. Design*, **9–11**, 355 (1998)
202. D.J.Wild, P.Willett. *J. Chem. Inf. Comput. Sci.*, **36**, 159 (1996)
203. P.Constans, R.Carbo. *J. Chem. Inf. Comput. Sci.*, **35**, 1046 (1995)

204. J.C.Duca, A.J.Hopfinger. *J. Chem. Inf. Comput. Sci.*, **41**, 1367 (2001)
205. B.C.P.Allen, G.H.Grant, W.G.Richards. *J. Chem. Inf. Comput. Sci.*, **43**, 134 (2003)
206. S.L.Dixon, K.M.Merz. *J. Med. Chem.*, **44**, 3795 (2001)
207. R.D.Brown, Y.C.Martin. *J. Chem. Inf. Comput. Sci.*, **36**, 572 (1996)
208. H.Matter. *J. Med. Chem.*, **40**, 1219 (1997)
209. V.J.Gillet, P.Willett, J.Bradshaw. *J. Chem. Inf. Comput. Sci.*, **43**, 338 (2003)
210. M.Thimm, A.Goede, S.Hougardy, R.Preissner. *J. Chem. Inf. Comput. Sci.*, **44**, 1816 (2004)
211. M.Whittle, P.Willett, W.Kraffke, P.van Noort. *J. Chem. Inf. Comput. Sci.*, **43**, 449 (2003)
212. J.D.Holliday, N.Salim, M.Whittle, P.Willett. *J. Chem. Inf. Comput. Sci.*, **43**, 819 (2003)
213. M.Cotta-Ramusino, R.Benigni, L.Passerini, A.Giuliani. *J. Chem. Inf. Comput. Sci.*, **43**, 248 (2003)
214. M.Whittle, V.J.Gillet, P.Willett, A.Alex, J.Loesel. *J. Chem. Inf. Comput. Sci.*, **44**, 1840 (2004)
215. S.F.B.Abolmaali, C.Ostermann, A.Zell. *J. Mol. Model.*, **9**, 66 (2003)
216. N.Salim, J.D.Holliday, P.Willett. *J. Chem. Inf. Comput. Sci.*, **43**, 435 (2003)
217. М.И.Скворцова, И.В.Станкевич, И.И.Баскин, В.А.Палюлин, Н.С.Зефирова. *Докл. АН*, **350**, 786 (1996)
218. M.I.Skvortsova, I.I.Baskin, I.V.Stankevich, V.A.Palyulin, N.S.Zefirov. *J. Chem. Inf. Comput. Sci.*, **38**, 785 (1998)
219. А.В.Пасюков, М.И.Скворцова, В.А.Палюлин, Н.С.Зефирова. *Докл. АН*, **374**, 786 (2000)
220. P.J.Peruzzo, D.J.G.Marino, E.A.Castro, A.A.Toropov. *Internet Electron. J. Mol. Des.*, **2**, 334 (2003); <http://www.biochempress.com>
221. R.P.Sheridan, S.K.Kearsley. *Drug Discov. Today*, **17**, 903 (2002)
222. E.J.Vidlock, V.K.Chung, M.A.Mohan, T.M.Strok, D.J.Austin. *J. Am. Chem. Soc.*, **126**, 3730 (2004)
223. C.Qin, X.Bu, X.Wu, Z.Guo. *J. Combinat. Chem.*, **5**, 353 (2003)
224. G.M.Makara. *J. Med. Chem.*, **44**, 3563 (2001)
225. Y.-K.Kim, M.A.Arai, T.Arai, J.O.Lamenza, E.F.Dean, N.Patterson, P.A.Clemons, S.L.Schreiber. *J. Am. Chem. Soc.*, **126**, 14740 (2004)
226. D.K.Agrafiotis, D.N.Rassokhin. *J. Chem. Inf. Comput. Sci.*, **42**, 117 (2002)
227. D.K.Agrafiotis. *J. Chem. Inf. Comput. Sci.*, **41**, 159 (2001)
228. S.J.Yu, S.M.Keenan, W.Tong, W.J.Welsh. *Chem. Res. Toxicol.*, **15**, 1229 (2002)
229. C.H.Reynolds, A.Tropsha, L.B.Pahler, R.Druker, S.Chakravorty, G.Ethiraj, W.Zheng. *J. Chem. Inf. Comput. Sci.*, **41**, 1470 (2001)
230. A.M.M.Jorgensen, J.T.Pedersen. *J. Chem. Inf. Comput. Sci.*, **41**, 338 (2001)
231. C.T.Clein, D.C.Kaiser, G.Ecker. *J. Chem. Inf. Comput. Sci.*, **44**, 200 (2004)
232. S.Monteiro, M.A.Picarra-Pereira, P.R.Mesquita, V.B.Loureiro, A.Teixeira, R.B.Ferreira. *J. Agric. Food Chem.*, **49**, 3999 (2001)
233. S.V.Trepalin, V.A.Gerasimenco, A.V.Kozyukov, N.P.Savchuk, A.A.Ivaschenko. *J. Chem. Inf. Comput. Sci.*, **42**, 249 (2002)
234. W.H.B.Sauer, M.K.Schwarz. *J. Chem. Inf. Comput. Sci.*, **43**, 987 (2003)
235. M.Johnson. In *Chemical Information Systems: beyond the Structure Diagram*. (Eds D.Bawden, E.Mitchell). Ellis Horwood, New York, 1990. P. 149
236. H.Friedman, B.W.Silverman. *Technometrics*, **31**, 3 (1989)
237. В.Н.Вапник, А.Я.Червоненкис. *Теория распознавания образов*. Наука, Москва, 1974
238. А.А.Горелик, В.А.Скрипкин. *Методы распознавания*. Высшая школа, Москва, 1984
239. Н.Г.Загоруйко. *Прикладные методы анализа данных и знаний*. Изд-во Новосибирского института математики, Новосибирск, 1999
240. W.C.Herndon, G.Rum. In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*. (Eds F.Sanz, J.Giraldo, F.Manant). Prous Science, Barcelona, 1995. P. 380
241. O.Mekenyan, J.Ivanov, S.Karabunarliev, B.Hansen, G.Ankley, S.Bradbury. In *Quantitative Structure — Activity Relationships in Environmental Sciences-VII*. (Eds F.Chen, G.Schürmann). Setac Press, Pensacola, FL, 1997. P. 299
242. W.G.Richard. In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*. (Eds F.Sanz, J.Giraldo, F.Manant). Prous Science, Barcelona, 1995. P. 354
243. S.Tomić, R.P.Gabdouline, B.Kojić-Prodić, R.Wide. *J. Comput.-Aid. Mol. Design*, **12**, 63 (1998)
244. S.C.Basak. In *Practical Applications of Quantitative Structure — Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*. (Eds W.Karcher, J.Devillers). Kluwer Academic, Dordrecht; Boston, 1990. P. 83
245. M.Randić, A.Sablić, S.Nicoljić, N.Trinajstić. *Int. J. Quantum Chem., Quantum Biol. Symp.*, **15**, 267 (1988)
246. J.Takahashi, M.Sukekawa, S.Sasaki. *J. Chem. Inf. Comput. Sci.*, **32**, 639 (1992)
247. М.Ю.Горбачев, А.С.Димогло, И.Б.Берсукер, П.Ф.Влад, М.Н.Колца. *Теорет. эксперим. химия*, **22**, 355 (1986)
248. I.B.Bersuker, A.S.Dimoglo, M.Yu.Gorbachev, P.F.Vlad. *Nouv. J. Chim.*, **9**, 211 (1985)
249. S.C.Basak, B.D.Gute, G.D.Grunwald. *Croat. Chim. Acta*, **69**, 1159, (1996)
250. S.C.Basak, G.D.Grunwald. *Chemosphere*, **31**, 2529 (1995)
251. H.W.Kroto, J.R.Heath, S.C.O'Brien, R.F.Curl, R.E.Smalley. *Nature (London)*, **318**, 162 (1985)
252. Д.А.Бочвар, Е.Г.Гальперн. *Докл. АН СССР*, **209**, 610 (1973)
253. E.Osawa. *Kagaku (Kyoto)*, **25**, 854 (1970); *Chem. Abstr.*, **74**, 75698 (1971)
254. Р.Фейнман. *Характер физических законов*. Мир, Москва, 1968

MOLECULAR SIMILARITY CONCEPT AND ITS USE FOR PREDICTION OF THE PROPERTIES OF CHEMICAL COMPOUNDS

M.I.Skvortsova, I.V.Stankevich, V.A.Palyulin, N.S.Zefirov

N.D.Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences

47, Leninsky prosp., 119991 Moscow, Russian Federation, Fax +7(495)135–5328

A.N.Nesmeyanov Institute of Organoelement Compounds, Russian Academy of Sciences

28, Ul. Vavilova, 119991 Moscow, Russian Federation, Fax +7(495)135–5085

Department of Chemistry, M.V.Lomonosov Moscow State University

Leninskie Gory, 119992 Moscow, Russian Federation, Fax +7(495)939–0290

Key methods for determination of the degree of molecular similarity used in the modern theoretical chemistry are surveyed. Methods for predicting the properties of chemical compounds based on the statement «similar structures have similar properties» are described. Examples of using these approaches for modelling the structure–property relationships for organic compounds are given.

Bibliography — 254 references.

Received 11th November 2005