

**Министерство образования и науки Украины
Харьковский национальный университет им.В.Н.Каразина**

В.В.ИВАНОВ, Л.А.СЛЕТА

**РАСЧЕТНЫЕ МЕТОДЫ ПРОГНОЗА БИОЛОГИЧЕСКОЙ
АКТИВНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ**

Харьков - 2003

УДК 615.015.11

Иванов В.В, Слета Л.А.

**РАСЧЕТНЫЕ МЕТОДЫ ПРОГНОЗА БИОЛОГИЧЕСКОЙ
АКТИВНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ.**

Харьков: ХНУ, 2003–71с.

Учебное пособие является систематическим изложением основ методологии прогноза биологической активности органических молекул. Рассмотрены базовые параметры характеризующие молекулярную структуру, а также распространенные статистические методы обработки экспериментальных данных по биоактивности.

Для аспирантов и студентов старших курсов специальностей химия, биохимия, биофизика и медицинская физика.

Рис. 5. Библиография: 7 назв.

Рецензенты:

Доктор химических наук Дорошенко А.О.

Рекомендовано к печати методической комиссией
химического факультета

Харьковского национального университета им. В.Н.Каразина
Протокол № 5 от 14 мая 2003 г.

© Иванов В.В., Слета Л.А.

©Харьковский национальный университет им.В.Каразина, 2003 г.

СОДЕРЖАНИЕ

Введение.....	4
1. Дескрипторный метод описания структуры молекул.....	6
1.1. Индексы электронной структуры.....	6
1.2. Задачи.....	15
1.3. Топологические и теоретико-информационные индексы.....	21
1.4. Задачи.....	31
2. Регрессионные модели биологической активности.....	33
2.1. Метод наименьших квадратов.....	34
2.2. Эмпирические константы заместителей. Уравнения Гаммета и Тафта.....	36
2.3. Аддитивная модель Фри-Вильсона.....	39
2.4. Метод Хэнча. Оценка липофильности.....	41
2.5. Задачи.....	46
3. Статистические методы классификации молекул по их биологической активности.....	53
3.1. Логико-комбинаторный подход.....	54
3.2. Кластерный анализ.....	59
3.3. Линейный дискриминационный анализ.....	61
3.4. Задачи.....	64
4. Заключение.....	68
5. Список литературы.....	70
6. Приложение. Некоторые термины и параметры биологической активности.....	71

ВВЕДЕНИЕ

Современная технология ставит перед учеными задачу разработки методов надежного предсказания физико-химических и биологических видов активности различных классов химических соединений. Эта проблема имеет общее название *количественное соотношение структура - активность* (*Quantitative Structure – Activity Relationship, QSAR*) и в настоящее время является многопрофильной задачей, включающей в себя различные подходы не только в рамках фармакологической химии, но и физики и математики.

Для того, чтобы найти соотношение между заданным видом активности органических соединений и их структурой, последнюю необходимо представить в численном виде. Численные характеристики, выражающие структурные особенности молекулы, получили общее наименование *дескрипторов молекулярной структуры (или просто дескрипторы)*. В связи с этим главная проблема QSAR – выбор такого (желательно минимального) набора дескрипторов, который достаточен для описания заданного свойства. Хорошие перспективы при решении данной проблемы имеют методы квантовой химии, с помощью которых можно проводить расчеты электронной структуры и геометрии самых разнообразных химических систем вплоть до моделирования эффектов среды и взаимодействия молекулы с рецептором.

Особенностью проблематики QSAR является необходимость широкого внедрения математических методов обработки экспериментальных данных, позволяющих строить различные типы функциональных зависимостей. Здесь наряду с хорошо известным регрессионным анализом применяются *факторные методы*, позволяющие выделять наиболее существенные параметры, описывающие систему, а затем проводить классификацию химических соединений по типам

биологической активности. В целом методология QSAR позволяет заменить поиск соотношений “структура – активность” анализом соотношений “дескрипторы – активность” и на конечном этапе исследований получать модельные функции вида:

$$Y = Y(d_1, d_2, \dots, d_N),$$

где искомое свойство Y выражается через значения дескрипторов молекулярной структуры d_i . Подобные уравнения фактически являются схемами для расчета самых различных свойств органических соединений.

Область применения построенных моделей зависит как от их качества с точки зрения статистики, так и от представительности базы данных, использованной для их построения.

В процессе моделирования важно контролировать прогностическую способность построенных моделей. Один из наиболее надежных подходов к решению этой задачи состоит в разделении исходной базы данных на *обучающую выборку*, использующуюся для построения моделей, и *контрольную выборку*. Последняя необходима для независимого контроля предсказательной способности моделей. Такой подход позволяет в большинстве случаев избежать случайных корреляций.

Целью настоящего пособия является краткое изложение основных понятий и методологии QSAR, а также описание различных систем параметров молекул и методов оценки биологической активности на основе статистического анализа эмпирических данных. Для каждого из таких подходов предлагается набор соответствующих задач. Все объекты, описанные в этих задачах, и соответствующие данные об их биологической активности взяты из оригинальных работ.

1. ДЕСКРИПТОРНЫЙ МЕТОД ОПИСАНИЯ СТРУКТУРЫ МОЛЕКУЛ

В последние десятилетия для описания структуры молекул были предложены тысячи различных дескрипторов, которые включают индексы, характеризующие топологию, электронную структуру, энергию, форму молекул и т.д.

Строение молекулы может быть описано с разной степенью детализации. Так, *брутто-формула* показывает, сколько и какие атомы содержатся в молекуле. *Структурная формула* учитывает лишь типы атомов и характер их связывания, игнорируя при этом геометрические характеристики. Задание координат ядер атомов в пространстве позволяет строить трехмерные *геометрические модели молекул*. С помощью квантовохимических расчетов можно получить дополнительную информацию о распределении электронной плотности (*электронная структура молекулы*). В ряде случаев для характеристики молекул используют физико-химические параметры, отражающие молекулярное строение косвенным образом.

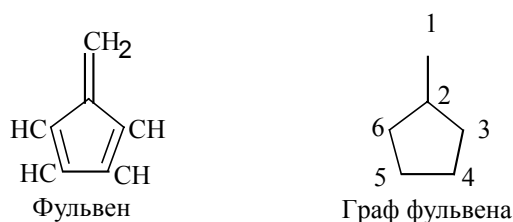
1.1. ИНДЕКСЫ ЭЛЕКТРОННОЙ СТРУКТУРЫ

На первом этапе исследования QSAR необходимо вычислить определенный набор дескрипторов для всех структур в имеющейся базе данных.

Электронные дескрипторы могут быть вычислены с использованием методов квантовой химии. В данном разделе мы рассмотрим возможности применения простейшего подхода - полуэмпирического метода молекулярных орбиталей Хюккеля (МОХ). Метод МОХ основан на предположении, что каждый атом углерода поставляет в общую систему

один π -электрон, а гетероатомы – один или два (так называемое π -электронное приближение).

В методе Хюккеля обычную структурную формулу органического вещества заменяют его *молекулярным графом* или *хюккелевским графом*. Хюккелевский граф наряду с *топологической матрицей* передает информацию о молекулярной структуре сопряженных и ароматических соединений. Так, для молекулы фульвена граф имеет вид:



Топологическая матрица (или матрица смежности) графа G представляет собой симметричную матрицу порядка N , которая определяется следующим образом. Матричный элемент G_{rs} равен 1, если вершины графа (атомы) связаны друг с другом, в противном случае матричный элемент G_{rs} равен 0, т.е. топологическая матрица состоит из 1 и 0. Например, топологическая матрица графа фульвена имеет вид:

$$G = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Хюккелевские молекулярные орбитали (МО) являются собственными функциями φ_i эффективного одноэлектронного гамильтониана, обозначаемого h :

$$h \varphi_i = \varepsilon_i \varphi_i.$$

Величина орбитальной энергии ε_i представляет собой собственное значение энергии, связанное с собственной функцией φ_i . Каждая МО выражается в виде линейной комбинации атомных орбиталей:

$$\varphi_i = \sum_{r=1}^N c_{ir} \chi_r,$$

где суммирование проводится по всем атомам r сопряженной молекулы, χ_r – атомная $2p_z$ – орбиталь атома с номером r , а c_{ir} – вклад r -й атомной орбитали в i -ю МО (так называемый орбитальный коэффициент). При этом атомные орбитали χ_r предполагаются ортонормированными:

$$\langle \chi_r | \chi_s \rangle = \delta_{rs}.$$

Полная π -электронная энергия E_π определяется выражением

$$E_\pi = \sum_{i=1}^N n_i \varepsilon_i,$$

где n_i – число заполнения орбитали φ_i , принимающее значения 2, 1 или 0.

Минимизация E_π с помощью вариационного метода приводит к системе *секулярных (или вековых) уравнений*

$$\sum_{s=1}^N (h_{rs} - \varepsilon_i \delta_{rs}) c_{is} = 0, \quad s = 1, \dots, N,$$

где h_{rs} – матричные элементы гамильтониана в базисе атомных орбиталей:

$$h_{rs} = \langle \chi_r | \mathbf{h} | \chi_s \rangle.$$

Секулярные уравнения имеют нетривиальные решения только при условии обращения в нуль *секулярного (векового) детерминанта*

$$\det | \mathbf{h} - E_\pi \mathbf{I} | = 0,$$

где \mathbf{I} – единичная матрица.

Упрощение секулярного детерминанта достигается путем введения следующих *хюккелевских приближений*:

- $h_{rr} = \langle \chi_r | \mathbf{h} | \chi_r \rangle = \alpha,$

где α - кулоновский интеграл, который предполагается одинаковым для всех орбиталей, центрированных на идентичных атомах r ,

$$\bullet \quad h_{rs} = \langle \chi_r | h | \chi_s \rangle = \begin{cases} 0, & \text{если атомы } r, s \text{ не связаны} \\ \beta, & \text{если атомы } r, s \text{ связаны} \end{cases}$$

где β называется *резонансным интегралом*, в простейшем варианте теории он предполагается одинаковым для всех связей между атомами одного типа. С учетом хюккелевских приближений матрица h приобретает вид

$$h = \alpha I + \beta G,$$

где G – топологическая матрица молекулы.

Секулярный детерминант легко преобразовать к виду:

$$\det \left| \frac{\varepsilon_i - \alpha}{\beta} I - G \right| = 0.$$

Численные величины $x = \frac{\varepsilon_i - \alpha}{\beta}$, входящие в секулярный детерминант, определяют энергию π -электронов на отдельных МО и образуют набор собственных значений рассматриваемого хюккелевского графа, т.е.

$$\varepsilon_i = \alpha + x_i \beta, \quad i = 1, 2, \dots, N.$$

Набор собственных значений $x_1, x_2, x_3, \dots, x_N$ матрицы G называют *спектром графа*.

Поскольку матрицы h и G коммутируют, они обладают общим набором собственных векторов. Собственные векторы топологической матрицы G идентичны хюккелевским молекулярным орбиталям.

На основе решений метода Хюккеля в рамках π -электронного приближения разработана система дескрипторов молекулярной структуры (табл.1.1).

Таблица 1.1 Наиболее распространенные квантовохимические индексы π -электронной структуры

Тип дескриптора	Дескриптор	Предсказываемое физико-химическое свойство
Индексы электронной структуры	Электронная плотность – ρ_r	Электрофильные и нуклеофильные центры, Дипольный момент
	Порядок связи – ρ_{rs}	Длина, силовая постоянная, прочность химической связи
	Индекс свободной валентности – F_r	Реакционная способность в радикальных реакциях замещения
	Суперделокализуемость – $S_e(r), S_n(r)$	Электрофильные, нуклеофильные центры
Энергетические индексы	Поляризуемость молекулы	Взаимодействие электронной оболочки с электрическим полем, реакционная способность
	Полная энергия – E_π	Устойчивость конформации
	Энергия связи – E_b	Энергия разрыва связи
	Энергия высшей занятой МО – $\varepsilon_{(HOMO)}$	Энергия ионизации
	Энергия низшей вакантной МО – $\varepsilon_{(LUMO)}$	Энергия сродства к электрону
	Энергия электронного перехода – λ	Электронные спектры поглощения
Энергия делокализации – E_D	Устойчивость сопряженной системы	

Электронная плотность характеризует вероятность пребывания электронов в данной области молекулы (у данного атома r)

$$\rho_r = \sum_i n_i c_{ir}^2,$$

где n_i – число электронов на i -той МО, c_{ir} – орбитальный коэффициент разложения i -той молекулярной орбитали по r -той атомной орбитали (атом r). С этой величиной связан также заряд на атоме (q_r):

$$q_r = Z_r - \rho_r,$$

где Z_r – число π -электронов атома, отданных в систему сопряжения.

π -Порядок связи характеризует степень π -электронного связывания.

Эту величину для связи r-s можно вычислить следующим образом:

$$\rho_{rs} = \sum_i n_i c_{ir} c_{is}.$$

Индекс свободной валентности F_r характеризует реакционную способность углеродного атома r и согласно определению вычисляется следующим образом:

$$F_r = N_{\max} - N_r,$$

где $N_{\max} = \sqrt{3}$ – максимальная для атома углерода величина, характеризующая способность к образованию π -связи, а N_r – сумма порядков всех связей атома r:

$$N_r = \sum_{(rs)} \rho_{rs}$$

Суперделокализуемость атома r (индекс Фукуи) различают электрофильную – $S_e(r)$ и нуклеофильную – $S_n(r)$:

$$S_e(r) = 2 \sum_a^{\text{vac.}} \frac{c_{ar}^2}{0.5(\epsilon_{(\text{HOMO})} + \epsilon_{(\text{LUMO})}) - \epsilon_a},$$

$$S_n(r) = 2 \sum_i^{\text{occ.}} \frac{c_{ir}^2}{\epsilon_i - 0.5(\epsilon_{(\text{HOMO})} + \epsilon_{(\text{LUMO})})},$$

где $\epsilon_{(\text{HOMO})}$ – энергия высшей заполненной МО, $\epsilon_{(\text{LUMO})}$ – энергия нижней вакантной МО. Эти величины интерпретируются как индексы, характеризующие способность к делокализации (стабилизации) электронов в переходном комплексе при атаке электрофильным/нуклеофильным реагентом. Особенно широкое применение получила суперделокализуемость в связи с оценками канцерогенной активности молекул.

Важное значение имеют индексы, характеризующие *поляризуемость* молекулярной системы. Обычно различают так называемую *электрическую дипольную поляризуемость* α и набор *коулсоновских поляризуемостей* π .

Электрическая дипольная поляризуемость характеризует изменение распределения электронной плотности молекулы при помещении её в однородное электрическое поле. Эта величина может быть получена из эксперимента по *молекулярной рефракции* или вычислена как производная дипольного момента ($\bar{\mu}$) по напряженности электрического поля (\vec{E}) из выражения

$$\bar{\mu} = \alpha \cdot \vec{E}.$$

Обычно эта величина используется в виде *средней поляризуемости* $\bar{\alpha}$:

$$\bar{\alpha} = \frac{1}{3}(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}),$$

или (реже) анизотропии поляризуемости:

$$\tilde{\alpha}^2 = \frac{1}{2}[(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{xx})^2].$$

Здесь $\alpha_{xx}, \alpha_{yy}, \alpha_{zz}$ – компоненты тензора дипольной поляризуемости.

Молярная рефракция (MR) также широко используется в качестве дескриптора. В области оптических частот она выражается известной формулой Лорентца-Лоренца:

$$MR = \frac{n^2 - 1}{n^2 + 2} \frac{MW}{d} = \frac{4}{3} \pi N_A \bar{\alpha},$$

где n – показатель преломления вещества, MW – молекулярный вес, а d – плотность.

Известны следующие типы *коулсоновских поляризуемостей*.

- *Самополяризуемость атома* (π_f) характеризует изменение электронной плотности атома при изменении электроотрицательности (кулоновского интеграла) того же атома:

$$\pi_r = \frac{\partial \rho_r}{\partial \alpha_r} = 4 \sum_i^{\text{occ.}} \sum_a^{\text{vac.}} \frac{c_{ir}^2 c_{ar}^2}{\epsilon_i - \epsilon_a}.$$

Сумма самополяризуемостей по всем атомам также может служить хорошим дескриптором “подвижности” электронной оболочки молекулы:

$$\Pi = \sum_r \pi_r.$$

- *Атом-атомная поляризуемость* (π_{rs}) характеризует изменение электронной плотности на атоме (r) при изменении электроотрицательности (кулоновского интеграла) другого атома (s):

$$\pi_{sr} = \pi_{rs} = \frac{\partial \rho_r}{\partial \alpha_s} = 4 \sum_i^{\text{occ.}} \sum_a^{\text{vac.}} \frac{c_{ir} c_{ar} c_{is} c_{as}}{\epsilon_i - \epsilon_a}.$$

Кроме перечисленных известны также поляризуемости типа атом-связь ($\pi_{r,st} = \frac{\partial \rho_r}{\partial \beta_{st}}$) и связь-связь ($\pi_{rs,tu} = \frac{\partial \rho_{rs}}{\partial \beta_{tu}}$), которые, однако, менее популярны в проблеме QSAR.

Энергия разрыва химической связи:

$$E_b = E - E',$$

где E – полная энергия молекулярной системы, а E' – суммарная энергия составных частей, на которые система распадается.

Энергия электронного перехода в различных квантовохимических подходах оценивается по-разному. В простейшем методе Хюккеля её получают как разность энергий МО. Так, энергия длинноволнового перехода вычисляется по формуле:

$$\lambda = \epsilon_{(\text{НВМО})} - \epsilon_{(\text{ВЗМО})}$$

Энергия делокализации, (E_D) является фундаментальной характеристикой сопряженной системы:

$$E_D = E_\pi - \sum_i E_i,$$

где второй член соответствует энергии системы изолированных двойных связей данной молекулы.

Вышеперечисленные квантовохимические индексы нашли широкое распространение в проблеме QSAR. Известно множество корреляционных функций, которые связывают наблюдаемый био- или физико-химический эффект с параметрами электронной структуры молекулы. В качестве примера приведем два уравнения. Так, из данных о токсичности ($\lg LD_{50}$, см. приложение) серии из 20 нитрилов было получено следующее корреляционное уравнение (Chem.Rev., 1996, **96**, 1027):

$$-\lg LD_{50} = -1,69 \frac{\alpha}{\Delta E} + 0,47, \quad (n = 20, r = 0,87, \sigma = 0,2),$$

где α – поляризуемость молекулы, а ΔE – ее энергия протонизации, оцененная методами квантовой химии как разность энергии нейтральной и протонированной формы. Во втором уравнении описывается ингибирование ацетилхолинестеразы производными бензилпиперидина:

$$-\lg IC_{50} = 2,21C_4 - 6,65\mu + 1,18\mu^2 - 162,9\varepsilon_{(\text{HOMO})} - 8,58\varepsilon_{(\text{HOMO})}^2, \\ (n = 16, r = 0,939, \sigma = 0,25)$$

где μ – дипольный момент, C_4 – орбитальный коэффициент в высшей заполненной молекулярной орбитали, а $\varepsilon_{(\text{HOMO})}$ – ее энергия (уравнение взято из работы J.Med.Chem., 1992, **35**, 584).

Метод Хюккеля и другие полуэмпирические методики (CNDO, INDO, AM1, PM3), которые используются в расчетах QSAR, реализованы в множестве пакетов квантовохимических программ (GAUSSIAN, HyperChem и т.д). Программа HyperChem позволяет оценить такие полезные величины как эффективная площадь поверхности молекулы, молекулярный объем, поляризуемость, рефракция. Ряд квантовохимических программ был разработан также и на кафедре технической химии Харьковского национального университета им. В.Н.Каразина.

1.2 ЗАДАЧИ

Задания 1-19 выполните, не используя квантовохимических вычислительных программ.

1. Изобразите молекулярные графы и запишите топологические матрицы для следующих молекул и радикалов: а) этилена, аллила и бутадиена; б) циклопропенила, циклобутадиена, пентадиенила и бензола; в) нафталина и азулена; г) антрацена и фенантрена.

2. Запишите детерминант Хюккеля для следующих молекул и радикалов: а) этилена, аллила и бутадиена; б) циклопропенила, циклобутадиена, пентадиенила и бензола; в) нафталина и азулена; г) антрацена и фенантрена.

3. Вычислите энергии МОХ для молекулы этилена. Постройте энергетическую диаграмму МО основного и возбужденного состояний этилена. Вычислите соответствующие им полные электронные энергии.

4. Вычислите энергии МОХ формальдегида. Полученные результаты сравните с соответствующими энергиями МОХ этилена. (При вычислениях используйте: $\alpha_0 = \alpha + \beta$; $\beta_{CO} = \beta$, где α и β - параметры для атома углерода.)

5. Вычислите энергии МОХ для радикала аллила. Постройте энергетические диаграммы МО для радикала аллила, его катиона и аниона. Вычислите соответствующие энергии связывания. Обсудите полученные результаты.

6. Постройте энергетическую диаграмму МО и вычислите энергию делокализации для молекулы бутадиена, если для связывающих МО значения x равны $-1,618$ и $-0,618$. Какой физический смысл имеет энергия делокализации?

7. Постройте энергетическую диаграмму МО и вычислите энергию делокализации для молекулы бензола, если для связывающих МО параметр x принимает значения $-2, -1$ и -1 .

8. Постройте энергетическую диаграмму МО и вычислите энергию делокализации для молекулы нафталина, если для связывающих МО x принимает значения:

$$-2,303; -1,618; -1,3028; -1,000; -0,618.$$

9. Постройте энергетическую диаграмму МОХ для радикала бензила, если для занятых МО получены следующие значения $x = -2,101; -1,259; -1,000; 0$.

10. Общее решение векового детерминанта в методе МОХ для циклического углеводорода с N атомами имеет вид: $x_k = -2 \cos(2\pi k)/N$, где $k=1, 2, \dots, N$. Используя этот результат, вычислите орбитальные энергии и энергии делокализации для а) радикала циклопропенила; б) молекулы циклобутадиена; в) радикала циклопентадиенила.

11. Используя соответствующий хюккелевский детерминант, составьте хюккелевские уравнения для нахождения орбитальных коэффициентов молекулы этилена. Вычислите орбитальные коэффициенты для связывающей и разрыхляющей МО этилена. Запишите для них аналитические выражения, изобразите их графически.

12. В рамках метода МОХ вычислите электронные плотности на атомах, порядок связи и индексы свободной валентности для молекулы этилена. Результаты расчета представьте на молекулярной диаграмме.

13. В рамках метода МОХ вычислите электронные плотности на атомах, порядки связей и индексы свободной валентности для молекулы бутадиена. Постройте молекулярную диаграмму бутадиена. Используйте следующие выражения для занятых молекулярных орбиталей:

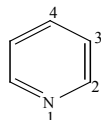
$$\varphi_1 = a(\chi_1 + \chi_4) + b(\chi_2 + \chi_3),$$

$$\varphi_2 = b(\chi_1 - \chi_4) + a(\chi_2 - \chi_3),$$

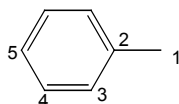
где численные коэффициенты имеют следующие значения:

$$a=0,372, b=0,602.$$

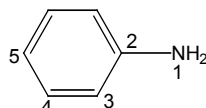
14. Вычислите индексы свободной валентности на атомах углерода в молекуле пиридина, если порядки связей равны $p_{12} = 0,654$, $p_{23} = 0,669$ и $p_{34} = 0,665$.



15. Вычислите индексы свободной валентности на атомах углерода в радикале бензиле, если порядки связей равны $p_{12} = 0,635$, $p_{23} = 0,523$, $p_{34} = 0,705$, $p_{45} = 0,635$.



16. Вычислите индексы свободной валентности на атомах углерода в молекуле анилина, если порядки связей равны $p_{12} = 0,291$, $p_{23} = 0,637$, $p_{34} = 0,673$, $p_{45} = 0,663$.



17. Отличаются ли электронные плотности на атомах в молекулах: а) этилена и бутадиена; б) бензола и нафталина; в) нафталина и азулена; г) антрацена и фенантрена?

18. Постройте молекулярную диаграмму для молекулы бензола, если известно, что порядок углерод-углеродной связи равен 0,667. Вычислите атом-атомные поляризуемости в молекуле бензола.

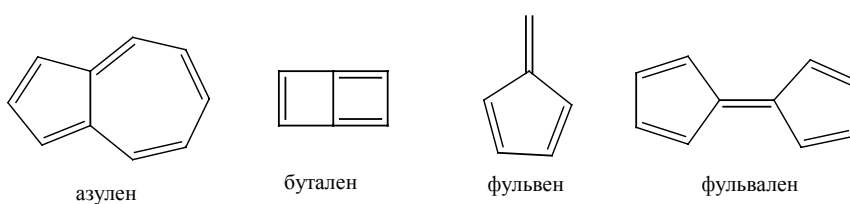
19. В рамках метода МОХ обсудите электронную структуру: а) триметилен метила; б) метиленициклопропена; в) дициклобутадиена; г) пентадиенила.

При выполнении заданий №№ 20—34 проведите компьютерный расчет молекул методом МО Хюккеля.

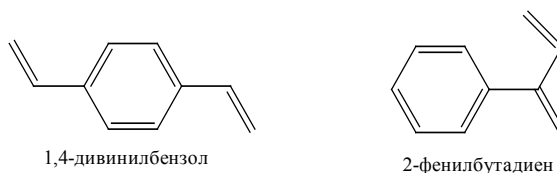
20. Сравните электронные структуры метиленициклопропена и циклопропенона (энергии МО, заряды, порядки связей).



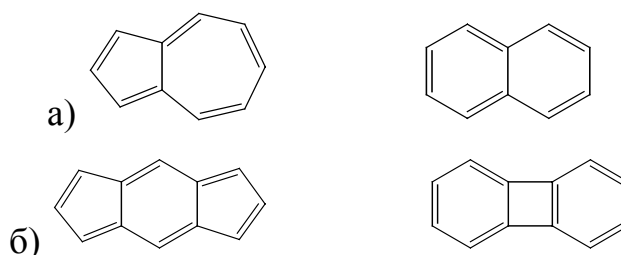
21. Вычислите энергии делокализации следующих небензоидных структур:

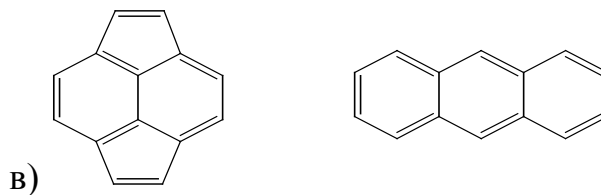


22. Покажите, что 1,4-дивинилбензол и 2-фенилбутадиен являются изоспектральными молекулами (им соответствуют одинаковый спектр энергий хюккелевских молекулярных орбиталей).

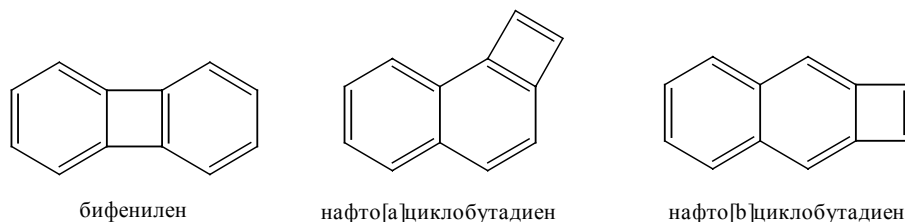


23. Сравните электронные структуры альтернантного и неальтернантного углеводородов. Вычислите заряды, индексы свободной валентности, самополяризуемости и суперделокализуемости углеродных атомов.





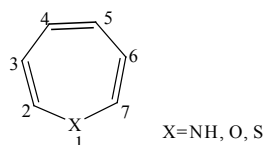
24. Вычислите энергию делокализации и предскажите устойчивость в ряду следующих углеводородов:



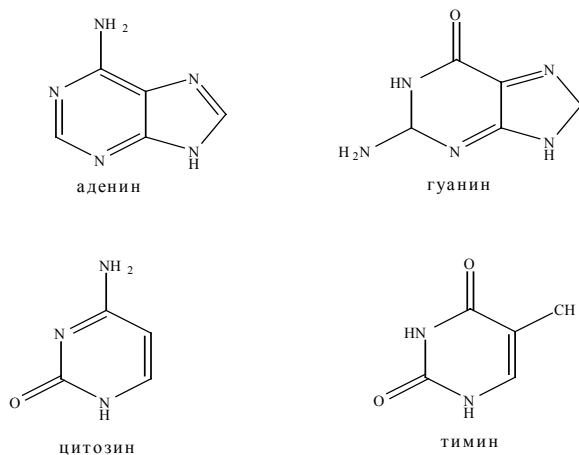
25. Обсудите электронную структуру калицена. Какая из структурных формул наиболее адекватно передает его свойства?



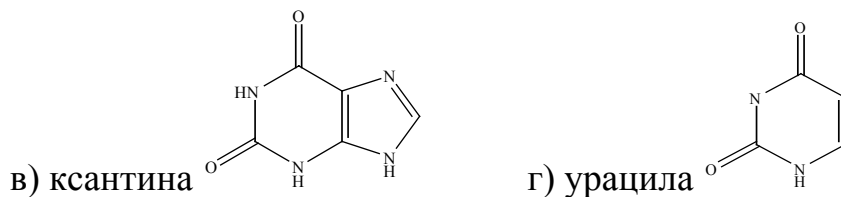
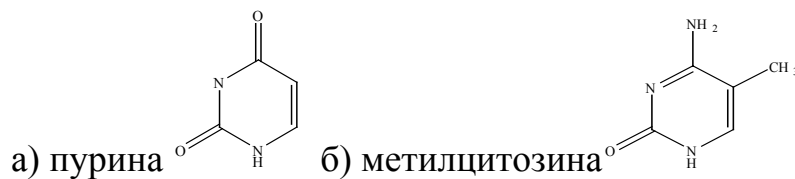
26. Вычислите порядки связей 2-7 в гетеронине:



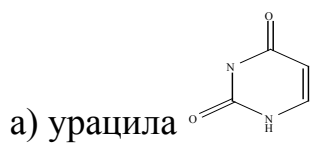
27. Сравните распределение зарядов на атомах и орбитальные энергии следующих нуклеотидных оснований

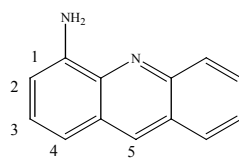


28. Вычислите распределение зарядов на атомах и орбитальные энергии

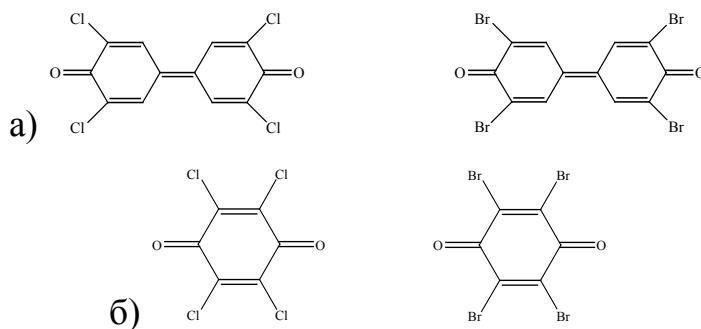


29. Проанализируйте индекс свободной валентности и электрофильную суперделокализуемость атомов углерода для

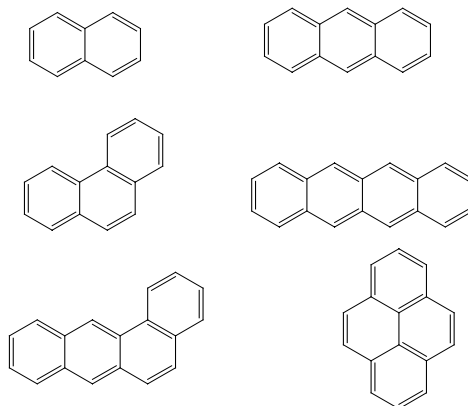


б) аминокридина  , аминогруппа может занимать положения 1, 2, 3, 4 и 5.

30. Сравните заряды на однотипных атомах кислорода и оцените акцепторные свойства следующих соединений:



31. Укажите атомы углерода, для которых индексы свободной валентности и электрофильной суперделокализуемости имеют наибольшие значения в молекулах следующих соединений:



32. Вычислите самополяризуемости и сверхделокализуемости углеродных атомов бутадиена, если молекулярные орбитали (в порядке возрастания энергии) имеют вид:

$$\begin{cases} \varphi_1 = a(\chi_1 + \chi_4) + b(\chi_2 + \chi_3), \\ \varphi_2 = b(\chi_1 - \chi_4) + a(\chi_2 - \chi_3), \\ \varphi_3 = b(\chi_1 + \chi_4) - a(\chi_2 - \chi_3), \\ \varphi_4 = a(\chi_1 - \chi_4) - b(\chi_2 - \chi_3). \end{cases}$$

$a = 0,372$, $b = 0,602$. Оцените изменение электронной плотности на первом углеродном атоме при увеличении его кулоновского интеграла на 0,1.

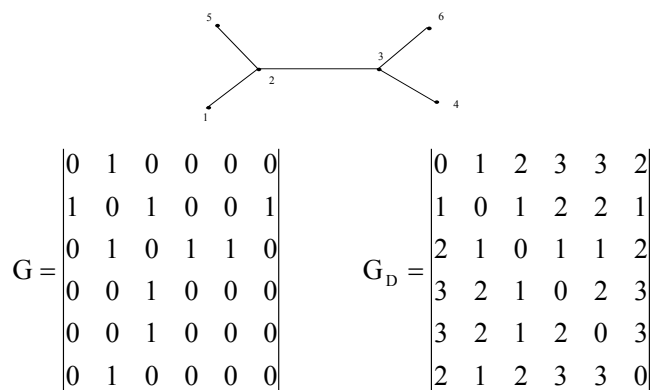
33. Оцените энергию нижайшего электронного возбуждения, первую энергию ионизации и сродство к электрону в ряду молекул: этилен, бутадиен, гексатриен, октатетраен.

1.3 ТОПОЛОГИЧЕСКИЕ И ТЕОРЕТИКО-ИНФОРМАЦИОННЫЕ ИНДЕКСЫ

Топологический подход к описанию молекулярной структуры основан на анализе структурной формулы молекулы и вычислении соответствующих индексов, которые остаются неизменными при

изменении нумерации атомов или связей (так называемые *топологические инварианты*). При этом обычно “легкие” водородные атомы в расчете не учитываются и рассматривается каркас молекулы, построенный из “тяжелых” атомов: C, N, O, S...

Развитие топологического подхода тесно связано с применением теории графов. Кроме *топологической матрицы* (G) здесь часто используется *матрица расстояний* G_D . Каждый элемент в матрице расстояний $g_D(ij)$ равен минимальному числу ребер, соединяющих одну вершину (i) с другой (j). В качестве примера приведем граф и соответствующие матрицы G и G_D для молекулы 2,3-диметилбутана:



Используя матрицы графа (G и G_D) можно получить ряд топологических дескрипторов.

- *Число путей в молекулярном графе определенной длины.*

Число путей между вершинами длины “1” (P_1), т.е. число химически связанных пар атомов, число путей длины “2” или индекс Гордона-Скантлбери (P_2), число путей длины “3” (P_3) получают простым подсчетом количества единиц, двоек или троек в одном из треугольников симметричной матрицы G_D .

- *Индекс Винера W равен сумме связей, существующих между всеми парами “тяжелых” атомов в графе молекулы с n вершинами:*

$$W = \frac{1}{2} \sum_{i,j}^n g_D(ij).$$

- *Индекс Рандича* $\chi^{(1)}$ характеризует молекулярную связность:

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

где – v_i и v_j степени вершин графа, т.е. соответствующие количества связей вершин i и j . Суммирование проводят по всем парам связанных вершин. Известны также обобщенные индексы Рандича, где суммирование проводится по всем цепям маршрутов длины k между вершинами i и j :

$$\chi^{(k)} = \sum_{(i,\dots,l,\dots,j)} (v_i \dots v_l \dots v_j)^{-1/2}$$

- *Индекс среднеквадратичных расстояний* D^2

$$D^2 = \frac{\sum_i P_i i^2}{\sum_i P_i},$$

где P_i – число пар вершин в матрице G_D , расстояние между которыми (длина) равно i .

- *Индексы загребской группы.*

$$M_1(G) = \sum_i v_i^2,$$

$$M_2(G) = \sum_{(i,j)} (v_i v_j).$$

В первом выражении суммирование проводится по всем вершинам, а во втором – по всем парам связанных атомов.

- *Наибольшее собственное значение* λ_{\max} матрицы смежности (G) как топологический индекс был предложен в качестве меры количества разветвлений в структуре графа.

Особенной группой топологических индексов являются так называемые теоретико-информационные индексы, основанные на использовании известной универсальной формулы Шеннона для оценки

неоднородности (информации) любой системы. Предположим, что структура молекулы определенным образом разбита на подмножества. Тогда, вычислив вероятность попадания того или иного элемента в заданное подмножество, можно оценить информацию о распределении всех элементов по подмножествам (формула Шеннона):

$$\text{Информация} = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n} \text{ (бит)},$$

где n_i – количество элементов в подмножестве i , а n – суммарное количество всех элементов системы, равное

$$n = \sum_i n_i .$$

В этом выражении логарифм берется по основанию 2, что соответствует информации выраженной в битах. На основе формулы Шеннона введены следующие теоретико-информационные индексы.

- *Неоднородность распределения расстояний между вершинами в графе*, I_D используется для характеристики разветвленности молекулярной структуры. Здесь степень неоднородности структуры, оценивается на основе анализа распределения расстояний между вершинами графа. Тогда $n_i = P_i$ – количество связей длины “ i ”, а n – общее количество кратчайших расстояний между всеми парами атомов структуры. Часто бывает удобно использовать *общее количество информации* в пересчете на граф – TI_D :

$$TI_D = n \cdot I_D$$

- *Информационное содержание графа относительно окрестностей k -го порядка* – IC_k , представляет собой информационное содержание в расчете на одну вершину:

$$IC_k = -\sum_i p_i \log_2 p_i ,$$

где p_i - вероятность того, что выбранная случайным образом вершина графа (атом) попадет в i -тое подмножество, причем вычисляется она с учетом окружения.

- TIC_k – полное информационное содержание – мера сложности в расчете на один граф.

$$TIC_k = n \cdot IC_k$$

- SIC_k – структурное информационное содержание

$$SIC_k = IC_k / \log_2 n .$$

- BIC_k – информационное содержание связывания

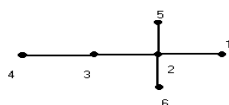
$$BIC_k = IC_k / \log_2 N_b ,$$

где N_b – полное число ребер (ковалентных связей) в молекулярном графе.

- CIC_k – комплементарное информационное содержание

$$CIC_k = \log_2 n - IC_k .$$

В качестве примера рассмотрим вычисление индексов I_D и IC для молекулы 2,2-диметилбутана (остальные информационные индексы легко вычисляются из IC). Граф молекулы можно представить в виде:



Для вычисления индекса I_D подсчитаем число путей между вершинами длины “1”, “2”, “3” и т.д. $P_1=5$, $P_2=7$, $P_3=3$. Путей длины более “3” в этом графе нет, поэтому

$$n = P_1 + P_2 + P_3 = 5 + 7 + 3 = 15 .$$

И, следовательно, индекс I_D равен

$$I_D = - \left(\frac{5}{15} \log_2 \frac{5}{15} + \frac{7}{15} \log_2 \frac{7}{15} + \frac{3}{15} \log_2 \frac{3}{15} \right) = 1,506 \text{ (бит)}$$

В отличие от индекса I_D индекс IC_k вычисляется на основе анализа вершин (атомов). Для IC_0 подсчитывается количество неэквивалентных атомов без учета ближайших соседей. Поскольку в этой молекуле все атомы одного сорта – атомы углерода, то все они эквивалентны и, следовательно, относятся к одной группе.

$$IC_0 = -\frac{6}{6} \log_2 \frac{6}{6} = 0.$$

Индекс IC_1 вычисляется с учетом ближайших соседей. Анализ графа показывает, что при таком рассмотрении можно выделить группы атомов (см. табл 1.2)

Таблица 1.2 Группы атомов в молекуле 2,2-диметилбутан с учетом ближайшего окружения.

Группа	Атомы группы	Число атомов в группе
1	1, 5, 6, 4	4
2	2	1
3	3	1

Следовательно, значение индекса IC_1 равно

$$IC_1 = -\left(\frac{4}{6} \log_2 \frac{4}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6}\right) = 1,252 \text{ (бит)}$$

Индекс IC_2 вычисляется с учетом ближайших и следующих за ними соседей. Четыре группы эквивалентности атомов с учетом соседей “второго порядка” приведены в табл. 1.3.

Таблица 1.3 Группы атомов в молекуле 2,2-диметилбутан с учетом окружения второго порядка.



Группа	Атомы группы	Число атомов в группе
1	1, 5, 6	3
2	2	1

3	3	1
4	4	1

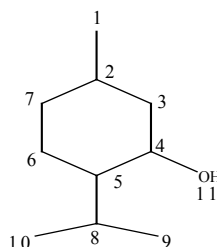
$$IC_2 = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 1,792 \text{ (бит)}$$

В табл.1.4 приведены для сравнения некоторые топологические индексы двух изомеров гексана – н-гексана и 2,2-диметилбутана.

Таблица 1.4 Некоторые топологические индексы для двух изомеров гексана

Индекс		
	(н-гексан)	(2,2-диметилбутан)
N_C	6	6
P_1	5	5
P_2	4	7
P_3	3	3
W	35	28
I_D	2,149	1,506
IC_0	0	0
IC_1	0,918	1,252
IC_2	1,585	1,792
$\chi^{(1)}$	2,914	2,561

Теперь рассмотрим более сложную систему – молекулу ментола, молекулярный граф которой имеет вид (как и в предыдущих случаях, атомы водорода не рассматриваем):



Нулевой порядок (IC_0).

С учетом окрестностей нулевого порядка все десять углеродных атомов эквивалентны. Следовательно, одиннадцать “тяжелых атомов” системы

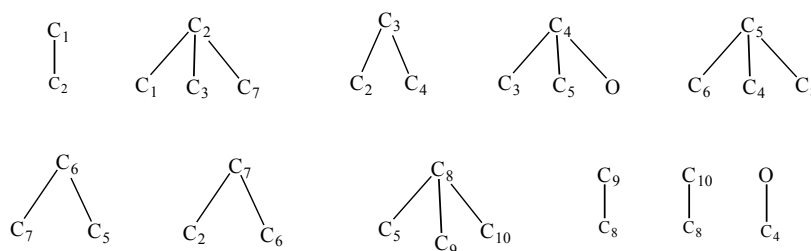
могут быть сгруппированы в две группы – все атомы углерода и кислород.

Информационное содержание в этом случае равно

$$IC_0 = -\left(\frac{10}{11} \log_2 \frac{10}{11} + \frac{1}{11} \log_2 \frac{1}{11}\right) = 0,439 \text{ (бит)}.$$

$$TIC_0 = 4,829, \quad SIC_0 = 0,127, \quad VIC_0 = 0,127, \quad CIC_0 = 3,020$$

Первый порядок (IC_1). Рассмотрим окрестности первого порядка (ближайшие соседи) для всех атомов



Очевидно, что имеется 5 групп эквивалентности, а именно

C_1, C_9, C_{10} , – 3 атома;

C_2, C_5, C_8 , – 3 атома;

C_3, C_6, C_7 , – 3 атома;

C_4 – 1 атом;

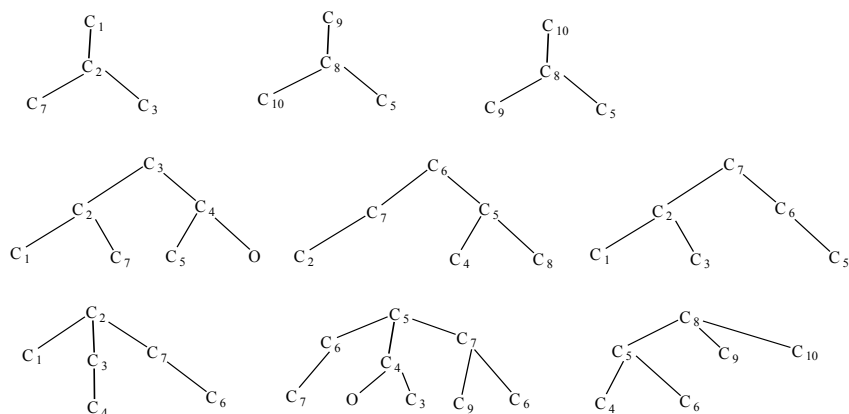
O – 1 атом.

Используя эти данные, можно вычислить IC_1 :

$$IC_1 = -\left(\frac{3}{11} \log_2 \frac{3}{11} + \frac{3}{11} \log_2 \frac{3}{11} + \frac{3}{11} \log_2 \frac{3}{11} + \frac{1}{11} \log_2 \frac{1}{11} + \frac{1}{11} \log_2 \frac{1}{11}\right) = 2,163 \text{ (бит)}$$

$$TIC_1 = 23,793, \quad SIC_1 = 0,625, \quad VIC_1 = 0,625, \quad CIC_1 = 1,296$$

Второй порядок (IC_2). Окрестности второго порядка (учет соседей через один атом) для всех эквивалентных атомов в первом порядке выглядят следующим образом:



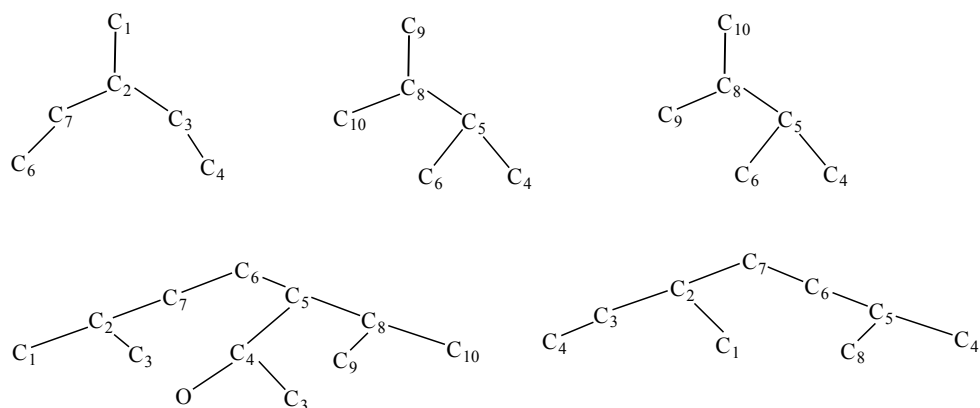
Итак, имеем следующие группы эквивалентности:

- C_1, C_9, C_{10} – 3 атома;
- C_6, C_7 – 2 атома;
- а также шесть групп по одному атому.

$$IC_2 = -\left(\frac{3}{11} \log_2 \frac{3}{11} + \frac{2}{11} \log_2 \frac{2}{11} + 6 \cdot \frac{1}{11} \log_2 \frac{1}{11}\right) = 2,845 \text{ (бит)}$$

$$TIC_2 = 31,295, \quad SIC_2 = 0,822, \quad VIC_2 = 0,822, \quad CIC_2 = 0,614$$

Третий порядок (IC_3). Рассматриваем все эквивалентные атомы во втором порядке и выписываем окрестности через два соседних атома.



- Атомы C_9 и C_{10} включаются в одну группу;
- остальные группы по одному атому.

$$IC_3 = -\left(\frac{2}{11}\log_2\frac{2}{11} + 9\cdot\frac{1}{11}\log_2\frac{1}{11}\right) = 3,278 \text{ (бит)}$$

$$TIC_3 = 36,058, \quad SIC_3 = 0,948, \quad VIC_3 = 0,948, \quad CIC_3 = 0,181$$

Очевидно, что дальнейшее расширение окрестностей не приведет к изменению информационного содержания, поскольку атомы 9 и 10 топологически неразличимы по соображениям симметрии.

Перечисленные топологические и информационные индексы широко применяются для построения корреляционных зависимостей описывающих самые различные биологические и физико-химические свойства молекул. Ряд примеров использования этих индексов приведен в последующих разделах, а здесь, в качестве иллюстрации, мы запишем два простых корреляционных соотношения. Первое связывает температуру кипения алканов $T(^{\circ}C)$ с индексом Рандича $\chi^{(1)}$:

$$T(^{\circ}C) = 57,85\chi^{(1)} - 97,9$$

Применение этого уравнения для расчета $T(^{\circ}C)$ изомеров гексана демонстрирует удовлетворительное согласие с экспериментальными величинами (табл. 1.5).

Таблица 1.5. Индекс Рандича и температура кипения изомеров гексана

	2-метил- пентан	2,2-диметил- бутан	2,3-диметил- бутан	н-гексан
$\chi^{(1)}$	2,77	2,56	2,64	2,91
$T(^{\circ}C)$	62,3	50,2	55,0	70,7
$T(^{\circ}C)$ эксп.	60,3	49,7	58,0	68,7

Во втором уравнении (J.Pharm.Sci., 1975, **64**, 1971) тот же индекс Рандича линейно связан с минимальной блокирующей концентрацией анестетиков

(спирты, кетоны, эфиры, амины, фенолы) для мускульных или нервных волокон ($\lg \text{MBC}$).

$$\lg \text{MBC} = -0,762\chi^{(1)} + 3,55 \quad (n = 36, r = 0,98, \sigma = 0,39).$$

Вычисление топологических и информационных индексов представляет собой относительно простую задачу ручного счета лишь для малых систем. В тех же случаях когда речь идет о крупных молекулах, содержащих десятки атомов, такой расчет может оказаться трудоемким. К сожалению, авторам известна только одна современная компьютерная программа DRAGON, которая позволяет вычислять более тысячи (!) всевозможных дескрипторов, включая топологические. Немаловажным ее достоинством является то, что она свободно распространяется (<http://www.disat.unimib.it/chm>).

1.4 ЗАДАЧИ

34. Изобразите молекулярные графы изомеров бутана. Для каждого из них вычислите число связей длины “1”, “2” и “3”. По каким топологическим индексам можно различать изомеры бутана?

35. Изобразите молекулярные графы изомеров пентана. Для каждого из них вычислите число связей длины “1”, “2” и “3”. По каким топологическим индексам можно различать изомеры пентана?

36. Изобразите молекулярные графы и вычислите индексы Винера для всех изомеров гексана.

37. Изобразите молекулярные графы и вычислите индексы Винера для изомеров гептана.

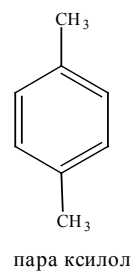
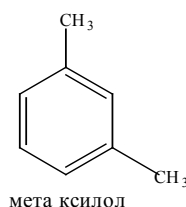
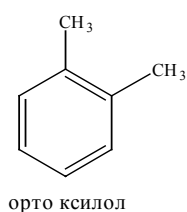
38. Вычислите индексы Рандича для изомеров гексана.

39. Вычислите индексы Рандича для изомеров гептана

40. Вычислите информационные индексы (I_D , IC_0 , IC_1 , IC_2) для изомеров гексана.

41. Вычислите информационные индексы (I_D , IC_0 , IC_1 , IC_2) для двух любых изомеров гептана.

42. Составьте таблицу топологических и информационных индексов для орто-, мета- и пара- ксилола. По каким индексам эти изомеры можно различать?



43. Составьте таблицу топологических и информационных индексов для любых трех изомеров гептана. По каким индексам их можно различать?

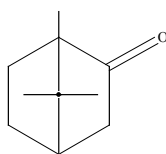
44. Изобразите молекулярные графы любых двух изомеров октана. Вычислите для них индекс Винера, индекс Рандича и информационный индекс (I).

45. Вычислите индексы IC_0 , IC_1 для молекул акролеина и бутадиена.



46. Вычислите индексы IC_0 , IC_1 , SIC_1 , VIC_1 , CIC_1 для насыщенных одноатомных спиртов: бутанол-1, 2-метилпропанол, 3-метилбутанол, пентанол-1, 3-метилбутанол-2.

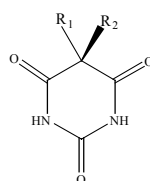
47. Вычислите индексы IC_0 , IC_1 , IC_2 , IC_3 , для молекулы камфоры



48. Вычислите индексы I_D и TI_D для следующих молекул: н-бутан, изобутан, н-пентан, 2-метилбутан, 2,2-диметилпропан, 2-метилпентан, 3-метилпентан, 2,3-диметилбутан, 2,2-диметилбутан.

49. Вычислите информационные индексы для молекул гексана, циклогексана.

50. Вычислите индексы IC_0 , IC_1 , IC_2 а также индексы Рандича ($\chi^{(1)}, \chi^{(2)}$) и Винера (W) для дизамещенных барбитуровой кислоты:



а) барбитал ($R_1 = R_2 = -C_2H_5$), в) фенобарбитал ($R_1 = -C_2H_5$, $R_2 = -C_6H_5$),

с) барбамил ($R_1 = -C_2H_5$, $R_2 = -C_5H_{11}(i)$).

2. РЕГРЕССИОННЫЕ МОДЕЛИ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ

Одним из наиболее популярных подходов к оценке биологической активности молекул является метод, основанный на построении регрессионной модели. Эта модель связывает численное выражение биоактивности (обычно – логарифм количества препарата, приводящего к заданному биоэффекту) и набор выбранных дескрипторов. При этом предполагается, что введенное в биообъект вещество с некоторыми потерями проходит к так называемому *целевому рецептору* – участку организма, воздействие на который и определяет биоэффект. Сам вид функции, связывающей биоактивность или биоэффект с дескрипторами, может варьироваться от линейного до нелинейного типа. Подбираемые

параметры, определяющие вклад дескрипторов, обычно находят с помощью метода наименьших квадратов (МНК).

2.1. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Задача о множественной регрессии заключается в нахождении коэффициентов в уравнении зависимости биоэффекта (y_i) для i -той молекулы от набора дескрипторов, описывающих молекулярную систему ($d_{1i}, d_{2i}, d_{3i}, \dots$):

$$y_i = k_0 + d_{1i} \cdot k_1 + d_{2i} \cdot k_2 + \dots,$$

где k_0, k_1, k_2, \dots – искомые коэффициенты регрессии. На практике нахождение коэффициентов k_i сводится к минимизации следующей величины (метод наименьших квадратов, МНК):

$$Y = \sum_i (y_i(\varepsilon) - y_i)^2,$$

где $y_i(\varepsilon)$ – экспериментальное значение биоэффекта i -той молекулы, а y_i – его теоретическая оценка. Система линейных уравнений для нахождения коэффициентов регрессии в самом общем виде имеет следующий вид:

$$\begin{cases} \frac{\partial Y}{\partial k_0} = 0 \\ \frac{\partial Y}{\partial k_1} = 0 \\ \frac{\partial Y}{\partial k_2} = 0 \\ \dots \end{cases}$$

Качество аппроксимации может быть оценено при помощи дисперсии адекватности:

$$\sigma^2 = \frac{1}{n - m} \sum_i (y_i(\varepsilon) - y_i)^2,$$

и коэффициента корреляции между $y(\varepsilon)$ и y :

$$r = \frac{\sum_i (y_i(\varepsilon) - \bar{y}(\varepsilon))(y_i - \bar{y})}{\sqrt{\sum_i (y_i(\varepsilon) - \bar{y}(\varepsilon))^2 \sum_i (y_i - \bar{y})^2}}$$

где n – число молекул в выборке, m – количество искомых коэффициентов регрессии, а \bar{y} и $\bar{y}(\varepsilon)$ - средние величины (по обучающей выборке) теоретической и экспериментальной биоактивности. Коэффициент r выражает зависимость (корреляцию) экспериментальных данных и их аппроксимацию методом МНК. В литературе по статистике корреляции переменных, обычно, оценивают как

“отличные” $|r| > 0.99$, “хорошие” $0.98 \leq |r| \leq 0.99$,

“удовлетворительные” $0.95 \leq |r| < 0.98$, “плохие” $|r| < 0.9$.

Следует отметить, что такие оценки являются довольно условными.

Для проверки значимости корреляции используют множество критериев. Среди них популярен так называемый F-критерий (предполагается, что задача соответствует нормальному распределению ошибок):

$$F = (n - m) \frac{r^2}{1 - r^2}.$$

С помощью этой величины по специальным таблицам оценивается вероятность того, что между биоэффектом и дескрипторами имеется статистическая связь.

Особенность проблематики QSAR заключается в том, что заранее неизвестно сколько и какие именно дескрипторы необходимы для описания заданного свойства. Поэтому зачастую возникает “соблазн” выбрать очень широкий набор параметров, что приводит к возникновению случайных корреляций. Эта ситуация существенна и тогда, когда число объектов (молекул) сравнительно мало по сравнению с числом дескрипторов. Одним из распространенных подходов в таких ситуациях является метод анализа главных компонент (*principal components analysis*, PCA). В этом методе анализируется структура матрицы корреляции между всеми параметрами задачи с целью идентификации новых переменных, которые суммируют информационное содержание широкого первичного

дескрипторного набора. К сожалению, ограниченный объем пособия не позволяет уделить должного внимания этой важной группе статистических методов.

В данном разделе мы привели лишь наиболее общее описание МНК, поскольку оно имеется в ряде учебников и монографий по статистике, кроме того этот метод и соответствующие оценки значимости включают в себя множество современных пакетов прикладных программ для персонального компьютера (EXEL, CurveExpert, STATISTICA, ORIGIN). Ряд программ разработан также на кафедре технической химии Харьковского национального университета им.В.Н.Каразина.

2.2. ЭМПИРИЧЕСКИЕ КОНСТАНТЫ ЗАМЕСТИТЕЛЕЙ.

УРАВНЕНИЯ ГАММЕТА И ТАФТА

При построении регрессионных зависимостей типа “биоактивность – структура” для рядов соединений, отличающихся только заместителями, удобно использовать подходы с эмпирически подобранными параметрами заместителей. Одним из таких подходов, который базируется на термодинамическом принципе *линейности свободных энергий*, является метод Гаммета. Он применяется для описания π -электронных эффектов заместителей и основан на анализе констант ионизации мета- и пара-замещенных бензойных кислот. Уравнение Гаммета связывает свободную энергию Гиббса для данной реакционной серии (ΔG) с аналогичной величиной, полученной для другой (стандартной) реакционной серии (ΔG_0), с тем же самым набором варьируемых параметров (заместителей):

$$\lg k - \lg k_0 = -\frac{1}{2,303 RT}(\Delta G - \Delta G_0),$$

где k_0 и k – константы скорости реакции соответственно для незамещенной и замещенной молекулы. Предполагая линейную связь между $\Delta G - \Delta G_0$ и параметрами заместителей, получаем *уравнение Гаммета*:

$$\lg k - \lg k_0 = \rho \sigma.$$

В этом уравнении параметр ρ - характеристика данной реакционной серии, а σ - величина, характеризующая заместитель. При этом обычно различают константы заместителей, введенных в пара – σ_p и мета – σ_m положения бензольного кольца. Известны также соответствующие величины для реакций в алифатическом ряду (индукционные постоянные – σ^*).

Таблица 2.1 Эмпирические константы заместителей

Заместитель	σ_p	σ_m	σ^*	E_s
ь				
-NH ₂	-0,660	-0,161	-	-
-N(CH ₃) ₂	-0,83	-0,211	-	-
-OCH ₃	-0,268	0,115	1,450	0,97
-OC ₂ H ₅	-0,24	0,1	1,366	0,86
-CH ₃	-0,170	-0,069	0	0
-C ₂ H ₅	-0,151	-0,07	-0,100	-0,27
-C ₃ H ₇ (i)	-0,126	-	-0,115	-0,56
-C(CH ₃) ₃	-0,197	-0,10	-0,300	-2,14
-SCH ₃	-0,047	0,15	-	-
-H	0	0	0,	1,24
-C ₆ H ₅	0,009	0,218	0,600	-0,90
-F	0,062	0,337	3,08	0,49
-Cl	0,227	0,373	2,92	0,18
-Br	0,232	0,391	2,78	0,06
-I	0,18	0,352	2,36	-0,20
-COCH ₃	0,502	0,376	1,65	-
-COOH	0,265	0,355	2,1	-
-CN	0,66	0,56	3,6	-
-NO ₂	0,778	0,710	3,9	-0,75

Однако, подчеркнем, что уравнение Гаммета описывает лишь электронные эффекты взаимодействия заместителя с реакционным центром. В тоже время известно много реакций, константы которых зависят также от пространственных (стерических) особенностей заместителей. В связи с

этим при изучении кислотного гидролиза эфиров Тафтом было получено выражение, по форме аналогичное уравнению Гаммета:

$$\lg k - \lg k_0 = \delta E_s,$$

в котором E_s – некоторая пространственная характеристика заместителя, а δ – мера чувствительности данной реакционной серии.

Параметры наиболее распространенных заместителей приведены в табл. 2.1. Эти параметры, как характеристики электронных и стерических эффектов заместителей, широко используются в качестве дескрипторов регрессионных моделей биологической активности.

2.3. АДДИТИВНАЯ МОДЕЛЬ ФРИ-ВИЛЬСОНА

В модели биологической активности по Фри-Вильсону рассматривается ряд соединений, у которых заместители находятся в различных положениях. Тогда, предполагая вклад от заместителей аддитивным, уравнение для биоактивности молекулы i (y_i) можно записать следующим образом:

$$y_i = \sum_i n_{ip} a_{ip} + y_0,$$

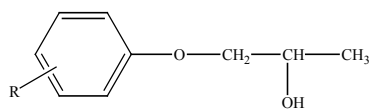
где a_{ip} – вклад в общую активность i -того заместителя, находящегося в p -том положении, n_{ip} – количество заместителей типа i в положении p (обычно 0 или 1, редко – 2), а y_0 – биоактивность незамещенной молекулы.

Численные значения коэффициентов a_{ip} и y_0 находят с помощью МНК.

При наличии в исследуемых системах некоторого набора заместителей в определенных положениях удобно построить *матрицу Фри-Вильсона*, отображающую присутствие (или отсутствие) соответствующих заместителей. Так, в системе производных феноксипропанолов, обладающей фунгистатическими свойствами¹ (QSAR,

¹ Фунгистат – химическое соединение, ингибирующее рост и размножение грибов.

2002, **20**, 422) возможны замещенные (-CH₃, -Cl) в орто-, мета-, и пара-положениях:



Соответствующая матрица Фри-Вильсона приведена в табл.2.2.

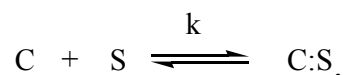
Таблица 2.2 Матрица Фри-Вильсона для производных феноксипропанолов

№	Заместители R	Орто			Мета		Пара			Активность
		H	CH ₃	Cl	H	CH ₃	H	CH ₃	Cl	
1	2-метил	1	1	0	2	0	1	0	0	2,46
2	2-хлор	1	0	1	2	0	1	0	0	2,84
3	4-хлор	2	0	0	2	0	0	0	1	2,81
4	2,6-дихлор	0	0	2	2	0	1	0	0	3,04
5	2,4-дихлор	1	0	1	2	0	0	0	1	3,35
6	2-метил-4-хлор	1	1	0	2	0	0	0	1	3,30
7	3-метил-4-хлор	2	0	0	1	1	0	0	1	3,30
8	2-метил-6-хлор	0	1	1	2	0	1	0	0	2,70
9	2,6-диметил-4-хлор	0	2	0	2	0	0	0	1	3,51
10	3,5-диметил-4-хлор	2	0	0	0	2	0	0	1	3,68
11	2,6-дихлор-4-метил	0	0	2	2	0	0	1	0	3,47

Обработка этого массива данных методом наименьших квадратов позволяет получить искомую функциональную зависимость активности от параметров структуры (см. задачу 58).

2.4. МЕТОД ХЭНЧА. ОЦЕНКА ЛИПОФИЛЬНОСТИ

В методе Хэнча рассматривается фармакокинетическая модель, согласно которой биологически активное вещество (С), взаимодействуя с целевым рецептором биообъекта (S), порождает соответствующий комплекс (C:S). В результате устанавливается следующее равновесие:



В соответствии с законами химической кинетики константа равновесия К выражается через отношение концентраций:

$$k = \frac{[C:S]}{[C][S]}.$$

Очевидно, что концентрация вещества С (см. Приложение) вблизи рецептора пропорциональна его количеству, введенному в биообъект C_0 :

$$[C] = A \cdot C_0.$$

Поэтому концентрация C_0 (обычно рассматривается логарифм обратной величины – $\lg 1/C_0$) может служить мерой биологической активности препарата. Так, если заданный или стандартный биоэффект (например, на 50% замедляется рост опухоли) наступает при очень малых значениях C_0 , то препарат высокоактивен, и наоборот, – большие значения C_0 свидетельствуют о его малой активности. Очевидно, что коэффициент пропорциональности А определенным образом должен быть связан с параметрами, характеризующими транспортировку препарата через клеточные мембраны к целевому рецептору. При этом, поскольку препарату на своем пути приходится преодолевать ряд водно-липидных слоев, в качестве такого параметра можно использовать величину, которая характеризует распределение вещества в различных фазах. В настоящее время для этой цели принято использовать *липофильность* – логарифм

коэффициента распределения вещества в смеси октанол-вода $\lg P$ (см. Приложение). Здесь величина P выражается через равновесные концентрации вещества в н-октаноле ($C_{\text{октанол}}$) и воде ($C_{\text{вода}}$):

$$P = \frac{C_{\text{октанол}}}{C_{\text{вода}}}.$$

Для оценки $\lg P$ существует множество теоретических подходов. Среди них особо популярным является аддитивный метод, в котором предполагается, что общая липофильность молекулы может быть разложена на структурные составляющие. Одна из первых таких схем, позволяющая оценивать липофильность заместителей, была предложена самим Хэнчем. Она основывалась на следующем соображении: если липофильность соединения R-X равна $\lg P_{RX}$, а незамещенной молекулы R-H – $\lg P_{RH}$, то липофильность заместителя π_X равна

$$\pi_X = \lg P_{RX} - \lg P_{RH}.$$

Следуя аддитивной модели, общая величина $\lg P$ должна выражаться через липофильности составляющих π

$$\lg P = \sum_i n_i \pi_i,$$

где n_i – число структурных фрагментов типа i для которых известна липофильность π_i .

Рассмотрим, каким образом $\lg P$ связан с биоактивностью в рамках фармакокинетической модели. Ответ на этот вопрос дает компьютерное моделирование прохождения вещества через множество водно-липидных слоев. Численное решение системы кинетических уравнений описывающих такие процессы приводит к так называемому

Таблица 2.3 Константы липофильности распространенных заместителей (по Хэнчу) в ароматическом и алифатическом радикалах

№	Заместитель	Заместитель в	
		ароматическом кольце $\pi_X = \lg P_{C_6H_5X} - \lg P_{C_6H_6}$	алифатическом фрагменте $\pi_X = \lg P_{RX} - \lg P_{RH}$
1	-NH ₂	-1,23	-1,19
2	-I	1,12	1,0
3	-SCH ₃	0,61	0,45
4	-COCH ₃	-0,55	-0,71
5	-CONH ₂	-1,49	-1,71
6	-COOCH ₃	-0,01	-0,27
7	-Br	0,86	0,60
8	-CN	-0,57	-0,84
9	-F	0,14	-0,17
10	-Cl	0,71	0,39
11	-COOH	-0,28	-0,67
12	-OCH ₃	-0,02	-0,47
13	-OC ₆ H ₅	2,08	1,61
14	-N(CH ₃) ₂	-0,18	-0,30
15	-OH	-0,67	-1,16
16	-NO ₂	-0,28	-0,85

параболическому типу зависимости $\lg \frac{1}{C_0}$ от $\lg P$:

$$\lg 1/C_0 \sim a (\lg P)^2 + b \lg P.$$

Типичный график такой зависимости показан на Рис. 1

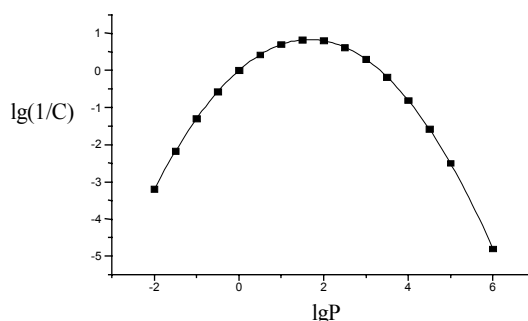


Рис. 1 Параболический тип зависимости активности $\lg(1/C)$ от липофильности $\lg P$

Очевидно, что при определенных значениях $\lg P$ достигается максимум биологической активности препарата $\lg 1/C_0$. Это обстоятельство широко используется при разработке новых биоактивных препаратов. Так, например, анализ экспериментальных данных говорит о том, что соотношение $\lg P \sim 2$ является оптимальным для того, чтобы лекарство легко достигало головного мозга. Иными словами, при данной величине липофильности концентрация препарата в головном мозге биообъекта максимальна. Соответствующие оптимальные величины $\lg P$ известны и для достижения других органов – печени, почек и т.д.

Параболическая зависимость $\lg 1/C_0$ от $\lg P$, не является единственной. Большое распространение в последнее время получила также *билинейная модель*:

$$\lg 1/C_0 \sim a \lg P + b \lg(\beta P + 1),$$

в которой числа a , b , и β - подгоночные параметры (Рис. 2).

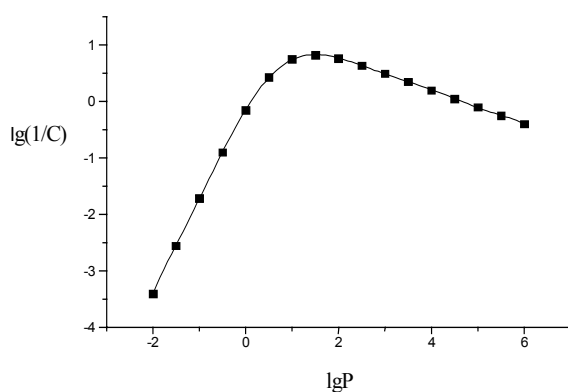


Рис. 2 Билинейная модель связи биоактивности и липофильности

Из рис.2 следует, что особенностью билинейной модели является наличие двух линейных участков вдали от максимума, что хорошо согласуется с множеством экспериментальными данными.

Кроме собственно липофильности ($\lg P$) в расчетах QSAR часто используются соответствующие (подкорректированные) величины для

катионной, анионной и цвиттерионной структур ($\lg P^+$, $\lg P^-$, $\lg P^\pm$). Если для препарата (ХН) предполагается ионное равновесие в растворе:



то, очевидно, что его распределение между несмешивающимися фазами, зависит от рК соединения и рН среды. В этом случае, вместо $\lg P$ целесообразно использовать величину

$$\lg D = \lg P - \lg(1 + 10^{-\text{pK} + \text{pH}}).$$

Соответствующие выражения получены и для других типов ионных равновесий.

Однако, не только величина липофильности оказывает влияние на эффективность воздействия вещества на организм. Очевидно, что взаимодействие молекулы с рецептором должно быть связано с параметрами электронной структуры молекулы (электронная плотность, дипольный момент, сверхделокализуемость и т.д.). Если речь идет о ряде соединений различающихся набором заместителей, то ясно, что биологическое действие может быть выражено (в стиле уравнения Гаммета) через эмпирические константы заместителей, описанные в разделе 2.2.

Таким образом, учитывая перечисленные выше факторы, можно записать *общее уравнение Хэнча* (с параболической зависимостью от $\lg P$), описывающее биоактивность препарата:

$$\lg 1/C_0 = a (\lg P)^2 + b \lg P + c \sum \sigma_i + d \sum E_{s_i} + \dots + \text{const},$$

где варьируемые параметры (a , b , c , d, \dots, const) могут быть найдены с помощью МНК-процедуры. Если вместо общей липофильности системы целесообразно использовать только вклады от заместителей (π_x), это уравнение приобретает вид:

$$\lg 1/C_0 = a \left(\sum \pi_i \right)^2 + b \sum \pi_i + c \sum \sigma_i + d \sum E_{s_i} + \dots + \text{const}.$$

Отметим также, что теоретическая оценка $\lg P$ для молекулярной системы может быть получена с использованием ряда пакетов прикладных программ, среди которых HyperChem, ACDLabs, DRAGON и т.д.

2.5. ЗАДАЧИ

51. Постройте регрессионную зависимость токсичности вещества (LD_{50}) от топологических параметров заместителя по данным табл. 2.4.

Таблица 2.4 Зависимость токсичности (LD_{50}) некоего вещества от заместителя

№	Заместитель	LD_{50} (мк/кг)	№	Заместитель	LD_{50} (мк/кг)
1	CH_3CH_2-	544	6	$(CH_3)_2CHCH_2-$	206
2	$CH_3CH_2CH_2-$	428	7	$(CH_3)_3C-$	197
3	CH_3-	644	8	$CH_3(CH_2)_4-$	150
4	$(CH_3)_2CH-$	301	9	$(CH_3)_3CCH_2-$	236
5	$CH_3CH_2CH_2CH_2-$	258	10	$CH_3(CH_2)_4CH_2-$	43

Используя коэффициент корреляции и дисперсию, оцените качество следующих наборов параметров: а) N_C, W, I_D ; б) IC_0, IC_1, IC_2 ; в) P_1, P_2, P_3 .

52. Известно, что аккумуляцию биологически активных веществ можно прогнозировать на основании данных об их растворимости в воде. Используя данные табл. 2.5 для ряда насыщенных спиртов, постройте функцию растворимости $\lg X$ от молекулярных параметров (N_C, IC_1, I_D). Оцените $\lg X$ и сравните расчетные данные с экспериментом для 3-метилбутанола-2 ($-\lg X = 1,926$) и пентанола ($-\lg X = 2,332$).

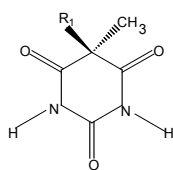
Таблица 2.5. Растворимость в воде насыщенных спиртов

№	Соединение	Растворимость в мольных долях, -lgX	№	Соединение	Растворимость в мольных долях, -lgX
1	бутанол-1	1,750	6	пентанол-2	2,025
2	2- метилпропанол	1,743	7	пентанол-3	1,961
3	бутанол-2	1,724	8	гексанол-1	2,957
4	3-метилбутанол	2,254	9	гексанол-2	2,961
5	2-метилбутанол	2,207	10	гексанол-3	2,542

53. Производные барбитуровой кислоты обладают снотворным действием различной продолжительности (данные приведены в табл.2.6). Вычислите функцию, описывающую длительность действия барбитуратов в зависимости от параметров заместителя. Используйте следующие наборы топологических индексов (для каждого оцените качество аппроксимации):

- а) P_1, P_2, P_3 ; б) $N_C, \chi^{(1)}, I_D$; в) P_1, P_2, P_3, I_D .

Таблица 2.6 Длительность снотворного действия барбитуратов



№	R ₁	Структура	Длительность
		заместителя R ₁	действия (мин.)
1	C ₂ H ₅ -	CH ₃ CH ₂ -	1400
2	C ₃ H ₇ -	CH ₃ CH ₂ CH ₂ -	1140
3	C ₃ H ₇ -	CH ₃ CH(CH ₃)-	1520
4	C ₄ H ₉ -	CH ₃ CH ₂ CH ₂ CH ₂ -	450
5	C ₄ H ₉ -	CH ₃ CH(CH ₃)CH ₂ -	540
6	C ₄ H ₉ -	CH ₃ CH ₂ CH(CH ₃)-	600
7	C ₅ H ₁₁ -	CH ₃ CH ₂ CH ₂ CH ₂ CH ₂ -	220
8	C ₅ H ₁₁ -	CH ₃ CH ₂ CH(CH ₃)CH ₂ -	190
9	C ₅ H ₁₁ -	(CH ₃) ₃ CCH ₂ -	200
10	C ₅ H ₁₁ -	CH ₃ CH ₂ CH(CH ₃ CH ₂)-	300
11	C ₆ H ₁₃ -	CH ₃ (CH ₂) ₅ -	45
12	C ₆ H ₁₃ -	CH ₃ (CH ₂) ₂ CH(CH ₃)CH ₂	210
		-	
13	C ₆ H ₁₃ -	CH ₃ CH ₂ C(CH ₃) ₂ CH ₂ -	60

54. Алифатические спирты ингибируют микросомальное пара-гидроксилирование анилина цитохромом. Вычислите линейную зависимость ингибиторной способности спирта pIC_{50}^2 от различных наборов теоретико-информационных индексов, если известны следующие данные для различных одноатомных спиртов (табл.2.7).

**Таблица 2.7 Ингибиторная способность спиртов при пара-
гидроксилировании анилина**

Соединение	pIC ₅₀	Соединение	pIC ₅₀
Метанол	-3,09	2,2-диметилпропанол-1	-0,67
Этанол	-1,10	Пропанол-2	-0,47
Пропанол-1	-0,48	Бутанол-2	-0,35
Бутанол-1	-0,05	Пентанол-2	-0,07
Пентанол-1	0,27	Гексанол-2	0,15
Гексанол-1	0,54	Гептанол-2	0,25
Гептанол-1	0,68	Пентанол-3	-0,37
2-метилпропанол-1	-0,39	Гексанол-3	-0,47
2-метилбутанол-1	-0,15	2-метилпентанол-3	-0,89
3-метилбутанол-1	-0,19	2,4-диметилпентанол-3	-1,38

В качестве дескрипторов используйте следующие наборы:

а) IC₀, IC₁, IC₂, б) IC₀, SIC₁, SIC₂, в) TIC₀, TIC₁, TIC₂

Какой из этих наборов удачнее описывает активность?

55. Токсичность 5-(диалкилвинил)-5-алкил производных барбитуровой кислоты LD₅₀ (мк/кг) в зависимости от характера заместителя представлена в табл. 2.8. Предполагая, что функция, описывающая токсичность LD₅₀ (мк/кг), в зависимости от структуры молекулы имеет вид:

$$LD_{50} = A + B \cdot X + C \cdot X^2,$$

² pIC₅₀ – отрицательный логарифм концентрации спирта (в миллимолях) приводящий к 50% ингибированию гидроксилирования анилина.

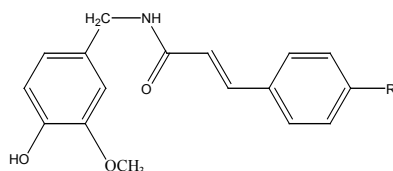
вычислите коэффициенты А, В и С методом МНК, если параметр Х равен: а) IC_0 , б) IC_1 , в) SIC_1 , г) SIC_2 , д) CIC_2 , е) индекс Рандича, ж) индекс Винера. Сравните качество описания токсичности LD_{50} с помощью этих дескрипторов.

Таблица 2.8. Токсичность производных барбитуратов

Диалкилвинил	Алкил	LD_{50} (мк/кг)
EtCH=C(Me)-	Me-	500
EtCH=C(Me)-	Et-	180
EtCH=C(Me)-	Pr-	270
EtCH=C(Me)-	i-Pr-	100
MeCH=C(Et)-	Me-	580
MeCH=C(Et)-	Et-	280
MeCH=C(Et)-	Pr-	270
MeCH=C(Et)-	i-Pr-	155
PrCH=C(Me)-	Me-	375
PrCH=C(Me)-	Et-	160
i-PrCH=C(Me)-	Me-	200
BuCH=C(Me)-	Me-	360
BuCH=C(Me)-	Et-	130
EtCH=C(Pr)-	Et-	130

56. Синтетические аналоги капсаицина³

³ Природный алкалоид капсаицин (амид дециленовой кислоты) - выделяют из перца.



обладают анальгетическим действием (J.Med.Chem., 1993, **36**, 2381). В табл. 2.9 приведено численное выражение биоактивности $\lg EC_{50}$ совместно с некоторыми параметрами заместителей (σ_p – константа Гаммета, MR_R – молекулярная рефракция, π_R – липофильность). Постройте линейно-регрессионную модель зависимости биоактивности от параметра заместителя. Какой из параметров лучше описывает активность? Сделайте прогноз биологической активности соединения с $R = -COCH_3$. Для *параболической* и *билинейной* модели Хэнча найдите коэффициенты регрессии. Какая из этих моделей лучше описывает биологический отклик? Используя σ_p константы заместителя (табл. 2.1) сделайте прогноз биоактивности для соединений с $R = -SCH_3$, $-C_6H_5$, $-F$.

Таблица 2.9 Биологическая активность аналогов капсаицина и параметры заместителя

R	σ_p	MR_R	π_R	$\lg EC_{50}$
-H	0	1,03	0	1,07
-Cl	0,227	6,03	0,77	0,09
-NO ₂	0,778	7,36	-0,28	0,66
-CN	0,66	6,33	-0,57	1,42
-C ₆ H ₅	0,009	25,36	1,96	-0,62
-N(CH ₃) ₂	-0,205	15,55	0,18	0,64
-I	0,278	13,94	1,12	-0,46
-COCH ₃	0,502	10,77	-0,55	?

57. Постройте матрицу Фри-Вильсона для набора заместителей R_1 и R_2 , приведенных в табл. 2.10 (QSAR, 2002, **20**, 422). Вычислите

коэффициенты a_{ip} и Y_0 линейной регрессии. Какова биоактивность дифторпроизводного? Постройте линейную

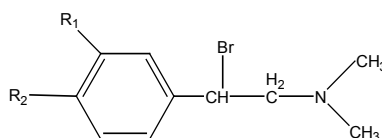
$$\lg C = a(\pi_{R_1} + \pi_{R_2}) + b(\sigma_{R_1} + \sigma_{R_2}) + \text{const}$$

и параболическую

$$\lg C = a(\pi_{R_1} + \pi_{R_2}) + b(\pi_{R_1} + \pi_{R_2})^2 + c(\sigma_{R_1} + \sigma_{R_2}) + \text{const}$$

зависимости $\lg C$ от парциальных липофильностей (π_{R_1}, π_{R_2}) и электронных констант заместителей.

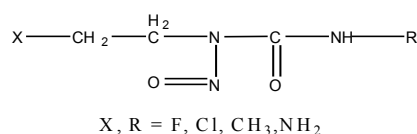
Таблица 2.10 Адреноблокирующие свойства в ряду замещенных



R_1	R_2	Активность -lgC	R_1	R_2	Активность -lgC
-H	-F	8,16	-Br	-F	8,57
-H	-Cl	8,68	-CH ₃	-F	8,82
-H	-Br	8,89	-Cl	-Cl	8,89
-H	-I	9,25	-Br	-Cl	8,92
-H	-CH ₃	9,30	-CH ₃	-Cl	8,96
-F	-H	7,52	-Cl	-Br	9,00
-Cl	-H	8,16	-Br	-Br	9,35
-Br	-H	8,30	-CH ₃	-Br	9,22
-I	-H	8,40	-CH ₃	-CH ₃	9,30
-CH ₃	-H	8,46	-Br	-CH ₃	9,52
-Cl	-F	8,19	-H	-H	7,46

58. Для матрицы Фри-Вильсона, приведенной в разделе 2.2 (табл.2.2) постройте линейную регрессионную модель. Какова биоактивность незамещенной молекулы ?

59. Активность производных нитрозомочевины:



против L-1210 лейкемии мышей выражается следующей формулой:

$$\lg \frac{1}{C_0} = -0,0568 (\lg P)^2 + 0,0689 \lg P + 4,527. \quad (N=22, r=0,922, \sigma=0,163)$$

При каких значениях липофильности можно ожидать максимум биологической активности? Постройте график зависимости теоретического значения $\lg 1/C_0$ от липофильности.

3. СТАТИСТИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ МОЛЕКУЛ ПО ИХ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ

В связи с успехами органического (комбинаторного) синтеза, позволяющего получать огромное количество новых структур, возрастает потребность проведения систематического *скрининга* (просеивания) соединений с целью отбора наиболее перспективных в плане их активности, дальнейшего исследования и, затем, применения. Это приводит к использованию статистических методов классификации структур по признакам активные - неактивные (слабоактивные). В наиболее распространенных методах классификации предполагается наличие определенной обучающей выборки, с помощью которой проводится “настройка” математической модели с целью явного или не явного выявления структурных параметров (или их совокупности), важных в контексте заданного типа активности. Популярными в связи с этим являются *дискриминационный* и *кластерный* анализы – методы ориентированные на многомерное исследование заданного набора структур, как совокупности параметров. Определенный интерес представляет также так называемый *логико-комбинаторный* подход, позволяющий в конкретных структурно-химических терминах описать и прогнозировать биоактивность молекул.

3.1. ЛОГИКО-КОМБИНАТОРНЫЙ ПОДХОД

Логико-комбинаторный подход (Розенблит, Голиндер, 1984) базируется на некоторых результатах *формальной логики* и статистики. Рассмотрим наиболее важные в контексте этой проблемы логические понятия.

1. *Сходство*. Если два или больше исследуемых случаев имеют общим одно обстоятельство, то это обстоятельство, с которым согласуются все явления, есть причина (или следствие) данного феномена. Схематически это можно выразить следующим образом. Некоторая структура имеет компоненты ABCD и обладает свойством X: $ABCD \rightarrow X$. Другая структура имеет компоненты AKLM и обладает тем же свойством X: $AKLM \rightarrow X$. Отсюда следует, что данное свойство связано с A: $A \rightarrow X$,

$$\frac{\begin{array}{l} ABCD \rightarrow X \\ AKLM \rightarrow X \end{array}}{A \rightarrow X}$$

2. *Различие*. Если случай, при котором данный феномен проявляется, и случай когда он не проявляется, схожи во всех обстоятельствах кроме одного, появляющегося только в первом случае, то это обстоятельство (которым только и различаются оба случая) и есть причина (или следствие) феномена.

$$\frac{\begin{array}{l} ABCD \rightarrow X \\ BCD \rightarrow Y \end{array}}{A \rightarrow X}$$

Эти два утверждения и являются основой для *логико-комбинаторного* метода анализа биоактивности. Перечислим его основные моменты:

1. Структура каждой молекулы μ описывается с помощью специального *вектора-описателя* S_μ .

2. Соединения обучающей выборки классифицируются по типам активности на N_k классов $A_k, k=1-N_k$. В самом простом случае предусматривается разбиение выборки на два класса *неактивные (или слабоактивные) – активные*.
3. *Признаки активности* $f_r(S)$ (или неактивности) определяются как некоторые фрагменты структуры, выраженные с помощью вектора-описателя S_μ , присутствие которых в молекуле с высокой вероятностью указывает на то, что соединение проявляет (или не проявляет) данную активность. Такие потенциальные признаки находят путем сравнения пар молекул внутри каждого класса с последующим выделением общего фрагмента:

$$f_r(S) = S_\mu \cap S_\nu.$$

4. Подсчитывается число появлений этого признака L_{kr} в каждом из классов N_k , а также суммарное количество проявлений этого признака во всех классах – L_r .

$$L_r = \sum_{k=1}^{N_k} L_{kr}$$

5. Проводится статистическая оценка проявления этих признаков для каждого класса. Поскольку точные значения вероятностей того, что соединение, обладающее данным признаком $f_r(S)$ относится к классу A_k неизвестны, для их оценки можно воспользоваться *методом Байеса*. В этом методе вероятности вычисляются так, что возможный ущерб от неверной оценки минимизируется. В случае разбиения массива объектов на два класса (активные — неактивные) формула для вероятности отнесения к классу A_k записывается очень просто:

$$P(A_k | f_r) = \frac{L_{kr} + 1}{L_r + 2}.$$

При этом *прогностическая ценность признака* определяется как

$$U_{kr} = \frac{P(A_k | f_r)}{P(A_k)}$$

Прогностическая ценность (U_{kr}) показывает, во сколько раз увеличиваются шансы нахождения соединения с заданным свойством в сравнении с методом простого перебора.

Проиллюстрируем этот метод следующим примером. Предположим, что структура соединений различается только заместителями R_1 , R_2 , R_3 и R_4 . При этом соединения имеют различную активность I (табл. 3.1).

Таблица 3.1 Терапевтический эффект (I) в различных замещенных

№	R_1	R_2	R_3	R_4	LD_{50}	ED_{50}	I
1	H	CH_3	Br	NH_2	5000	400	12,5
2	CH_3	NH_2	H	C_2H_5	2000	150	13,33
3	H	C_6H_5	Br	C_2H_5	3500	300	11,67
4	CH_3	NH_2	CH_3	NO_2	1800	100	18,00
5	H	NH_2	Cl	NO_2	2000	800	2,50
6	CH_3	C_6H_5	H	NH_2	1000	500	2,00
7	H	CH_3	CH_3	NO_2	70	40	1,75
8	CH_3	C_6H_5	Br	C_2H_5	200	70	2,86

Здесь указан $I = \frac{LD_{50}}{ED_{50}}$ – так называемый терапевтический эффект – величина, которая характеризует эффективность данного препарата. Основным структурным признаком в этой задаче для молекулы μ можно считать вектор S_μ :

$$S_\mu \equiv (R_1, R_2, R_3, R_4)$$

Так, для молекулы № 1 вектором-определителем является

$$S_1 \equiv (\text{H}, \text{CH}_3, \text{Br}, \text{NH}_2),$$

а для молекулы № 2 –

$$S_2 \equiv (\text{CH}_3, \text{NH}_2, \text{H}, \text{C}_2\text{H}_5).$$

Разделим весь набор по признаку I на два класса – активные молекулы (№ 1- №4, $11,67 \leq I \leq 18,0$) и неактивные (№ 5- № 8, $1,75 \leq I \leq 2,86$).

Далее проводим отбор признаков специфичных для данного класса. Например, векторы S_1 и S_3 характеризуют структуры активных молекул, а их пересечение $S_1 \cap S_3$ может служить ключевым признаком для классификации молекул:

$$1 \text{ вектор, } S_1: (\text{H}, \text{CH}_3, \text{Br}, \text{NH}_2)$$

$$2 \text{ вектор, } S_3: (\text{H}, \text{C}_6\text{H}_5, \text{Br}, \text{C}_2\text{H}_5)$$

$$\text{Пересечение векторов.} \quad (\text{H}, \emptyset, \text{Br}, \emptyset)$$

$$S_1 \cap S_3:$$

Здесь \emptyset – символ пустого множества. Признак $S_1 \cap S_3$ соответствует условию $(R_1=\text{H}) \& (R_3=\text{Br})$, где $\&$ – логическое “И”. Последовательное попарное сравнение структурных векторов S дает следующий набор признаков.

Таблица 3.2 Общий набор признаков

Номер признака r	Признак $f_r(S)$	L_{1r}	L_{2r}	$P(A_1 f_r)$	$P(A_2 f_r)$
1	$(R_1=H) \& (R_3=Br)$	2	0	0,75	0,25
2	$(R_4=C_2H_5)$	2	1	0,60	0,40
3	$(R_1=CH_3) \& (R_2=NH_2)$	2	0	0,75	0,25
4	$(R_1=H) \& (R_4=NO_2)$	0	2	0,25	0,75
5	$(R_1=CH_3) \& (R_2=C_6H_5)$	0	2	0,25	0,75

В табл. 3.2 кроме признаков приведены также числа их появлений в классе активных молекул L_{1r} и неактивных молекул L_{2r} . Соответствующие вероятности $P(A_1 | f_r)$ и $P(A_2 | f_r)$ дают оценку успешности классификации. Так, согласно этим данным, наличие признаков 1 или 3 с вероятностью 0,75 позволяют отнести неизвестную молекулу к классу активных систем. Признаки 4 и 5 с той же вероятностью позволяют отнести систему к неактивным молекулам.

Используя таблицу признаков, можно оценить активности новых соединений, которые не вошли в исходный набор, описывающий терапевтический эффект (табл. 3.3).

Таблица 3.3 Оценка активности новых соединений

№	R_1	R_2	R_3	R_4	Активность
9	CH_3	NH_2	Br	NH_2	Активное (признак 3)
10	CH_3	C_6H_5	H	C_2H_5	Неактивное (признак 5)
11	CH_3	CH_3	H	H	Отказ от прогноза
12	H	H	H	C_2H_5	Активное с малой вероятностью (признак 2)

Очевидно, что соединения № 9 и № 10 уверенно классифицируются как соответственно активное и неактивное. Соединение № 11 на основании пяти перечисленных признаков не может быть надежно классифицировано. Соединение № 12 (с вероятностью 0,60) также может быть отнесено к активным.

3.2. КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ связи биоактивности и структуры основывается на геометрическом представлении молекул в многомерном пространстве дескрипторов. Иными словами, набор из N дескрипторов ($d_i, i = 1, \dots, N$), которые однозначно характеризуют молекулу, можно интерпретировать как совокупность координат некой условной точки. Таким образом, молекула – это точка в пространстве N измерений. В качестве меры сходства (или близости) между разными молекулами можно принять расстояние R между соответствующими точками. Однако, поскольку это пространство не является реальным, физическим, то таковой метрикой могут служить самые разные функции, удовлетворяющие аксиомам расстояния. Эти аксиомы для любых трех объектов (молекул) X_i, X_j, X_k формулируются следующим образом:

1. $R(X_i, X_j) \geq 0$ для всех X_i, X_j из данного набора молекул.
2. $R(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$ т.е. когда речь идет об одной и той-же молекуле (неразличимость идентичных объектов).
3. $R(X_i, X_j) = R(X_j, X_i)$. (симметрия)
4. $R(X_i, X_j) \leq R(X_i, X_k) + R(X_k, X_j)$. (неравенство треугольника)

Несколько наиболее популярных типов расстояний приведены в табл. 3.4.

Естественно предположить, что ряд молекул-точек, обладающих близкими характеристиками (например, активные в отношении данного свойства) будут находиться на достаточно близком друг от друга расстоянии. Неактивные же молекулы будут образовывать отдельную группу. Такие группы принято называть *кластерами*, а соответствующий анализ данных, позволяющий разбивать совокупность объектов на отдельные кластеры, *кластерным анализом*.

Таблица 3.4. Метрические характеристики близости объектов (расстояния)

Название	Формула для вычисления расстояния
Евклидово расстояние	$R_2(X_i, X_j) = \sqrt{\sum_k^p (d_{ki} - d_{kj})^2}$
Линейное расстояние	$R_1(X_i, X_j) = \sum_k^p d_{ki} - d_{kj} $
Обобщенное расстояние Минковского	$R_n(X_i, X_j) = \sqrt[n]{\sum_k^p (d_{ki} - d_{kj})^n}$
Расстояние Махаланобиса ⁴	$R_m(X_i, X_j) = \sqrt{(d_{ki} - d_{kj})^+ W^{-1} (d_{ki} - d_{kj})}$

Таким образом главная цель кластерного анализа – нахождение групп (кластеров) схожих объектов. Распространенным способом представления результатов кластерных методов является дендрограмма (древовидная диаграмма), которая графически изображает иерархическую структуру, порожденную объединением объектов в кластеры. На рис. 3 показан пример дендрограммы отображающей разбиение трех объектов (молекул) на два кластера. Структуры А и В относятся к одному кластеру, тогда как структура С к другому.

⁴ W – внутригрупповая ковариационная матрица. См раздел 3.3, ⁺ – операция транспонирования

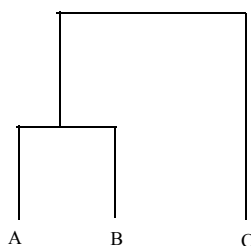


Рис. 3 Пример иерархической дендрогаммы

Различные варианты кластерного анализа программно реализованы в множестве статистических пакетов прикладных программ. (напр. STATISTICA).

3.3. ЛИНЕЙНЫЙ ДИСКРИМИНАЦИОННЫЙ АНАЛИЗ

Целью *линейного дискриминационного анализа* является вычисление функции, позволяющей классифицировать системы по заранее заданным группам. Чаще всего встречается необходимость классификации молекул на две группы (активные – неактивные). При этом функция D выбирается так, чтобы ее значения максимально различались для типичных представителей разных групп (активных – неактивных молекул). Наиболее распространенный, линейный вариант дискриминационной функции, строится как суперпозиция молекулярных дескрипторов (d_1, d_2, d_3, \dots), и значение D-функции для i -той молекулы выглядит следующим образом:

$$D_i = k_0 + d_{1i} \cdot k_1 + d_{2i} \cdot k_2 + \dots$$

Искомые коэффициенты разложения D: k_0, k_1, k_2, \dots (не путайте их с коэффициентами линейного регрессионного анализа !) можно вычислить, максимизировав отношение

$$\lambda(k) = \frac{\sum_{i,j} k_i T_{ij} k_j}{\sum_{i,j} k_i W_{ij} k_j},$$

где числитель характеризует межгрупповой разброс молекулярных параметров (дескрипторов), а знаменатель – разброс параметров внутри

групп. Эти величины построены с помощью ковариационных матриц (Т и W), описывающих соответствующие дисперсии:

$$W_{ij} = \sum_{\alpha=1}^2 \sum_{m=1}^{n_{\alpha}} (d_{i\alpha m} - \bar{d}_{i\alpha})(d_{j\alpha m} - \bar{d}_{j\alpha}),$$

$$T_{ij} = C_{ij} - W_{ij},$$

$$C_{ij} = \sum_{\alpha=1}^2 \sum_{m=1}^{n_{\alpha}} (d_{i\alpha m} - \bar{d}_i)(d_{j\alpha m} - \bar{d}_j).$$

При этом матрица С отвечает общей дисперсии дескрипторов. В этих выражениях предполагается, что \bar{d}_i – усредненная по всем (двум) группам величина дескриптора d_i , $\bar{d}_{i\alpha}$ – средняя величина дескриптора d_i по группе α . Величина $d_{i\alpha m}$ – i -тый дескриптор молекулы m , которая относится к группе α , а n_{α} – количество молекул в группе α . Таким образом, основная задача дискриминационного анализа состоит в нахождении наилучшего разделения молекул на группы в том смысле, что межгрупповая дисперсия должна быть максимальной, а внутригрупповая – минимальной.

С геометрической точки зрения дискриминационный метод близок к основам кластерного анализа. Каждая молекула представляется в виде точки в условном пространстве дескрипторов d . Тогда, при удачном выборе дескрипторного набора, молекулы, относящиеся к данной группе активности, будут находиться достаточно “близко” друг к другу (кластеризоваться). Вычислив значение D-функции для молекулы с неизвестной активностью (x), можно выяснить, к какому кластеру она относится и, следовательно, оценить ее активность. Схематически группы активных и неактивных молекул можно представить в виде рисунка (Рис.4).

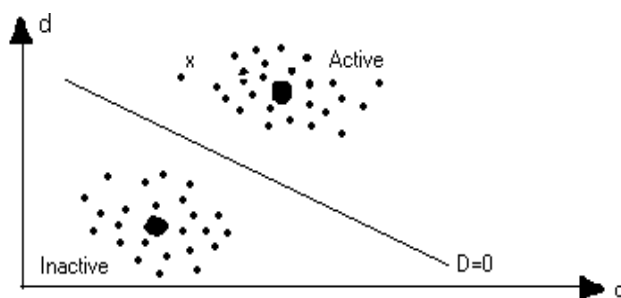


Рис. 4 Разделение молекул на активные и неактивные в пространстве дескрипторов (d). Крестиком помечена молекула с неизвестной активностью, жирная точка – центроид.

Линия, разделяющая кластеры, соответствует значению функции $D = 0$. Из рис.4 следует, что молекула, помеченная крестиком, ближе к группе активных.

Важным понятием дискриминационного анализа является понятие *центроида*. Под центроидом понимают систему, имеющую значения дескрипторов, которые вычисляются как средние по группе. Таким образом, центроид можно воспринимать как наиболее типичную (чаще всего гипотетическую) молекулу из заданной группы молекул. Значение дискриминационной функции для центроида, в свою очередь, может служить опорным при интерпретации расчетов D -функций реальных систем. На Рис. 4 центроид схематически показан в виде жирной точки.

Качество дискриминации по группе α обычно оценивается по отношению

$$\eta_{\alpha}(\%) = \frac{100 \cdot n}{n_{\alpha}},$$

где n - число молекул верно отнесенных с помощью D -функции к заданной группе.

3.4. ЗАДАЧИ

59. Согласно концепции Пульмана канцерогенная активность конденсированных углеводородов связывается с электронными свойствами так называемых К - и L - областей молекулы (Рис. 5).

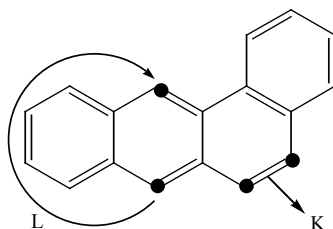


Рис. 5 К- и L- области в молекуле бензантрацена

В табл. 3.5 приведены максимальные значения электрофильных сверхделокализуемостей К- и L- областей некоторых углеводородов вместе с экспериментальными данными о наличии (+) или отсутствии (-) значительной канцерогенной активности.

На основе этих данных постройте дискриминационную функцию для распознавания канцерогенной активности углеводородов. Сделайте прогноз активности для трех последних в табл. 3.5 молекул (фенантрена, бенз(а)пентацена и дибенз(б,к)перилена). Каковы “координаты” центроида? Вычислите матрицу “расстояний” между объектами. Постройте дендрограмму описывающую “близость” различных молекул.

Таблица 3.5 Суперделокализуемости К- и L- областей и канцерогенная активность конденсированных углеводородов

№	Название	S _e (K)	S _e (L)	Активность
1	Трибензо(а,е,і)пирен	2,003	1,993	+
2	Бензо(г)хризен	1,970	1,970	+
3	Дибензо(а,н)антрацен	2,061	2,292	+
4	Нафто(2,3-в)пирен	2,140	2,046	+
5	Дибензо(а,е)пирен	1,995	1,995	+
6	Дибензо(а,н)пирен	1,997	1,947	+
7	бензо(а)пирен	2,028	1,879	+
8	Бензол	1,667	1,667	+
9	Нафталин	1,867	1,989	-
10	Антрацен	1,995	2,626	-
11	Диантрацено (2,33-в,н)	2,325	3,170	-
12	нафто(2,3-е)пирен Стирол	1,761	1,761	-
13	β-нафтостирола	1,894	2,124	-
14	Фенантрен	1,994	1,916	?
15	бенз(а)пентацен	2,182	3,353	?
16	дибенз(в,к)перилена	2,027	2,772	?

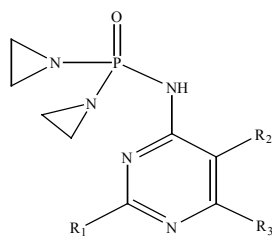
60. В табл. 3.6 приведена липофильность насыщенных углеводородов. Постройте дискриминационную функцию, с помощью которой можно было бы отделить соединения с высокой ($\lg P > 3,5$) липофильностью. В качестве дескрипторов выберите число углеродных атомов и какой-либо из параметров, характеризующих разветвленность

углеродного каркаса молекулы. Оцените эффективность дискриминации с помощью параметра η_α .

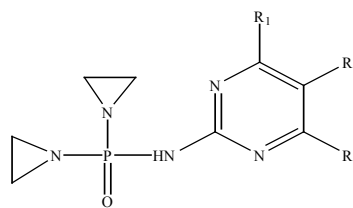
Таблица 3.6 Липофильности некоторых насыщенных углеводородов

Структура	C_5H_8	C_4H_{10}	$CH(CH_3)_3$	C_5H_{12}	$(CH_3)_2CH(C_2H_5)$	$C(CH_3)_4$
lg P	2,36	2,89	2,76	3,39	3,77	3,11
Структура	C_7H_{16}	C_8H_{18}	$(CH_3)_3C(C_2H_5)$	$C_{10}H_{22}$	$(CH_3)_2(CH)_2(CH_3)_2$	$C_{12}H_{26}$
lg P	4,66	5,18	3,82	5,01	3,42	6,1

61. Для двух типов фосфорильных производных пиримидина (структура А и структура В):



структура А



структура В

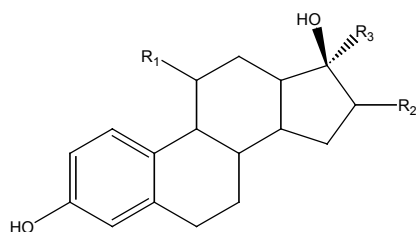
проведите разбиение на кластеры (табл. 3.7).

Таблица 3.7 Заместители фосфорильных производных пиримидина

№ структуры	R_1	R_2	R_3	№ структуры	R_1	R_2	R_3
A1	H	H	H	Б1	CH_3	H	H
A2	H	CH_3O	H	Б2	Cl	H	H
A3	H	H	Br	Б3	H	Cl	H
A4	CH_3	CH_3	H	Б4	H	Br	H
A5	CH_3	CH_3O	H	Б5	CH_3	H	CH_3
A6	H	CH_3	Br	Б6	CH_3O	H	Cl
A7	CH_3	H	Br	Б7	$N(C_2H_5)_2$	H	H

В качестве параметров структур можно выбрать различные характеристики заместителей (липофильности, σ -константы, электронные плотности на заместителях, дипольные моменты, энергии ВЗМО и НВМО и т.д., см. раздел 1,2). На сколько кластеров можно разбить всю совокупность молекул? Постройте дендрограмму, описывающую “близость” различных замещенных. Какой способ вычисления расстояний Вы используете? Зависит ли результат (кластеризация) от способа вычисления расстояний? Если зависит, то как?

62. Замещенные эстрадиолов относятся к классу стероидных гормонов:



В табл. 3.8 приведены данные, характеризующие относительное сродство замещенного (lgBA) с соответствующим биологическим рецептором (Chem.Rev., 1999, **99**, 723). Постройте матрицу Фри-Вильсона, описывающую активность замещенных эстрадиолов. Методом МНК найдите уравнение для биоактивности. Вычислите набор электронных дескрипторов описывающих эстрадиол (дип. момент, поляризуемость, энергии высшей занятой и нижней вакантной МО).

Проведите разбиение на два класса активности. Вычислите дискриминационную функцию описывающую активность приведенных молекул в терминах электронных дескрипторов. Каково качество Вашего прогноза по дискриминационной функции?

Таблица 3.8 Относительное сродство замещенных эстрадиолов

R1	R2	R3	lgBA	R1	R2	R3	lgBA
Et	H	H	2,60	H	OH	C≡CH	1,30
H	H	C≡CH	2,05	Et	OH	C≡CMe	1,26
H	H	H	2,00	OMe	H	C≡CMe	1,26
Et	H	C≡CH	1,94	OMe	H	C ₆ H ₅	1,85
Et	OH	Me	1,93	H	OH	Me	1,15
Et	H	Me	1,92	OMe	H	C≡CH	1,15
Et	OH	C≡CH	1,90	H	OH	C ₆ H ₅	0,90
Et	H	C≡CMe	1,82	OMe	H	H	0,90
H	H	Me	1,76	OMe	H	Me	0,78
H	H	C≡CMe	1,67	H	OH	C≡CMe	0,70
H	H	C ₆ H ₅	1,48	OMe	OH	C ₆ H ₅	0,70
Et	OH	H	1,45	OMe	OH	Me	0,60
OMe	H	HC≡CMe	1,42	OMe	OH	H	0,0
H	OH	H	1,32	OMe	OH	C≡CH	0,0
H	OH	C≡CH	1,30	OMe	OH	C≡CMe	-0,10

4. ЗАКЛЮЧЕНИЕ

В заключении авторам хотелось бы подчеркнуть несколько моментов. Прежде всего отметим, что описанные выше подходы и соответствующие им задачи призваны продемонстрировать лишь основы методологии QSAR. Исследование биологической активности молекул, как и любая другая нетривиальная задача, требует творческого понимания проблемы. В особенности это касается отбора дескрипторов, способных описать данный вид активности. Необходимо осознавать, что

универсального набора дескрипторов для любых систем и любых типов активности не существует. Уместно также вспомнить принцип Оккама: “Не приумножай сущностей более чем необходимо”, поскольку чрезмерно “раздутый” дескрипторный набор увеличивает риск случайных корреляций. В целом QSAR методология не предполагает глубокого понимания сущности конкретных биохимических процессов в организме, однако, имея какую-либо модель этих процессов, можно существенно облегчить формулировку самого вида количественного соотношения структура-активность.

В настоящем пособии не рассмотрены некоторые важные статистические подходы, поскольку их изложение требует специальных знаний. Среди них *факторный анализ* (один из его вариантов известен как *метод главных компонент*) – статистический метод, позволяющий проанализировать структуру взаимосвязей элементов дескрипторного набора и, что очень существенно, сжать его. Своеобразным гибридом МНК и *метода главных компонент* является частичный метод наименьших квадратов (*partial least squares*, PLS), использующийся в расчетах систем с большим количеством дескрипторов.

Необходимо отметить также, что QSAR – это бурно развивающаяся область хемометрии. За последнее десятилетие появилось множество новых подходов. Среди них так называемый *трехмерный QSAR* (3D-QSAR) – мощный метод, предполагающий зависимость биоактивности от стерических свойств молекул и их электростатических полей. Набирает популярность метод *искусственных нейронных сетей* – алгоритм, симулирующий функционирование нейронов. Он используется в проблеме молекулярного распознавания и классификации. *Генетические алгоритмы* – новый подход, позволяющий корректно обрабатывать данные, которые содержат больше переменных (дескрипторов), чем объектов (молекул).

Все эти методы наряду с проверенными известными подходами составляют теоретическую базу для целенаправленного поиска новых лекарственных препаратов. Авторы надеются, что данное пособие облегчит понимание студентами этой не простой и практически важной задачи.

5. СПИСОК ЛИТЕРАТУРЫ

1. Станкевич М.И., Станкевич И.В., Зефирова Н.С. Топологические индексы в органической химии // Успехи химии.- 1988.-Т.57, № 3.- С.337-365.
2. Раевский О.А. Дескрипторы молекулярной структуры в компьютерном дизайне биологически активных веществ // Успехи химии.- 1999.-Т.68, № 6.-С.555-575.
3. Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов.-М.:Наука, 1986.- 230с.
4. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989.- 213с.
5. Розенблит А.Б, Голиндер В.Е. Логико-комбинаторные методы в конструировании лекарств. – Рига:Зинатне, 1984.- 351с.
6. Стьюпер Э., Брюггер У., Джурс П. Машинный анализ связи химической структуры и биологической активности. – М.:Мир, 1982.- 235с.
7. Коваленко С.Н., Друшляк А.Г., Черных В.П. Основы комбинаторной органической химии. – Харьков, из-во НфаУ “Золотые страницы”, 2003.- 106с.

6. ПРИЛОЖЕНИЕ

НЕКОТОРЫЕ ТЕРМИНЫ И ПАРАМЕТРЫ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ

В качестве меры биологической активности обычно используют количество вещества на единицу биомассы (или его логарифм), необходимое для достижения заданного уровня биологического действия.

LD₅₀ – доза поражающая 50% экспериментальных организмов.

ED₅₀ – доза антагониста, снижающая на 50% действие стандартной дозы агониста.

МИС – концентрация ингибитора, подавляющая рост исследуемого биологического объекта до минимального уровня.

IC₅₀ – концентрации вещества, приводящая к 50%-ному ингибированию протекания ферментативной реакции.

Агонист – чужеродное организму вещество активизирующее рецепторы и вызывающее в биообъекте те же изменения, что и естественные (эндогенные) вещества. т.е. обладающее биоактивностью

Антагонист – вещество блокирующее рецепторы и тем самым ослабляющее или полностью предотвращающее действие другого вещества.

Липофильность (гидрофобность) – параметр, характеризующий распределение вещества между водно-липидной фазой. Ранее, для этой цели, использовался логарифм коэффициента распределения в системе оливковое масло – вода. В настоящее время общепринятой системой является н-октанол-вода.

Рецептор – активный фрагмент макромолекулы (белка) с которым взаимодействуют агонисты и антагонисты.