ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №20

Хемоинформатика
и
виртуальный скрининг

Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

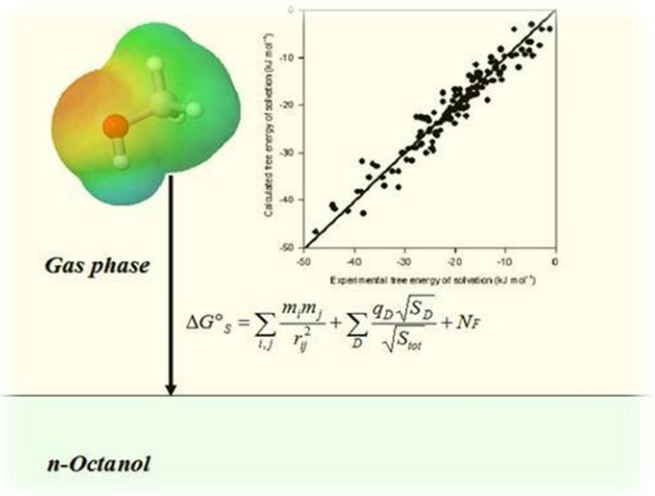
Сайт курса http://intbio.org/bioinf2018-2019

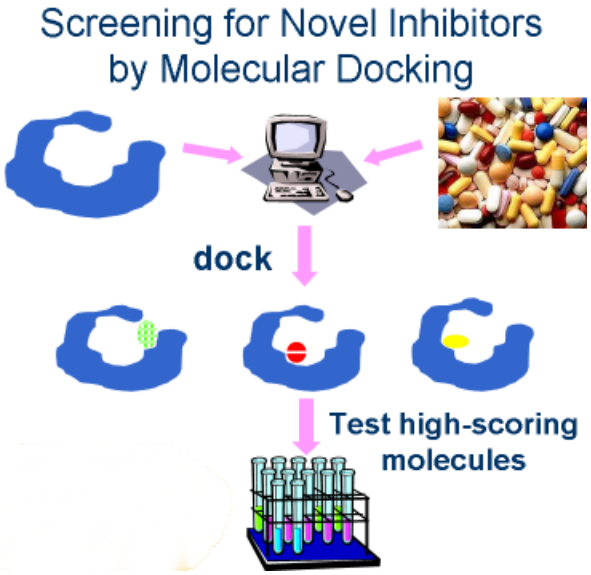
Хемоинформатика

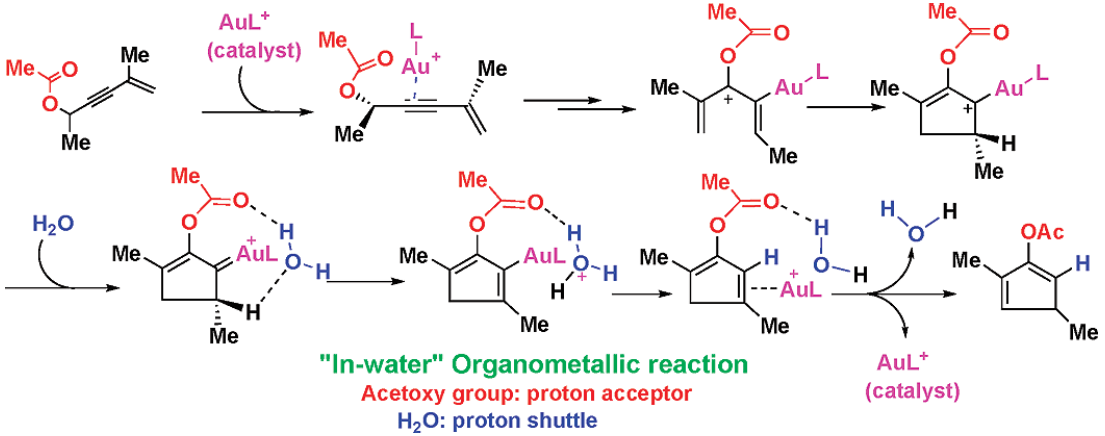
– это применение методов информатики для решения химических задач.

Область применения:

- Предсказание свойств химических соединений (QSPR)
- Поиск по химическому подобию, фармакофорный поиск, виртуальный скрининг
- Компьютерный синтез







Хемоинформатика и все-все-все...

Представление структуры молекул

Молекулярный граф – связный неориентированный граф, находящийся во взаимно-однозначном соответствии со структурной формулой химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а рёбрам графа — химические связи между этими атомами.

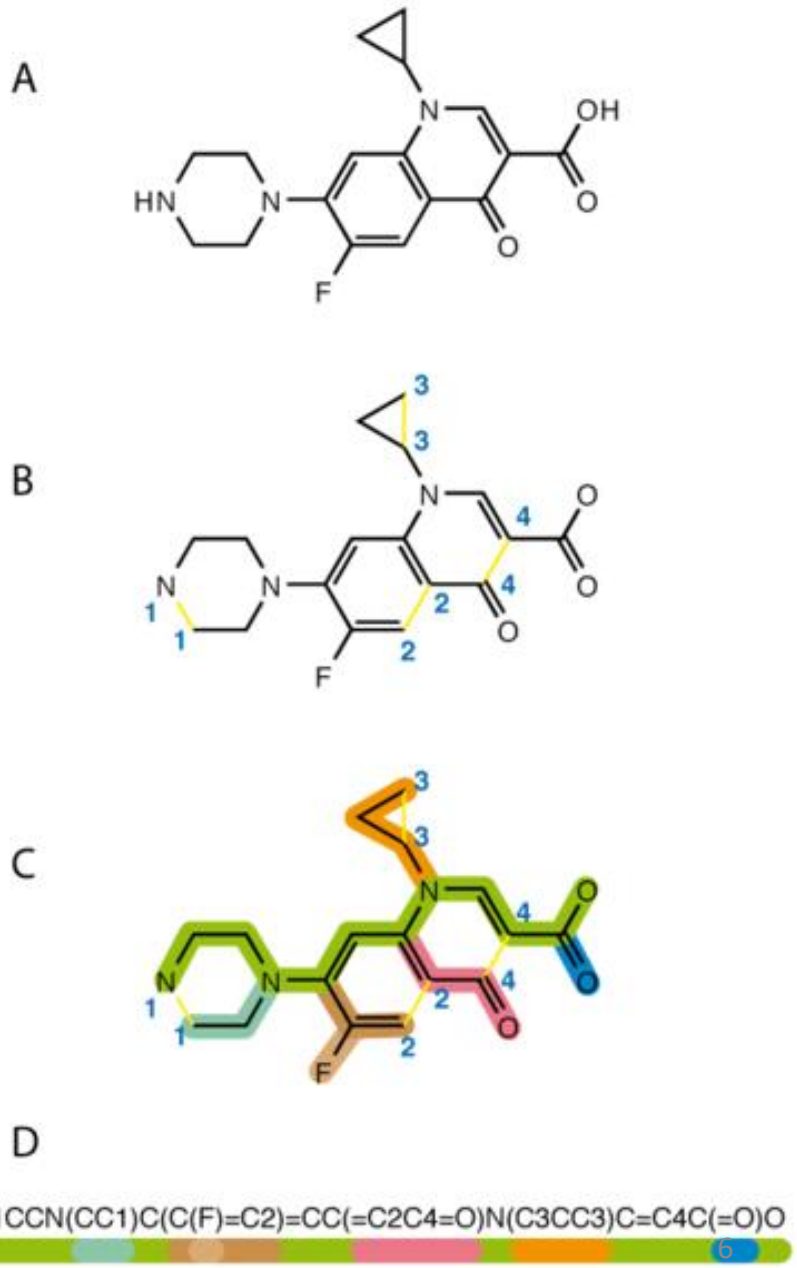
Способы записи:

- Линейные нотации (SMILES, SMARTS, SLN, InChI)
- Матрица смежности
- Структурные файлы (общие – MOL, SDF,… специальные - MOL2, HIN,…)
- Chemical Markup Language

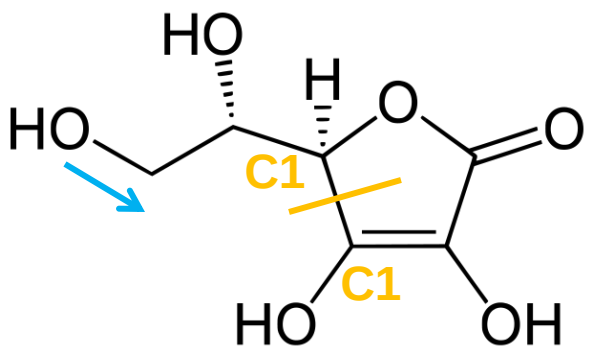
Линейные нотации. SMILES

SMILES (англ. *Simplified Molecular Input Line Entry Specification, спецификация упрощенного представления молекул в строке ввода*) — система правил (спецификация) однозначного описания состава и структуры молекулы химического вещества с использованием строки символов ASCII.

Вода	O
Этанол	CCO
Углекислый газ	O=C=O
Синильная кислота	C#N
Циклогексан	C1CCCCC1
Бензол	c1ccccc1

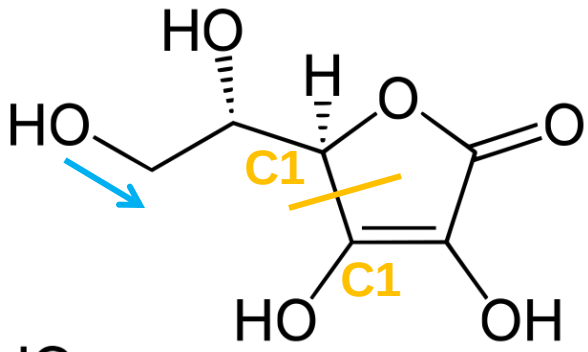


Линейные нотации. SMILES

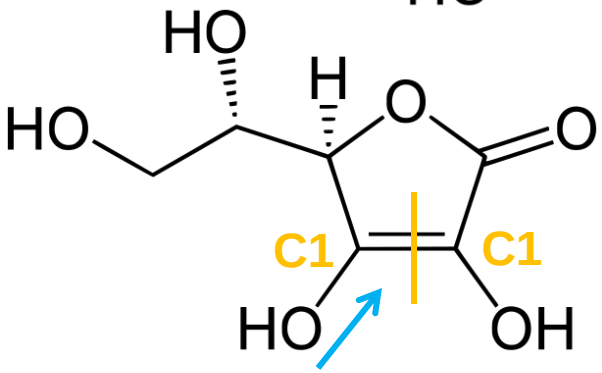


OCC(O)C1OC(=O)C(O)=C1(O)

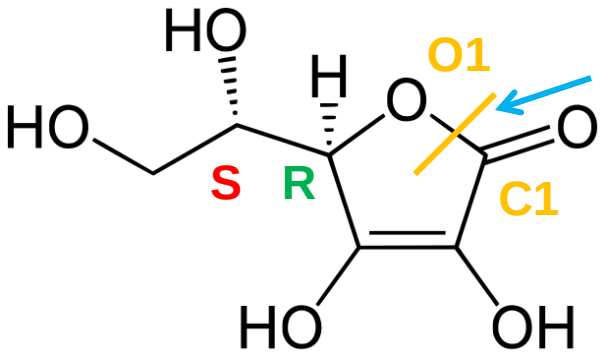
Линейные нотации. SMILES



OCC(O)C1OC(=O)C(O)=C1(O)



OC=1C(OC(=O)C=1O)[C@@H](O)CO



O=C1C(O)=C(O)[C@H](O1)[C@@H](O)CO

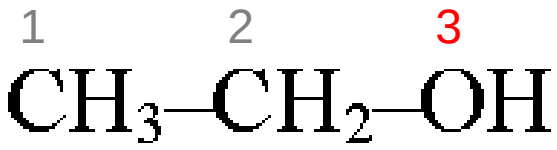
@ - перечисление заместителей против часовой стрелки
@@ - по часовой

https://jacobdurrant.com/2014/05/20/online-smiles-string-to-3d-visualizer/

Линейные нотации. InChI

InChI (International Chemical Identifier) — текстовый идентификатор химического соединения для стандартизации кодирования молекулярной информации и представления её в читаемом виде.

Этанол InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3

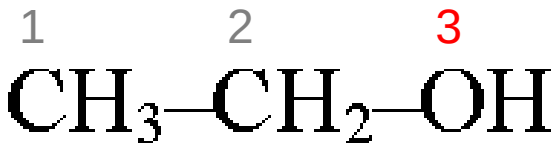


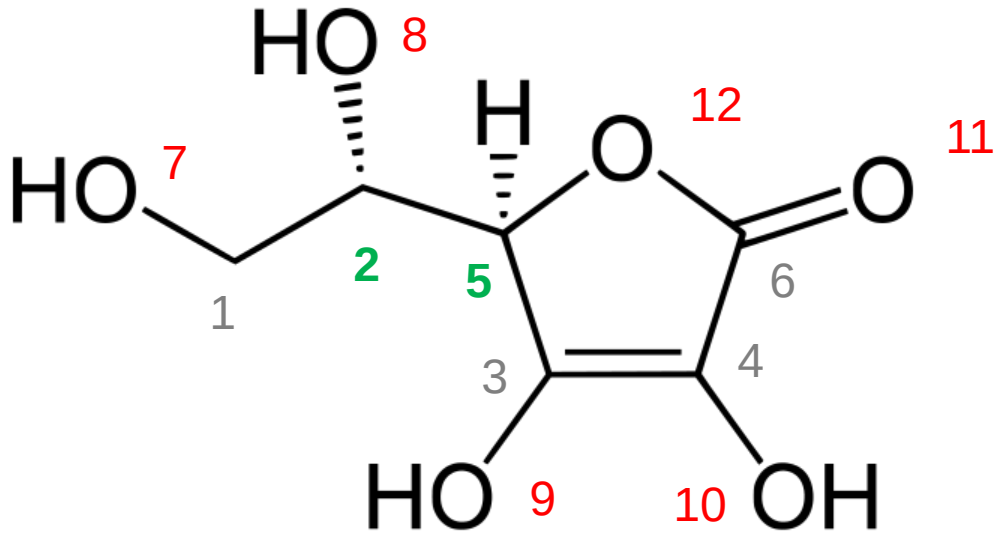
Линейные нотации. InChI

InChI (International Chemical Identifier) — текстовый идентификатор химического соединения для стандартизации кодирования молекулярной информации и представления её в читаемом виде.

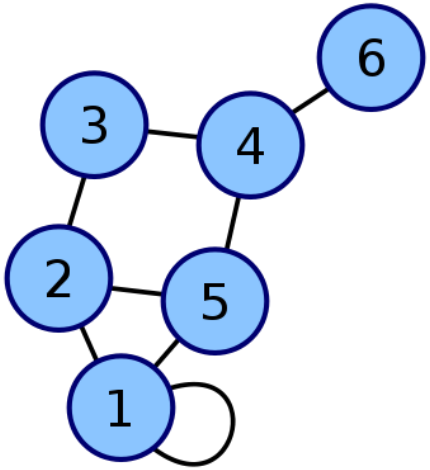
Этанол InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3

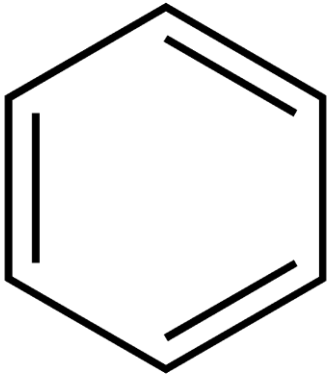
Аскорбиновая кислота InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1





Представление структуры молекул. Матрица смежности

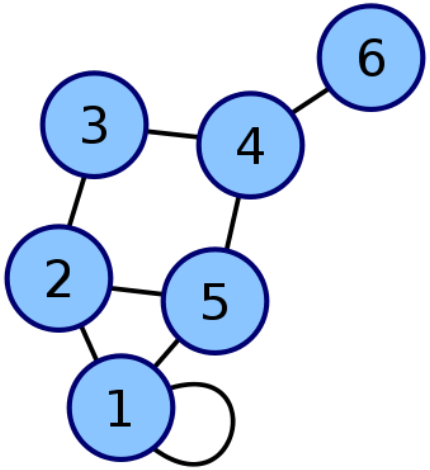


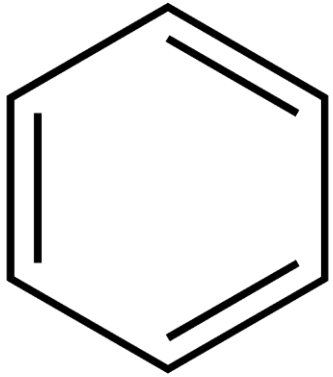


$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

?

Представление структуры молекул. Матрица смежности

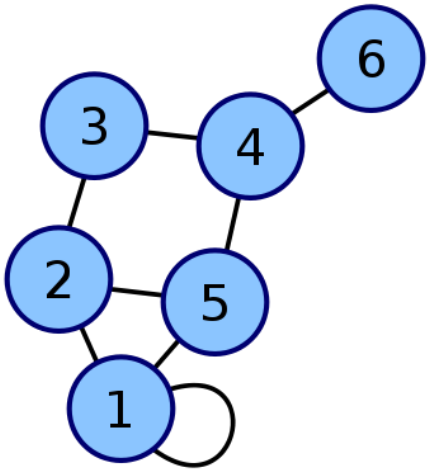


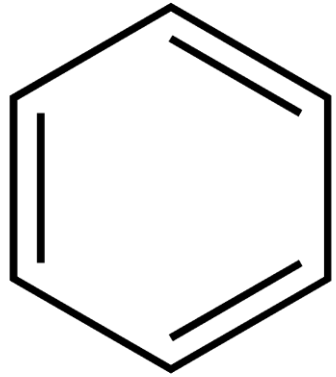


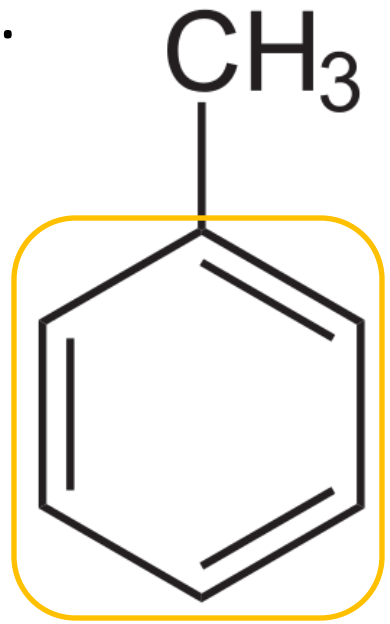
$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Представление структуры молекул. Матрица смежности







$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

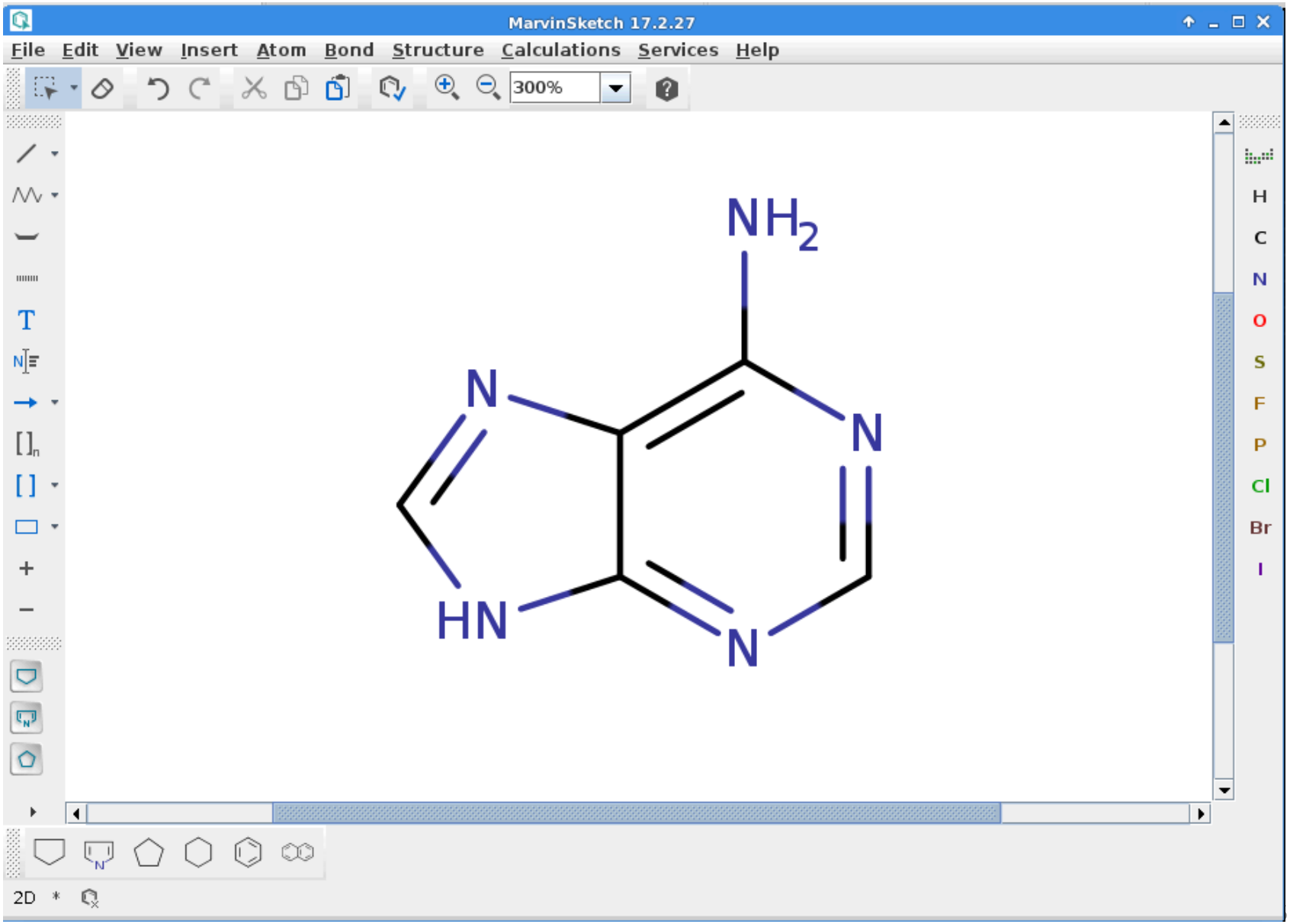
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

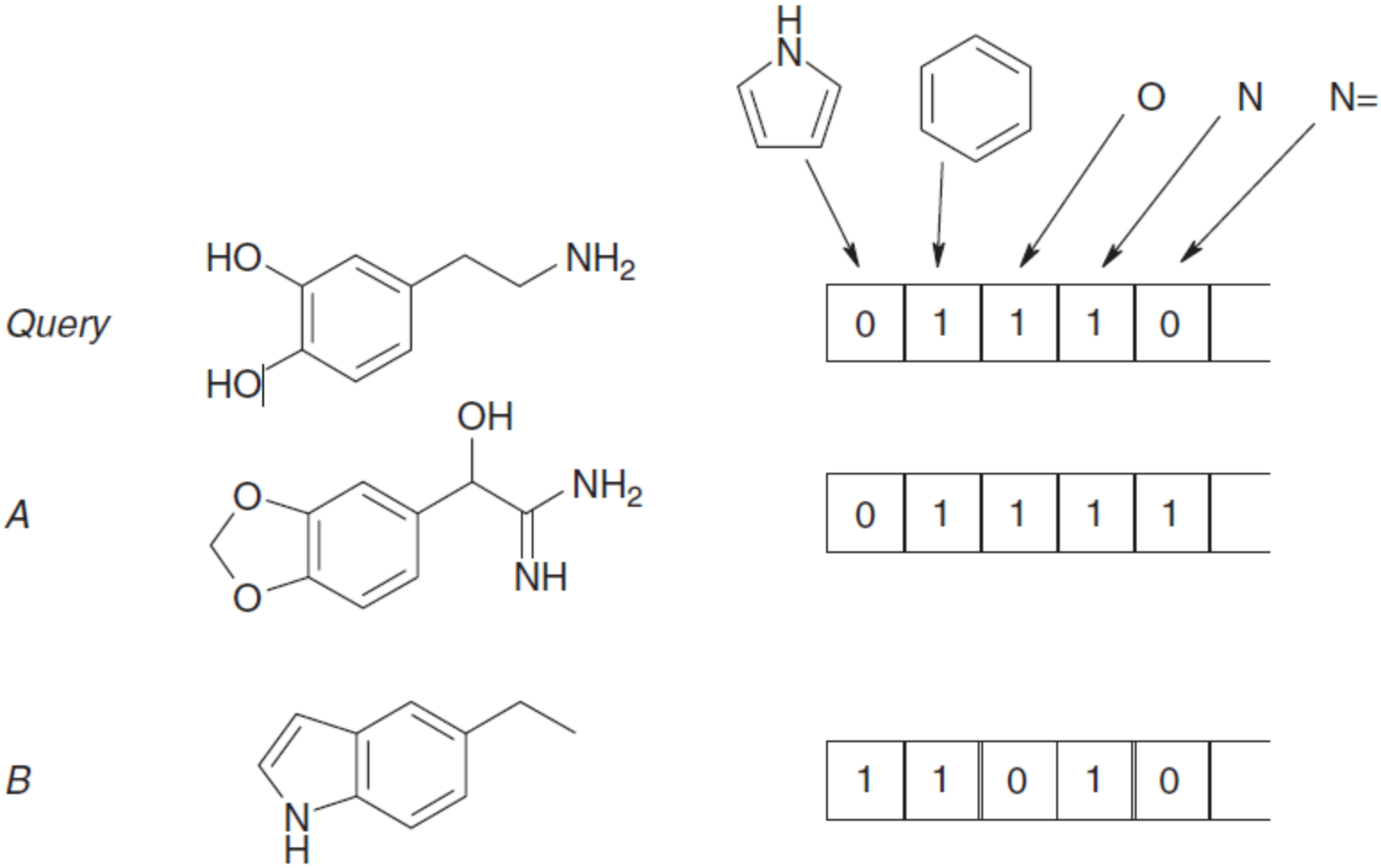
Структурные файлы. MOL

```
benzene
ACD/Labs0812062058
6  6  0  0  0  0  0  0  0  0  1 V2000
    1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  1  2  0  0  0  0
  4  2  2  0  0  0  0
  5  3  1  0  0  0  0
  6  4  1  0  0  0  0
  6  5  2  0  0  0  0
  M  END
```

Молекулярные редакторы. MarvinSketch



Представление структуры молекул. Битовые строки



Предсказание свойств соединений

Поиск количественных соотношений структура-свойство — процедура построения моделей, позволяющих по структурам химических соединений предсказывать их разнообразные свойства.

Основная гипотеза – сходные соединения имеют сходные свойства.

QSAR - *Quantitative Structure-Activity Relationship* – биологические свойства

QSPR - *Quantitative Structure-Property Relationship* – физические и физико-химические свойства

Примеры:

- Температуры плавления и кипения
- Вязкость
- Давление насыщенных паров
- Плотность
- Химические сдвиги в спектрах ^{1}H ЯМР
- Растворимость
- ...

Предсказание свойств соединений

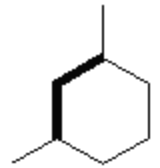
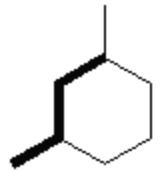
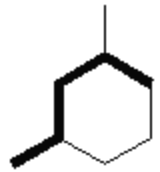
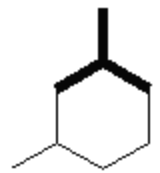
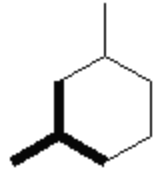
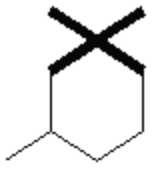
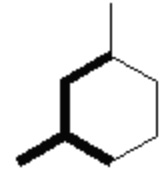
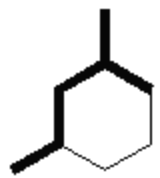
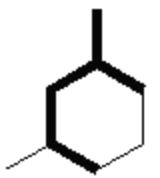
Молекулярные дескрипторы:

"*The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.*" (Todeschini and Consonni, 2000)

Молекулярные дескрипторы:

- Теоретические (число кратных связей, наличие молекулярных фрагментов,...)
- Экспериментальные (гидрофобность, поляризуемость, показатель преломления,...)

Дескрипторы инвариантны, т.е. не зависят от положения молекулы в пространстве.

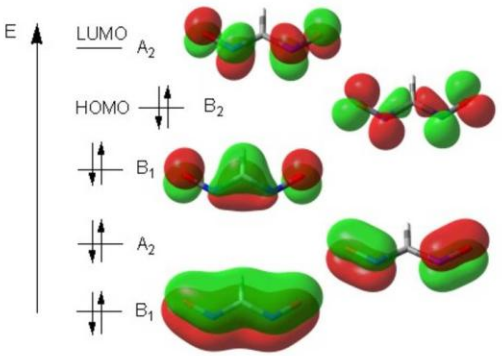
Subgraph	Examples		
Path of order...	 2	 3	 4
Cluster of order...	 3	 3	 4
Path/Cluster of order...	 1/3	 2/3	 2/3

Молекулярные дескрипторы

Фрагментные дескрипторы – отражают факт наличия фрагмента в молекулярном графе (*бинарные*) или число вхождений фрагмента (*целочисленные*)

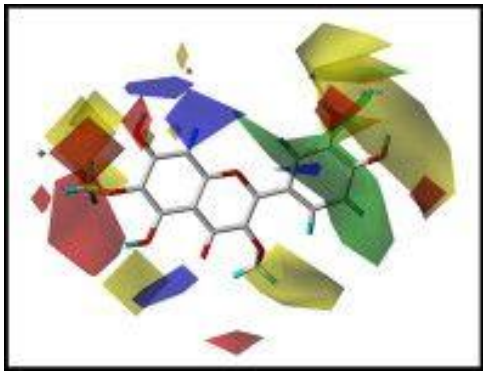
Физико-химические дескрипторы – соответствуют измеряемым физ-хим величинам (липофильность (LogP), молярная рефракция (MR), молекулярный вес (MW), молекулярные объемы и площади поверхностей,...)

Квантово-химические дескрипторы — величины, получаемые в результате квантово-химических расчетов (энергии граничных орбиталей, частичные заряды на атомах, порядки связей,...)



Дескрипторы молекулярных полей — величины, аппроксимирующие значения молекулярных полей путем вычисления энергии взаимодействия пробного атома, помещенного в узел решетки, с рассматриваемой молекулой

Другие…



Предсказание свойств соединений

В самой общей форме: значение свойства – это некая функция от некого набора дескрипторов.

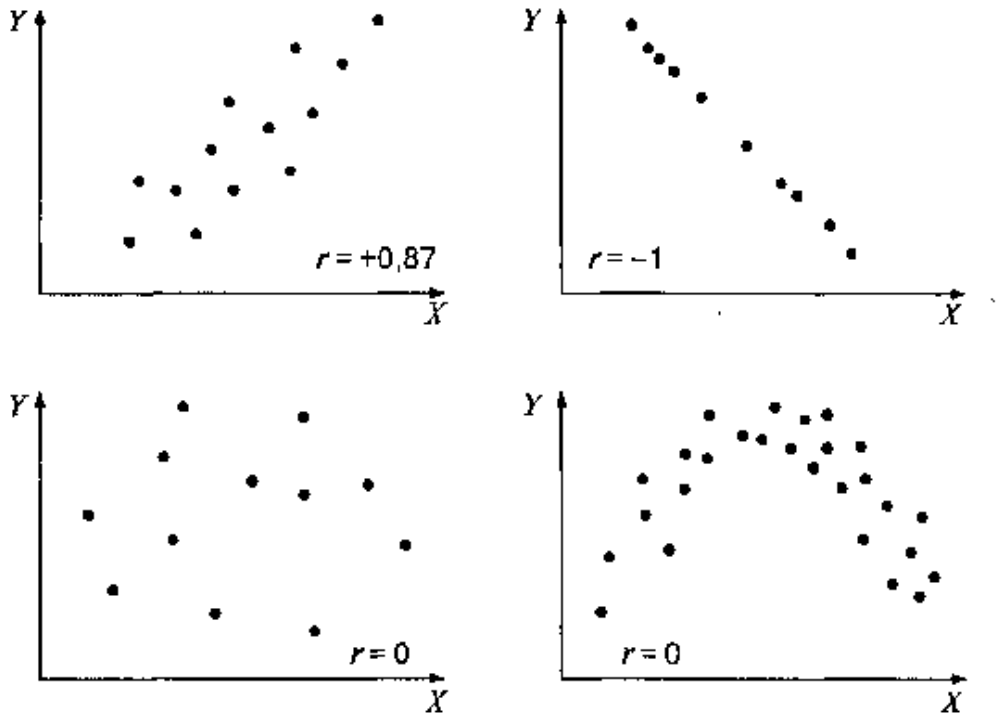
Цель: найти оптимальную функцию и оптимальный набор.

Выявленная связь должна быть проверена =>

- Сравнение модели с экспериментом (коэффициент корреляции)
- Обучающая выборка, тестовая выборка
- Перекрестная проверка (cross-validation) (для маленьких выборок)
- Рандомизация (для больших выборок)

Успех QSAR-модели зависит от точности исходных данных, выбора подходящих дескрипторов и статистических методов и полноценной проверки модели.

Сравнение модели с экспериментом. Коэффициент корреляции

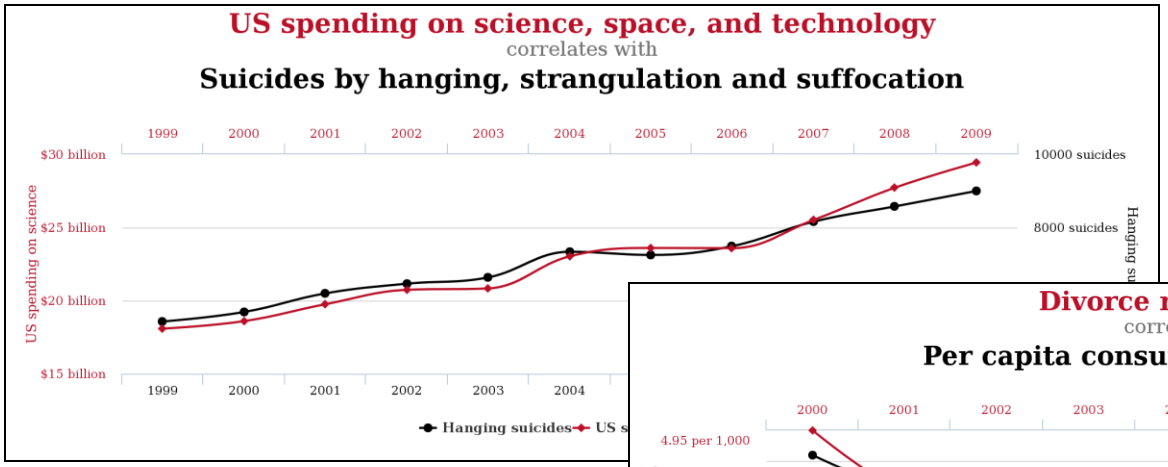


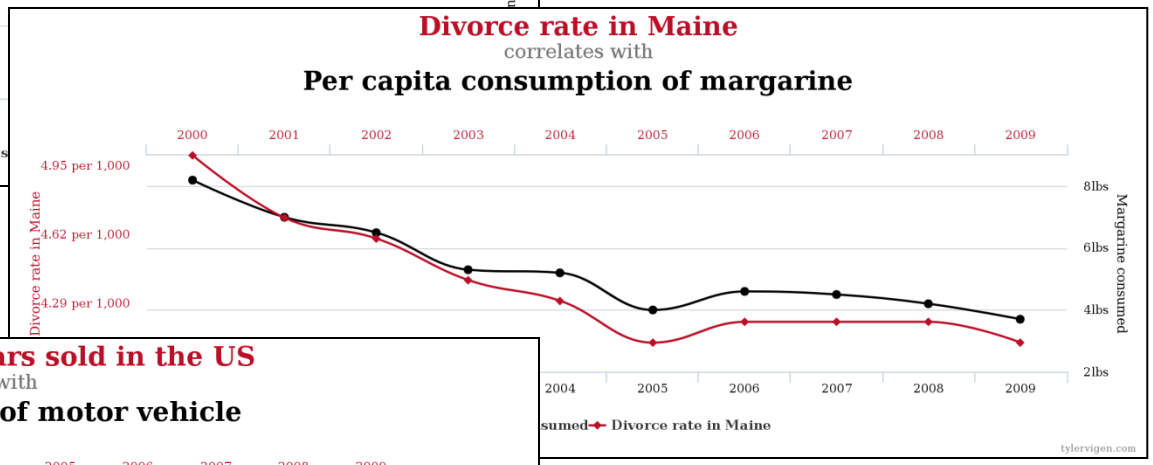
$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

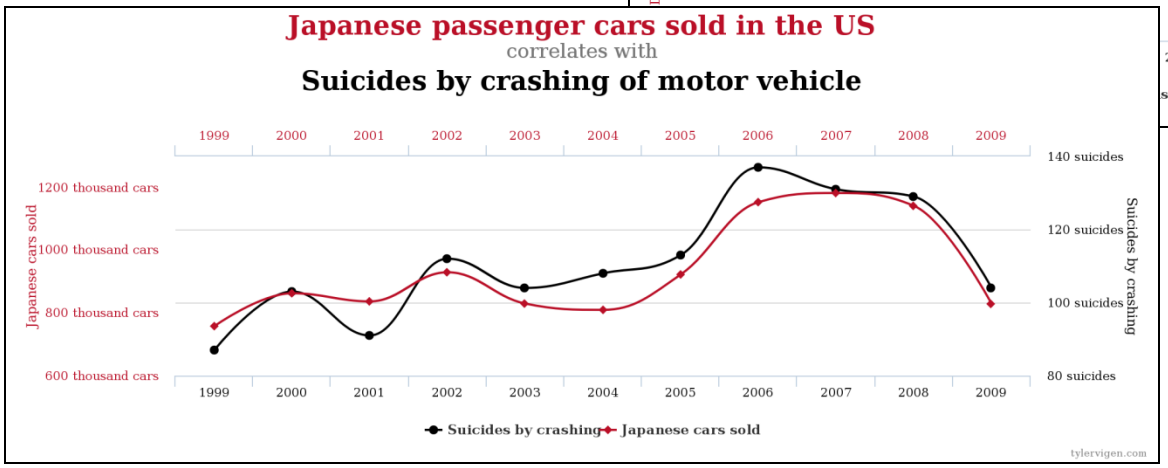
Отсутствие корреляции означает неадекватность выбранной модели

Неожиданные корреляции

http://tylervigen.com/spurious-correlations

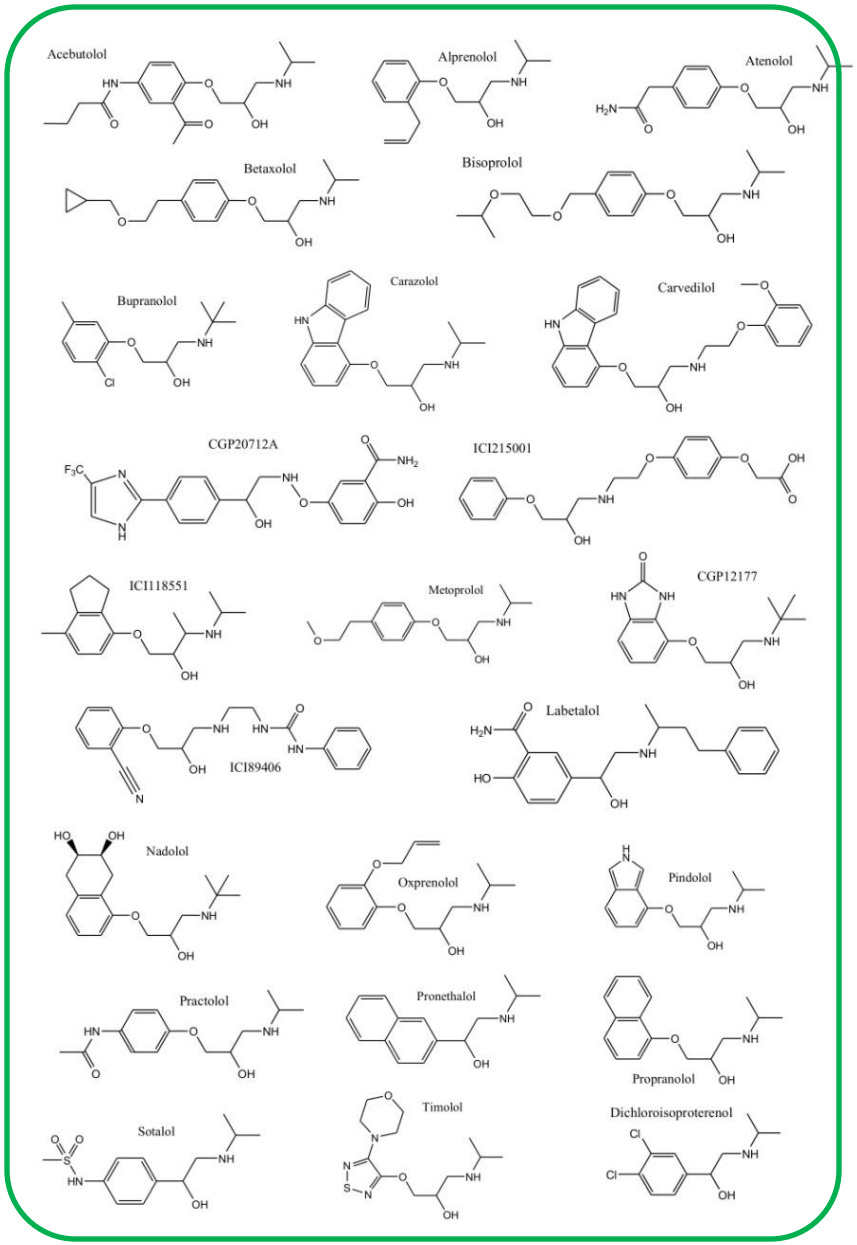


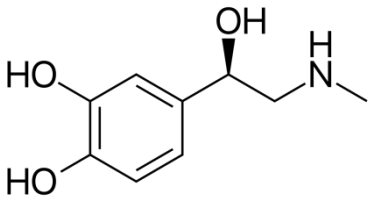




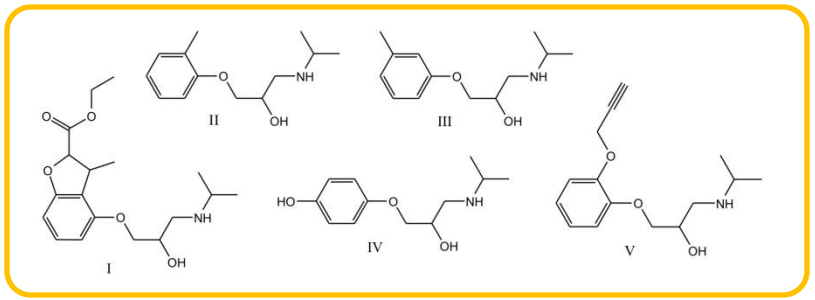
Наличие корреляции не всегда что-то означает. Хотя… ☺

Обучающая и тестовая выборки





Лиганды бета2-адренэргического рецептора

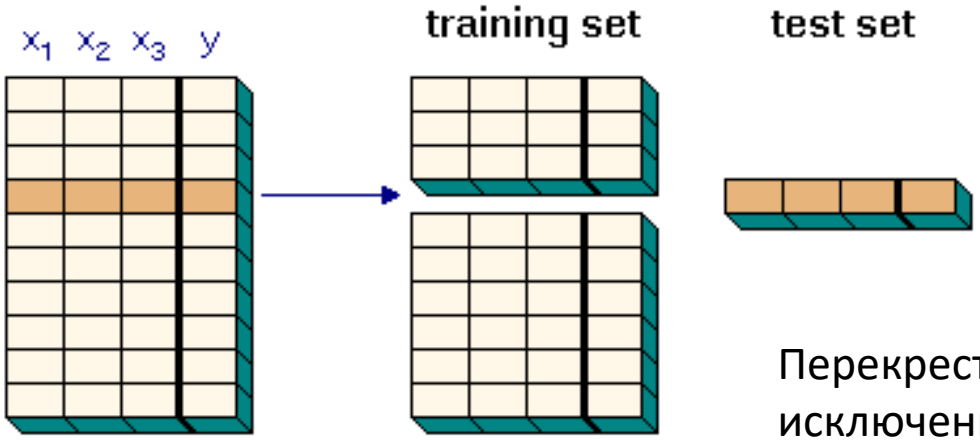


Выборки должны быть репрезентативными, т.е. отражать генеральную совокупность со всей возможной полнотой

r – коэффициент корреляции по обучающей выборке (хорошо, когда > 0,9)

q – коэффициент корреляции по тестовой выборке (приемлемо, когда > 0,6)

Перекрестная проверка и рандомизация



Перекрестная проверка с одним исключенным – leave-one-out cross-validation

Для изучения свойств деревьев необязательно рассматривать каждое дерево в лесу

Аналогично в клинических исследованиях

Какие свойства можно предсказывать?

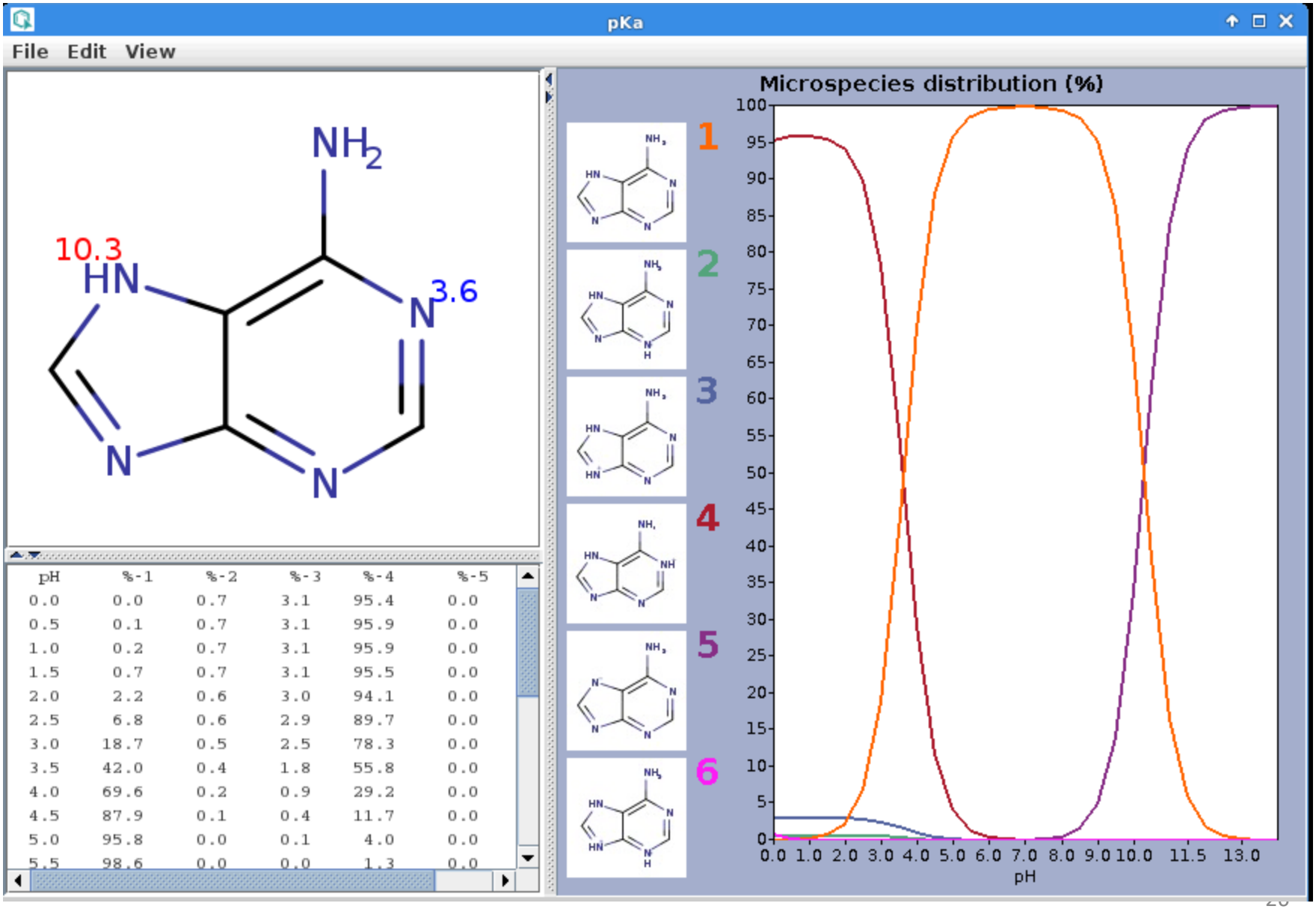
Физические свойства индивидуальных низкомолекулярных соединений
- Температура кипения (BP)
- Вязкость
- Плотность
- Показатель преломления
- Температура плавления (MP)
- Константы ионизации (кислотности или основности)
- ...

Спектроскопические свойства
- Положение длинноволновой полосы поглощения симметричных цианиновых красителей
- Химические сдвиги в спектрах ^{1}H ЯМР
- ...

Физические свойства, обусловленные межмолекулярными взаимодействиями молекул разного типа
- Растворимость в воде (LogSw)
- Коэффициент распределения *n*-октанол/вода (LogP)
- ...

Физические и физико-химические свойства полимеров
- Показатель преломления полимеров
- Коэффициент проницаемости через полиэтилен низкой плотности
- ...

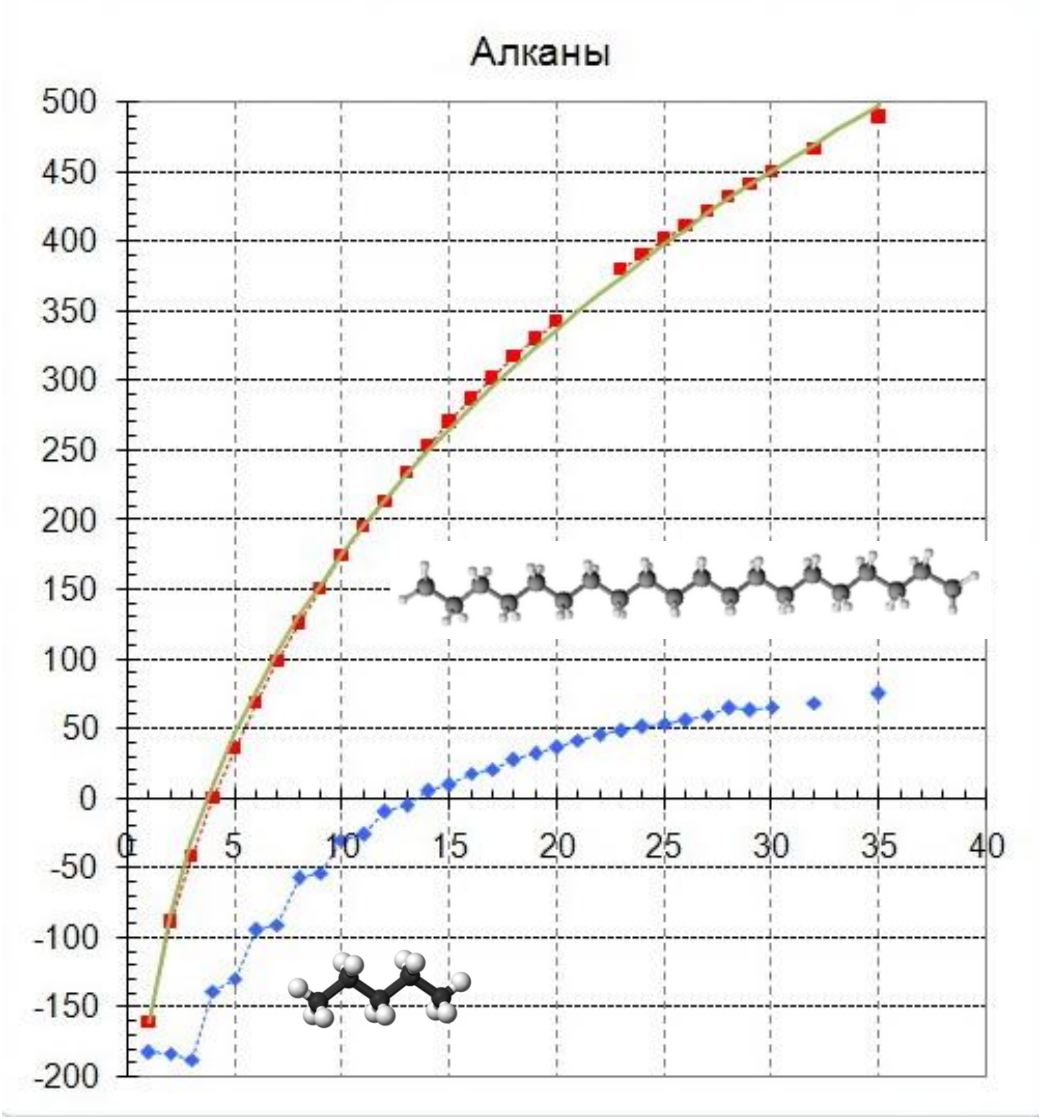
Предсказание pKa



Предсказание температур кипения

Предсказание температуры кипения линейных алканов

$$CH_3-(CH_2)_n-CH_3$$



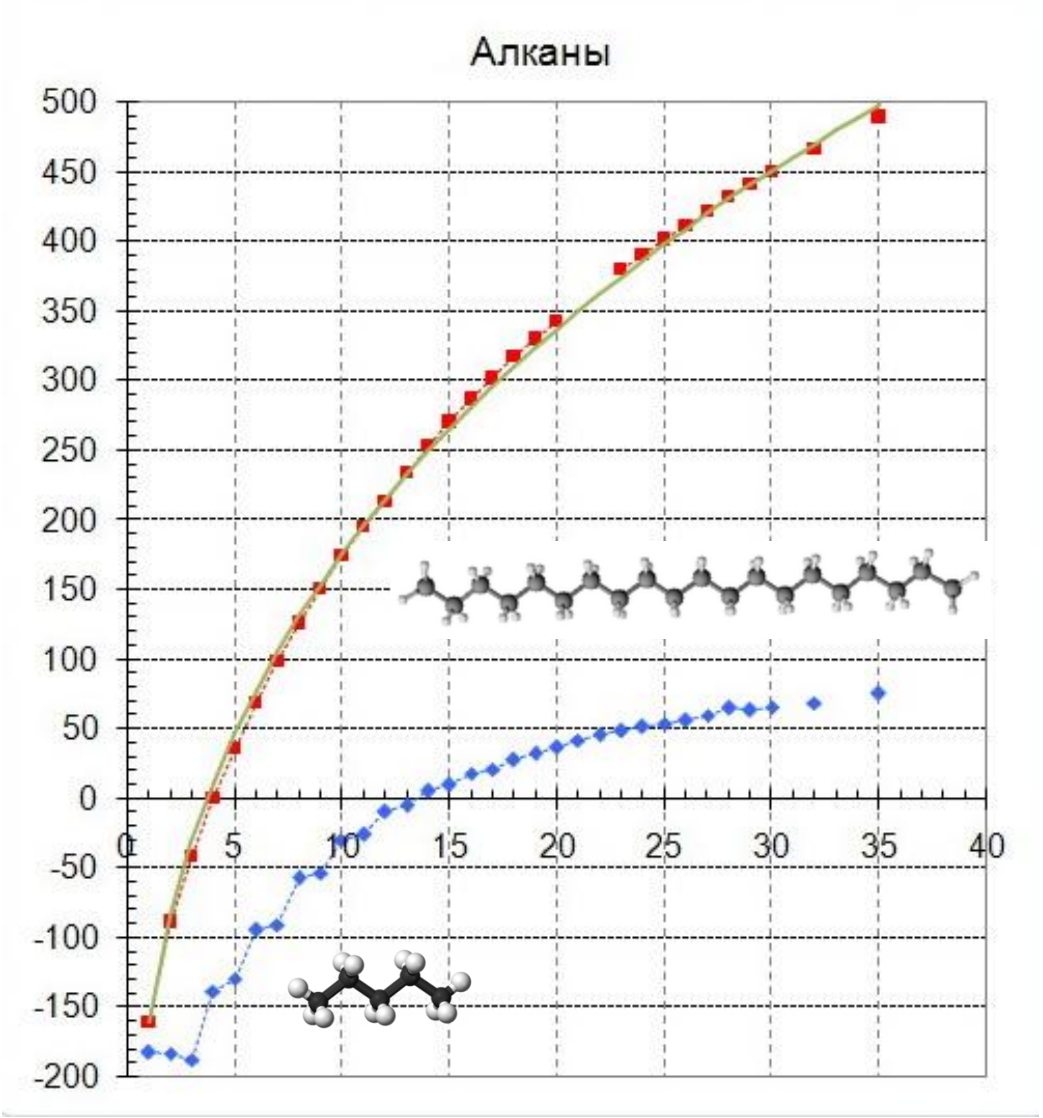
Предсказание температур кипения

Предсказание температуры кипения линейных алканов

$$CH_3-(CH_2)_n-CH_3$$

$$T(n)=295\cdot n^{0.33}-455$$

$$R^2=0.999645$$



Предсказание коэффициента гидрофобности

$$\log P_{ow} = \sum_i n_i \alpha_i$$

n_i – *число атомов типа i*

α_i – *вклад атома типа i*

Table I. Classification of atoms, and their contributions to octanol-water partition coefficient which is a measure of hydrophobicity.

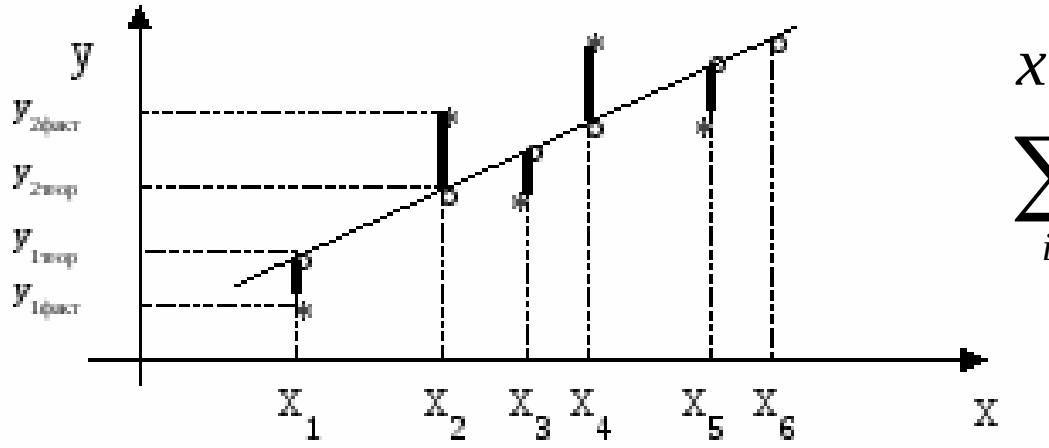
Type	Description[a]	Hydrophobic[b] Contribution	No. of Compounds	Frequency of Use	T-test	Molar Refraction[c]
	C in:					
1	:CH_3R, CH_4	−0.6037	360	548	100.00	2.3000
2	:CH_2R_2	−0.4295	216	454	100.00	2.3071
3	:CHR_3	−0.3426	45	50	100.00	2.4926
4	:CR_4	−0.1155	24	24	74.32	2.3000
5	:CH_3X	−1.0578	157	224	100.00	3.4006
6	:CH_2RX	−0.8188	257	402	100.00	3.2624
7	:CH_2X_2	−0.1540	5	5	51.00	3.6770
8	:CHR_2X	−0.5995	73	118	100.00	3.0137
9	:$CHRX_2$	0.0095	27	27	7.85	3.225
10	:CHX_3	0.5134	4	4	96.02	3.2401
11	:CR_3X	−0.4807	14	14	99.97	2.6140
12	:CR_2X_2	0.2853	2	2	58.14	3.1488
13	:CRX_3	0.5335	34	36	100.00	2.3010
14	:CX_4	1.1114	6	6	100.00	3.3559
15	:—CH	−0.1654	25	31	97.01	3.5071

http://www.vcclab.org/lab/alogps/

9000+ соединений, 115 типов атомов (Ghose et al., 1986, 1998)

Линейная регрессия

Метод наименьших квадратов (Гаусс, 1795; Лежандр, 1805)



$$x = \{x_i\},\ i = 1, 2, 3\ldots$$

$$\sum_i (y_i - f(x_i))^2 \to \min_x$$

$$y_i = a + bx_i + cz_i$$

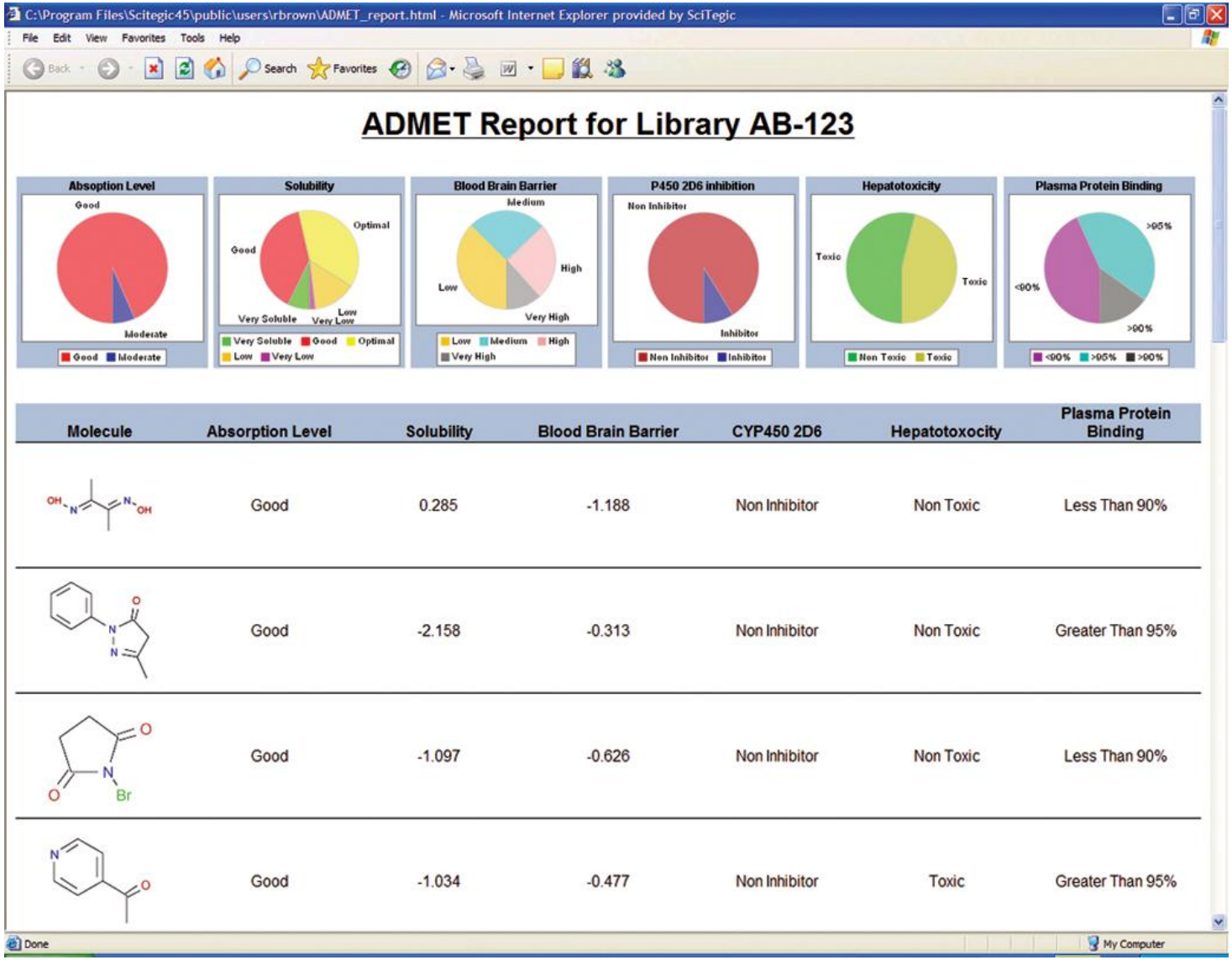
$$y_i = a + bx_i + cz_i + dx_i z_i$$

$$y_i = a + bx_i + cz_i + bcx_i z_i$$

Функция может быть любой, но линейной по коэффициентам !!!

Какие свойства можно предсказывать?

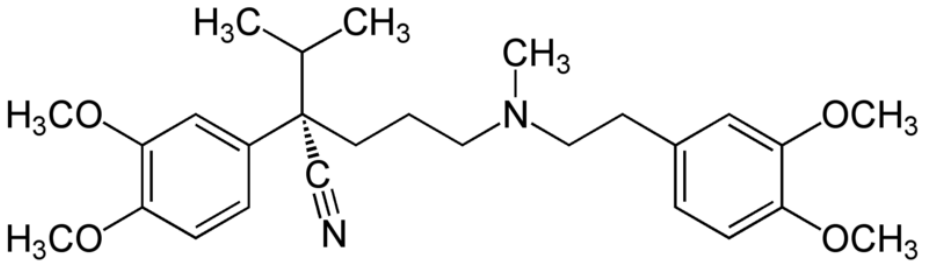
ADMET – Absorption, Distribution, Metabolism, Excretion, and Toxicity



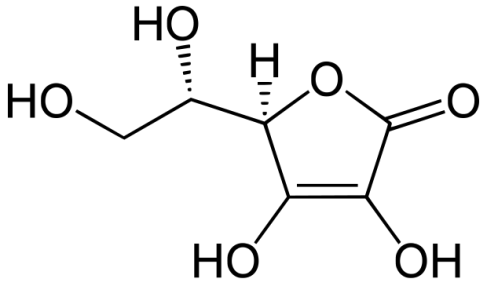
Предсказание биологической активности

«Правило пяти» (Lipinski, 1997) (Rule of thumb):

- Не более 5 доноров водородных связей
- Не более 10 акцепторов водородных связей
- Относительная молярная масса не более 500
- LogP не более 5



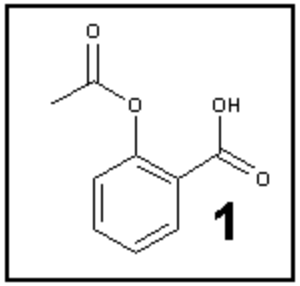
Верапамил
Mw=454 LogP=3,79

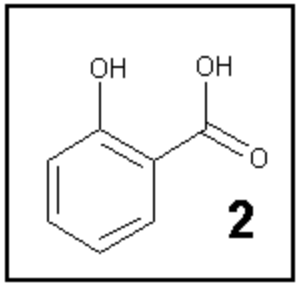


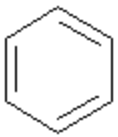
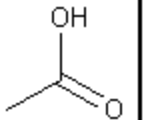
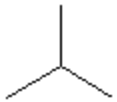
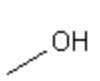
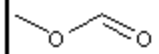
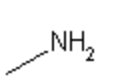
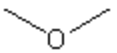
Аскорбиновая кислота
Mw=176 LogP=-1,9

Определение сходства молекул

Similarity Searching





1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

TANIMOTO COEFFICIENT = A / (A + B + C)
= 3 / (3 + 2 + 0) = 0.6 or 60%

О мерах сходства

Коэффициент Танимото (для битовых строк Xi и Yi) (1960):

$$S_T = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

Λ – логическое И
V – логическое ИЛИ

A	1	0	1	1	0	1	$\lvert A\rvert = 4$
B	1	1	0	1	0	0	$\lvert B\rvert = 3$
$A \wedge B$	1	0	0	1	0	0	$\lvert A \wedge B\rvert = 2$
$A \vee B$	1	1	1	1	0	1	$\lvert A \vee B\rvert = 5$

$$S_T(A, B) = \tfrac{2}{5}$$

О мерах сходства

Коэффициент Жаккара («коэффициент флористической общности»):

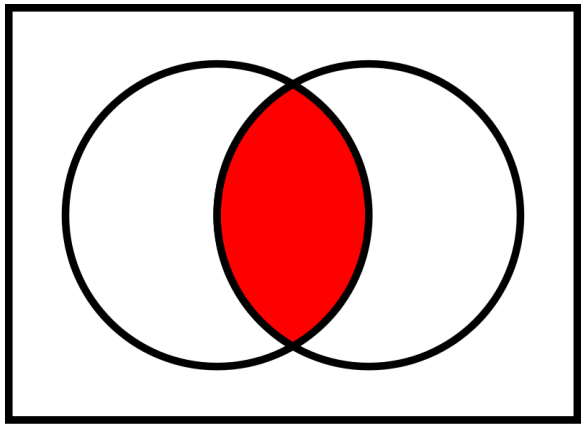
$$K_J = \frac{c}{a+b-c}$$

Первый предложенный коэффициент сходства!
(P. Jaccard, 1901)

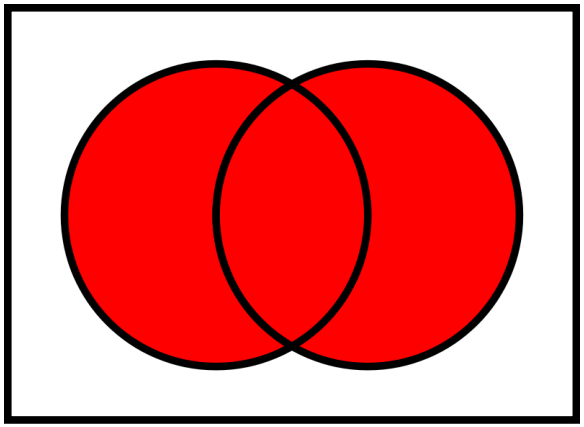
a — количество видов на первой пробной площадке,
b — количество видов на второй пробной площадке,
c — количество видов, общих для 1-ой и 2-ой площадок.

О мерах сходства

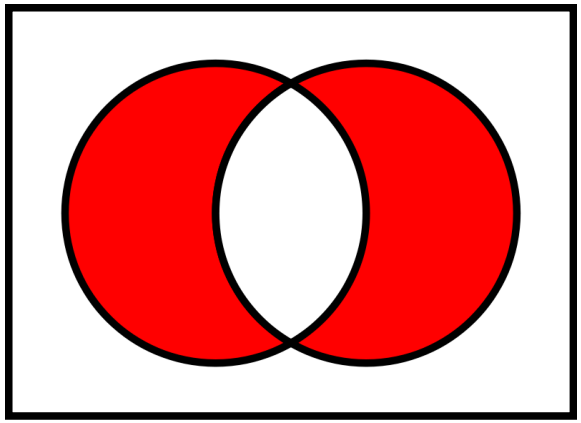
Диаграммы Венна («Eulerian Circles»,1880):



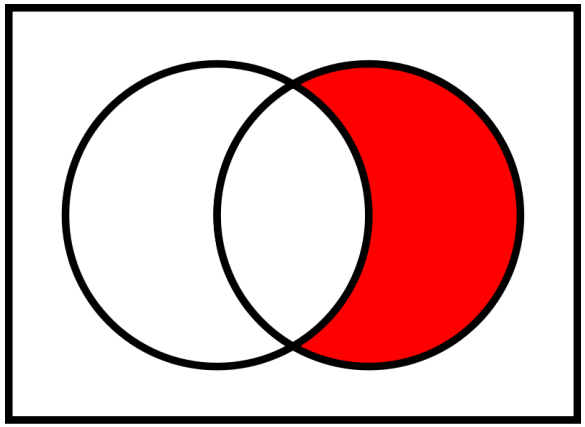
Пересечение



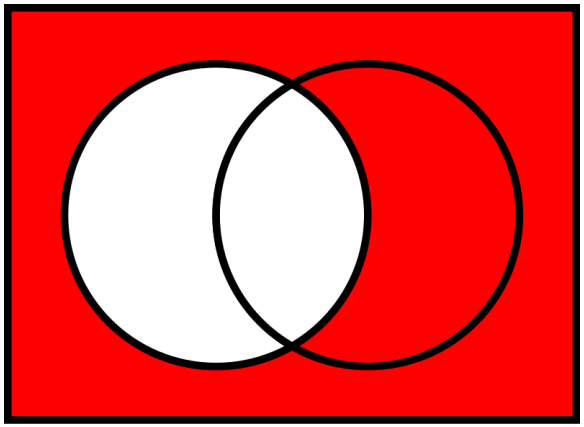
Объединение



Симметричная разность

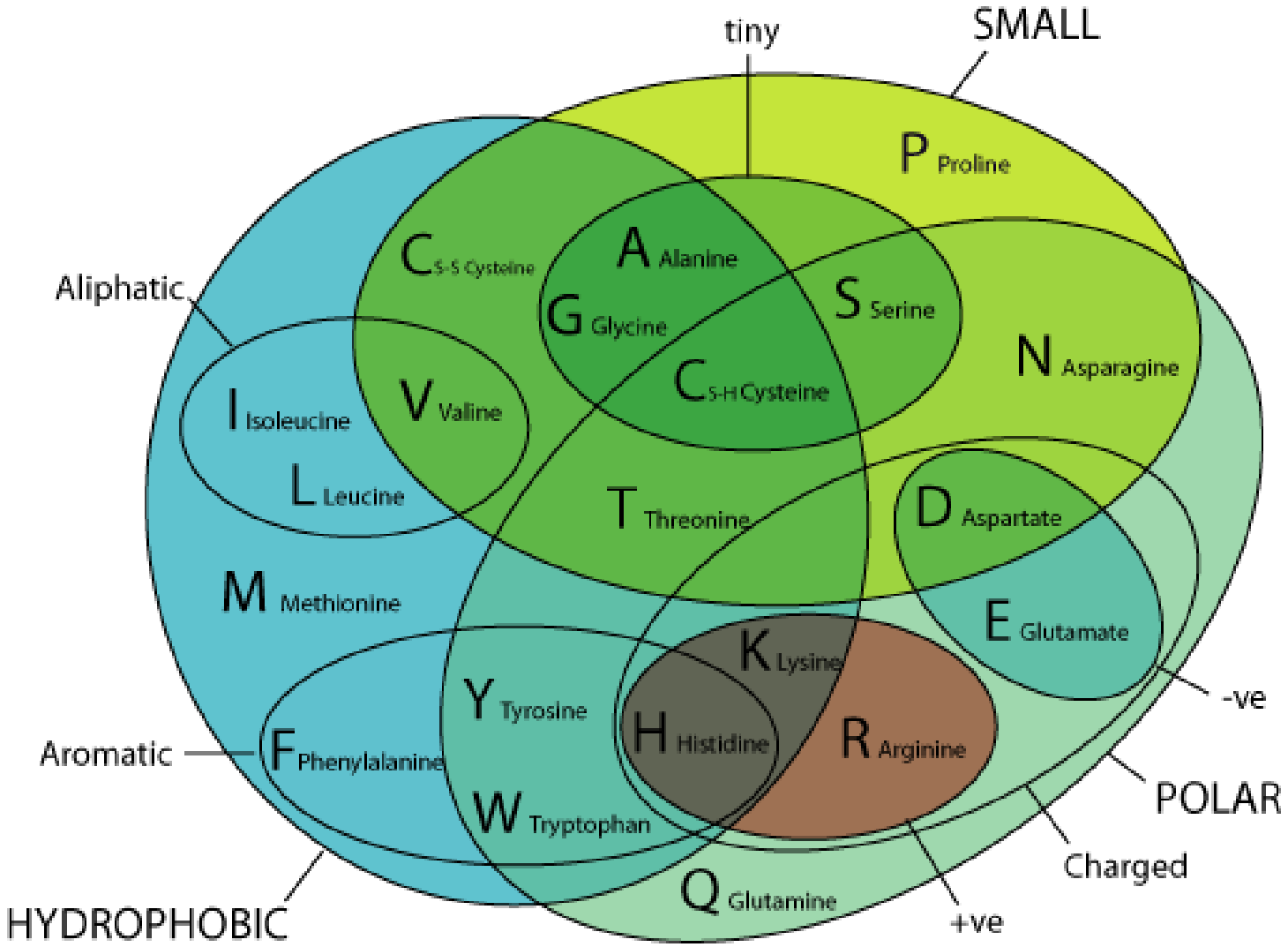


Относительное дополнение

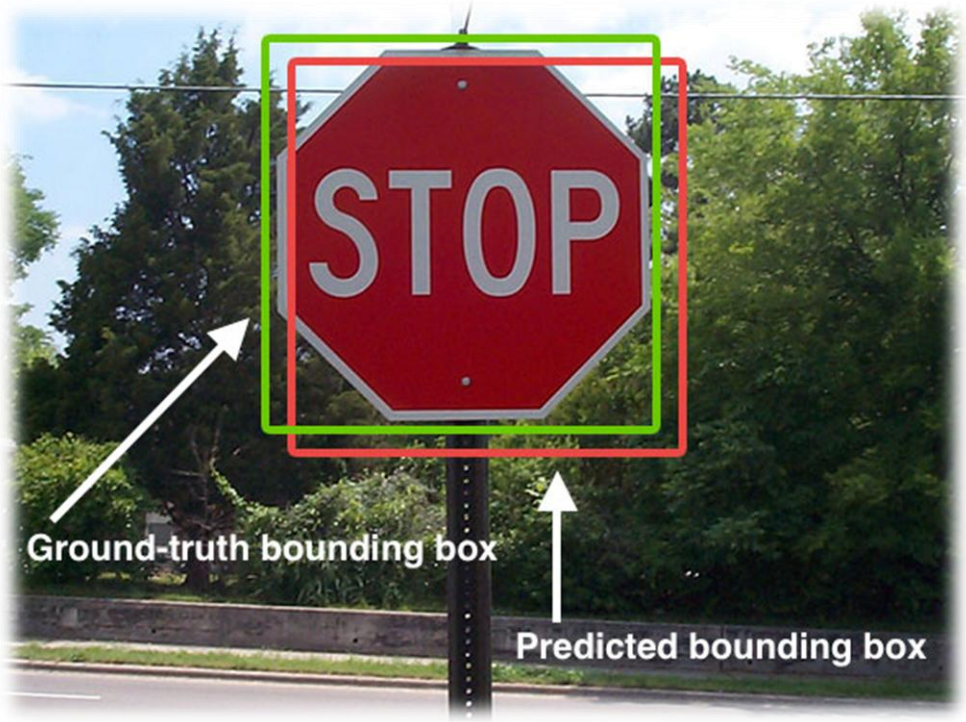


Абсолютное дополнение

О мерах сходства



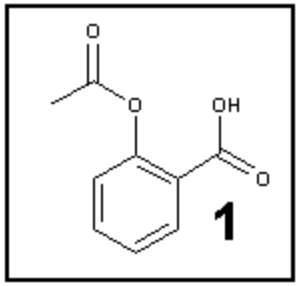
О мерах сходства

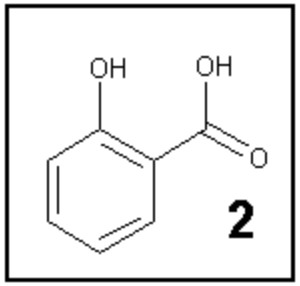


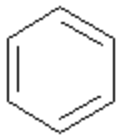
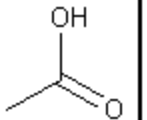
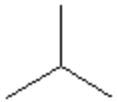
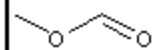
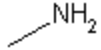
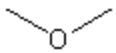
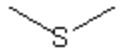
$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Определение сходства молекул

Similarity Searching





1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

TANIMOTO COEFFICIENT = A / (A + B + C)
= 3 / (3 + 2 + 0) = 0.6 or 60%

Определение сходства молекул