

А. А. Рагойша

Информационные технологии в химии

2-й семестр:

Структурные базы данных и структурный поиск информации

Лекция 1

Одномерные формы отображения
химического вещества в базах данных

Литература

1. Andrew R. Leach, Valerie J. Gillet.
An Introduction to Chemoinformatics. – Springer, 2007.
 2. Chemoinformatics: A Textbook.
Edited by Johann Gasteiger and Thomas Engel. – Wiley-VCH, 2003.
-
1. B. A. Bunin, B. Siesel, G. A. Morales, J. Bajorath.
Chemoinformatics: Theory, Practice, & Products. – Springer, 2007.
 2. Chemical Information for Chemists: A Primer.
Edited by Judith N. Currano and Dana L. Roth. - RSC Publishing, Cambridge, UK, 2014.

Одномерная форма отображения химического вещества

Способы многообразны, в том числе:

- Тривиальное название.
- Систематическое название.

Номенклатура ИЮПАК:

от названия к составу – однозначно,

от состава к названию – бывают варианты,

сложные правила,

длинные названия веществ.

- Молекулярная (брутто-) формула: $C_3H_6O_2$
- Рациональная формула: C_2H_5COOH
- Рациональная формула в полуразвернутом виде:
 CH_3CH_2COOH

Вещество в справочнике, каталоге

Блок идентификаторов

IUPAC name	Ethane	[hide]
Wikipedia		
Identifiers		
CAS number	74-84-0	✓
PubChem	6324	
EC number	200-814-8	
UN number	1035	
RTECS number	KH3800000	
SMILES	CC	[hide]
InChI	1/C2H6/c1-2/h1-2H3	[hide]
ChemSpider ID	6084	

14420

Ethane, 99+%

CH₃-CH₃

110g 172.00

C₂H₆ FW 30.07 [74-84-0] mp -172° bp -88° fp -135° RTECS KH3800000

EINECS 200-814-8 TSCA Merck 12,3767 BRN 1730716 UN 1035

EXTREMELY FLAMMABLE / KEEP COLD



R:12

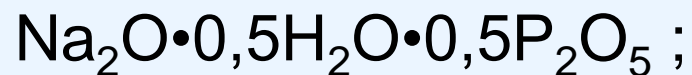
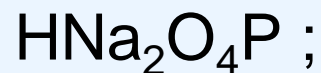
S:9-16-33

Каталог реактивов
Lancaster

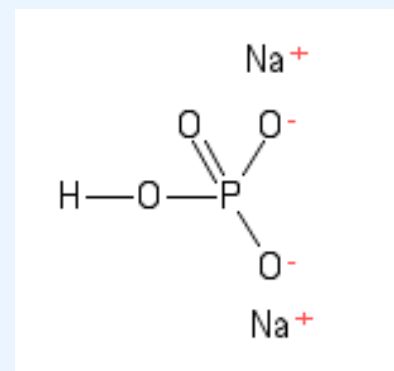


Это:

гидроортофосфат натрия,
натрий-гидрофосфат,
кислый фосфат натрия двузамещенный,
динатрийгидротетраоксофосфат(V),
disodium orthophosphate,
hidrogenoortofosfato de disodio,



и т. д.



Регистрационный номер

применяется для однозначного обозначения химического вещества в большой базе данных (и коммерческой, и некоммерческой).

Формат не стандартизирован.

Для ориентации пользователя –
буквенный идентификатор,

например: PubChem ID 6324

UN number, UN ID -

(United Nations)

цифровой код, который приписывается веществу или изделию, представляющему опасность на транспорте.

Пример: UN 1036

Вещество может иметь несколько UN ID (разбавленный – концентрированный раствор, низкая – высокая температура).

RTECS number -

(Registry of Toxic Effects of Chemical Substances)

цифровой код вещества в базе данных *RTECS*, содержащей сведения о токсичности веществ (США, коммерческая база данных).

Пример: RTECS KN3800000

EC number, EC-No, EC#

(The **E**uropean **C**ommunity number) –
цифровой код, который приписывается
химическому товару в странах ЕС.

Пример: EC number 200-814-8

EINECS #

E number

код, который присваивается
пищевым добавкам в странах ЕС

Примеры:

E260 (уксусная кислота)

E925 (хлор)

CAS Registry Number

(CAS = Chemical Abstracts Service)

CAS Registry Number

CASRN, CAS RN, CAS Number, CAS#

— номер, под которым химическое вещество (или смесь веществ) зарегистрировано в *Chemical Abstracts Service*.

Присваивается в хронологическом порядке, химический смысл не закладывается.

Формат: **7558-79-4**.

CAS Registry Number: ИСПОЛЬЗОВАНИЕ В ПОИСКЕ

В запросе в научных,
коммерческих
базах данных.

NIST →


В Google – тоже

Пример. Разные списки результатов:

Search for Species Data by CAS Registry Number

Please follow the steps below to conduct your search:


1. Enter a registry number (e.g., 74-82-8):
2. Select the desired units for thermodynamic data:



Web [+ Show options...](#) Results 1 - 10 of about 1,690,000 for ethane.

[Ethane - Wikipedia, the free encyclopedia](#)
Ethane is a chemical compound with chemical formula C₂H₆. It is the only two-carbon alkane that is an aliphatic hydrocarbon. At standard temperature and ...
[History](#) - [Chemistry](#) - [Production](#) - [Uses](#)
en.wikipedia.org/wiki/Ethane - [Cached](#) - [Similar](#)

[Ethane](#)
21 May 2005 ... The hydrocarbon **ethane** consists of two carbons



Web [+ Show options...](#) Results 1 - 10 of about 217,000 for "74-84-0".

[74-84-0, CAS Number: 74-84-0](#)
74-84-0 - chemical information, properties, structures, articles, patents and more chemical data.
www.chemindustry.com/chemicals/0188253.html - [Cached](#)

[ethane, CAS Number: 74-84-0](#)
ethane - chemical information, properties, structures, articles, patents and more chemical data.
www.chemindustry.com/chemicals/0188270.html - [Cached](#)

CAS Registry Number: особенности регистрации

Свой *CASRN* присваивается каждому химическому **объекту**, например:

Цис-1,2-дихлорэтен,
транс-1,2-дихлорэтен,

1,2-дихлорэтен (без указания изомерии)

C_2HDCl_2 (без указания изомерии)

цис- $C_2D_2Cl_2$

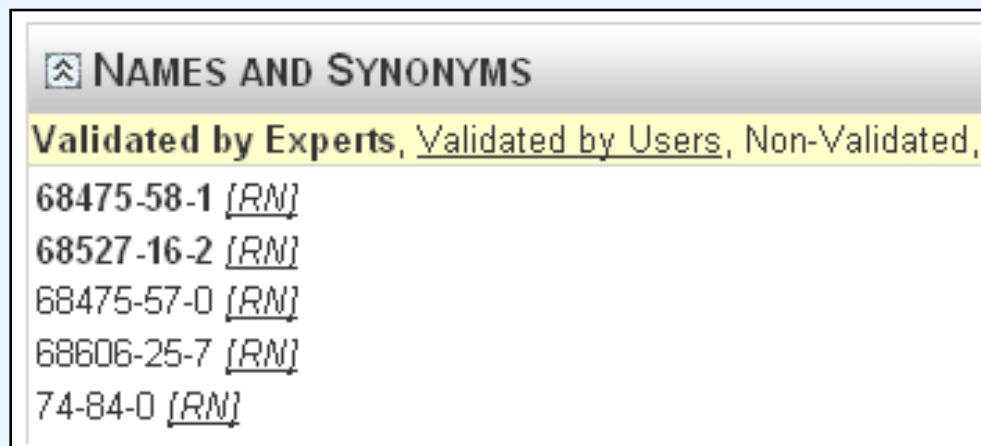
и т. д.

(в том числе, какая-нибудь особенно важная смесь, содержащая это вещество)

CAS Registry Number проблемы использования

а) Доступ к полному списку *CASRN* – платный.

б) Плодятся неточности в бесплатном WWW.




в) *CASRN* ↔ химический объект и
CASRN ↔ химическое вещество.

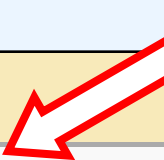
CAS Registry Number

где найти правильный код?

- а) В редактируемых научных базах данных.
- б) (В каталогах реактивов) (?).
- в) Официальный краткий список:
Common Chemistry (<http://www.commonchemistry.org/>)
- г) В Wikipedia (в статье о конкретном веществе).

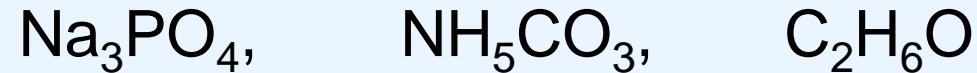
Identifiers	
CAS number	74-84-0 

важно!



Дополнительный материал:
CAS Registry Number и справочник Common Chemistry
http://www.abc.chemistry.bsu.by/bulchinf/2009_1_6-8.pdf

Молекулярная (брутто-) формула



В каком порядке располагают брутто-формулы в формульном указателе?

Брутто-формула: запись по системе Хилла (Hill, 1900)

Порядок расположения элементов:

а) Если вещество содержит углерод:

1. Углерод,
2. Водород,
3. Прочие элементы в алфавитном порядке.

Пример: C_7H_7ClO - это $C_6H_5-O-CH_2Cl$.

б) Если в веществе углерод отсутствует:

Все элементы, в том числе водород, перечисляются в алфавитном порядке.

Примеры: BaH_2O_2 - это $Ba(OH)_2$
 HNO_4Zn - это $(ZnOH)NO_3$

Система Хилла в формульном указателе

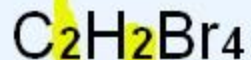
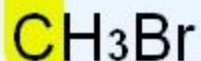
Элемент #1

- по алфавиту

- - по числу атомов;

Элемент #2 ...

Примеры:



Molecular Formula



C



C2



C2H3



C2 H3



C2 H3 Ag1 O2



C2 H3 Al1



C2 H3 Al1 Br3 N1



C2 H3 Al1 Br4 O1



C2 H3 Al1 Cl4 O1



C2 H3 Al1 I3 N1

Двумерная графическая формула
(2D-структура)

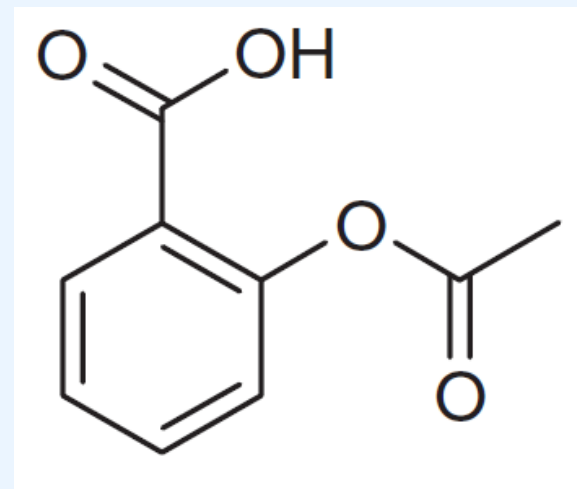
Топология структуры:

вид атомов и порядок их соединения друг с другом.

Топография структуры:

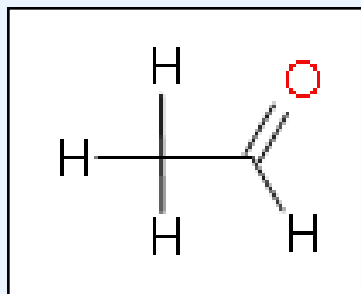
расположение атомов в пространстве.

Двумерная графическая формула отображает топологию структуры.



ацетилсалициловая кислота (аспирин)

Молекулярный графический редактор



GIF
BMP
JPG
etc.

Топология понятна для человека,
но не для компьютера.

Для компьютера:

Молекулярная структура состоит из отдельных элементов.

Каждый элемент имеет свойства

(атом: химический элемент, тип гибридизации,
координаты и т.д.;

химическая связь: длина, кратность и т.д.).

Молекулярный графический редактор создает объекты
с химически значимыми свойствами.

Набор свойств может варьироваться существенно.

Примеры молекулярных редакторов

ISIS/Draw  Symyx Draw (2D)

ChemDraw (2D) }
Chem3D (3D) } ChemOffice

ACD/ChemSketch (2D) }
ACD/3D (3D) } ACDLabs Freeware

Апплеты: JSME, Marvin JS, ...

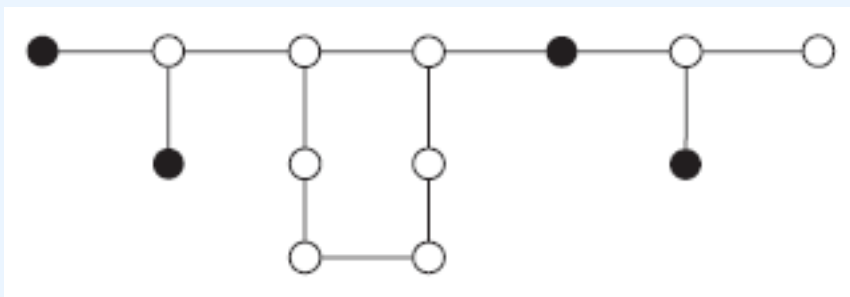
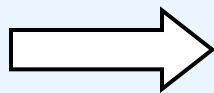
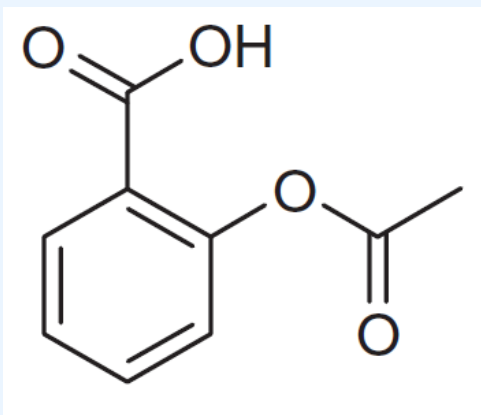
Молекулярный граф

Граф – совокупность объектов со связями между ними.

Молекулярный граф отображает структуру вещества: вершинами являются атомы, а ребрами – химические связи.

Вершины и ребра графа могут иметь (а могут и не иметь) свойства (вершины: название химического элемента; ребра: порядок связи и т. д.).

Абстрактная модель, отображает топологию молекулы (как правило, без атомов водорода).



молекулярный граф

Линейная нотация (линейная запись) –

одномерная форма отображения химического объекта
в виде строки буквенно-цифровых символов

Один из способов отображения
молекулярного графа в форме линейной нотации:

SMILES

SMILES

[СМАЙЛЗ]

на поисковом бланке

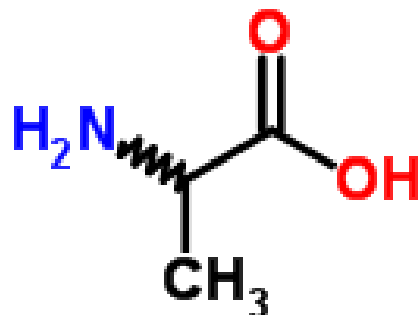
Systematic Name, Synonym, Trade Name,
Registry Number **SMILES** or InChI

Search

в результатах поиска

Пример:

O=C(O)C(N)C



3D

ChemSpider ID:	582
Empirical Formula:	C ₃ H ₇ NO ₂
Molecular Weight:	89.0932
Nominal Mass:	89 Da
Average Mass:	89.0932 Da
Monoisotopic Mass:	89.047678 Da

load save zoom

Systematic Name: 2-aminopropanoic acid

SMILES: O=C(O)C(N)C

SMILES

Simplified Molecular Input Line Entry System – это **система** компьютерной обработки массивов химической информации

Пример SMILES: O=C(O)C(N)C

Код SMILES в **текстовой** строке показывает **состав** вещества и **связи** между атомами.

Модель в основе кодирования: **метод валентных связей**.

Читабельный.

Компактный. *Пример:* Если в базе данных 23 тыс. структур, каждая по 20 атомов – на один атом в среднем используется 1,6 байт.

Разработчик: Daylight Chemical Information Systems

<http://www.daylight.com/smiles/>

Синтаксис SMILES. Атомы.

В общем случае,
атомы отображаются символами химических элементов и
записываются в квадратных скобках,
например: [As].

Без квадратных скобок
можно отображать атомы "органических" элементов
в "низших нормальных" валентных состояниях:

B(III)	C(IV)	N(III, V*)	O(II)	F(I)
		P(III, V),	S(II, IV, VI)	Cl(I)
				Br(I)
				I(I)

* На самом деле – N(IV)

Водород при **этих** атомах, насыщающий свободные
валентности, можно не указывать.

Гидриды

<i>Объект</i>	<i>Строка SMILES</i>
метан CH_4	C
аммиак NH_3	N
вода H_2O	O
сероводород H_2S	S
хлороводород HCl	Cl
арсин AsH_3	[AsH3]

Водород -
в неявной
форме

Водород -
в явной
форме

Ковалентная химическая связь

Соседние атомы записывают рядом.

Ковалентная связь отображается так:

одинарная никак (иногда -)

двойная =

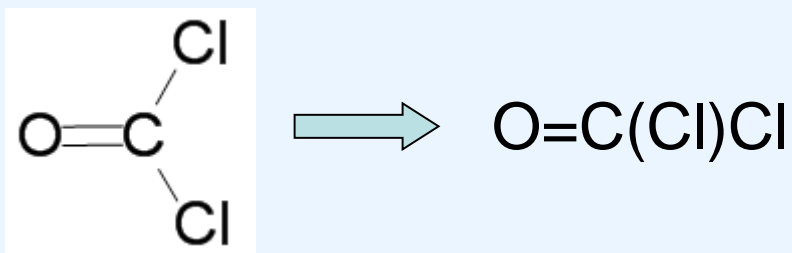
тройная #

Например:

<i>Объект</i>	<i>Строка SMILES</i>
этан CH_3CH_3	CC
пропан $\text{CH}_3\text{CH}_2\text{CH}_3$	CCC
углекислый газ	O=C=O
синильная кислота	C#N

Боковые цепи (заместители)

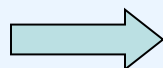
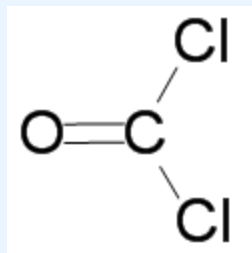
Боковую цепь указывают в круглых скобках после символа того атома, к которому она присоединена.



$\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_2 \\ \\ \text{H}_3\text{C}-\text{CH}_2-\text{N}-\text{CH}_2-\text{CH}_3 \end{array}$	$\begin{array}{c} \text{CH}_3 \quad \text{O} \\ \quad \\ \text{H}_3\text{C}-\text{CH}-\text{C}-\text{OH} \end{array}$	$\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_2 \quad \text{CH}_3 \\ \quad \\ \text{CH}_2 \quad \text{CH}_2-\text{CH}_3 \\ \quad \\ \text{H}_2\text{C}=\text{CH}-\text{CH}-\text{CH}-\text{CH}_2-\text{CH}_2-\text{CH}_3 \end{array}$
<chem>CCN(CC)CC</chem>	<chem>CC(C)C(=O)O</chem>	<chem>C=CC(CCC)C(C(C)C)CCC</chem>

Стандартная (каноническая) запись

Возможны многочисленные варианты записи:



O=C(Cl)Cl , ClC(Cl)=O
ClC(=O)Cl , C(Cl)(Cl)=O
и т. д.

Все они считаются правильными и каждый может использоваться как поисковый термин.

Один из вариантов является стандартным (каноническим). Стандартный генерируют по правилам, которые мы изучать не будем.

В базах данных информация хранится на основе канонических форм.

Компьютер сам преобразовывает запись пользователя в каноническую форму.

Ионы и ионные соединения

Заряд иона указывают внутри квадратных скобок.

<i>Объект</i>	<i>Строка SMILES</i>
Fe^{2+}	[Fe+2]
H_3O^+	[OH3+]
NH_4^+	[NH4+]

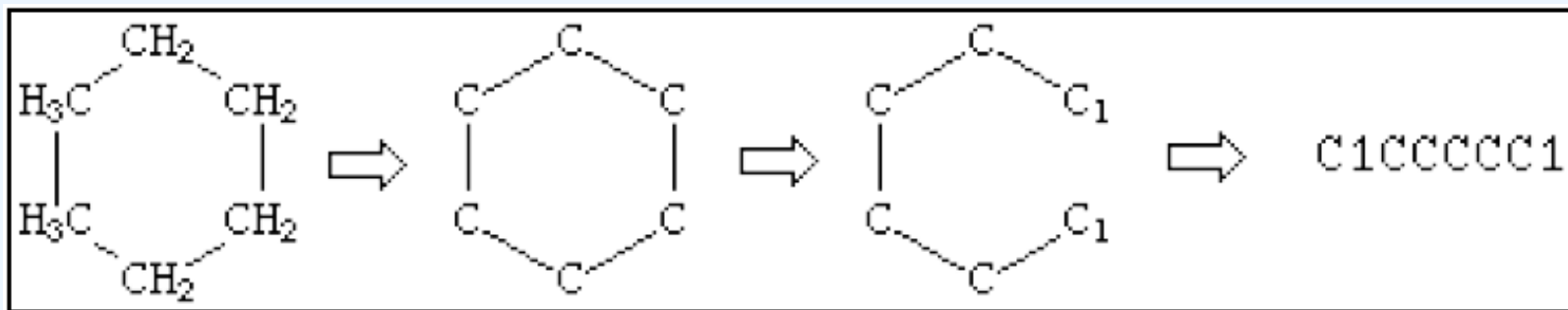
Обратим
внимание
на порядок
записи цифр
и знака "плюс"

Точкой отделяют автономные частицы.
Например, катион и анион.

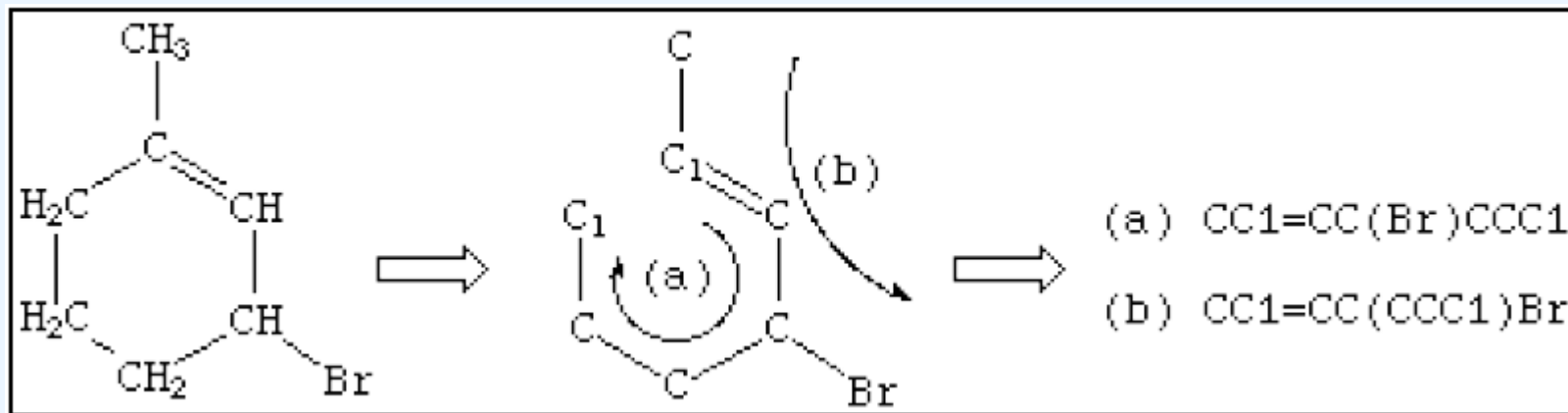
<i>Объект</i>	<i>Строка SMILES</i>
NaOH	[Na+].[OH-]

Циклы

Два атома цикла нумеруют одним и тем же числом и связь между этими атомами условно разрывают:



Допускаются варианты выбора основной цепи:



Ароматические соединения

Химические символы атомов, образующих ароматические связи, записывают **строчными** буквами.

Пример 1.

циклогексан

C1CCCCC1

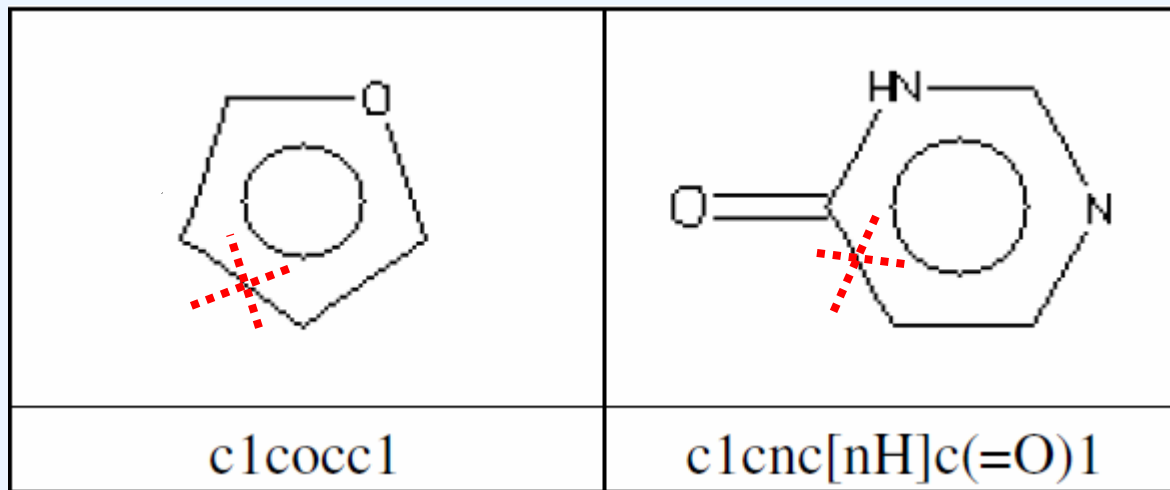
бензол

c1ccccc1

фенол

Oc1ccccc1

Пример 2.

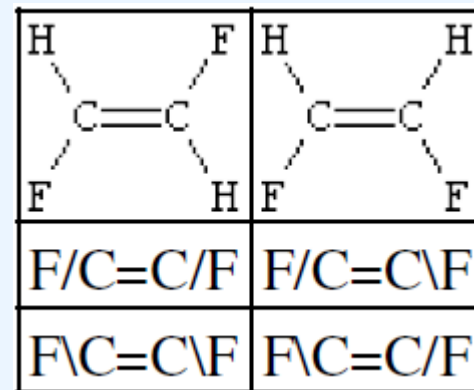


Пространственные изомеры

При двойной связи:

транс- \ = \ или / = /

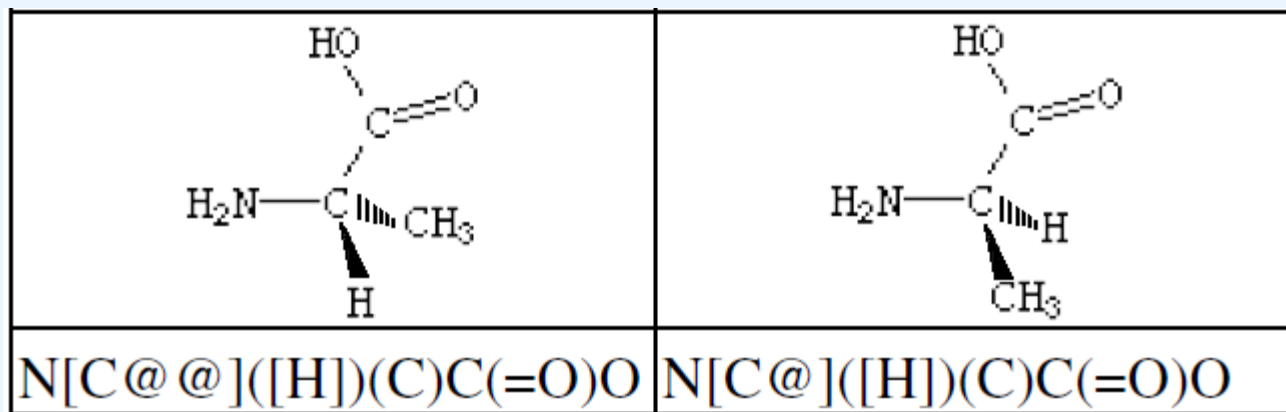
цис- \ = / или / = \



У тетраэдрического атома C:

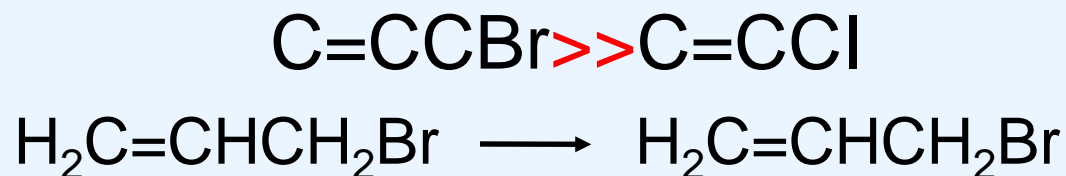
@ против часовой стрелки

@@ по часовой стрелке

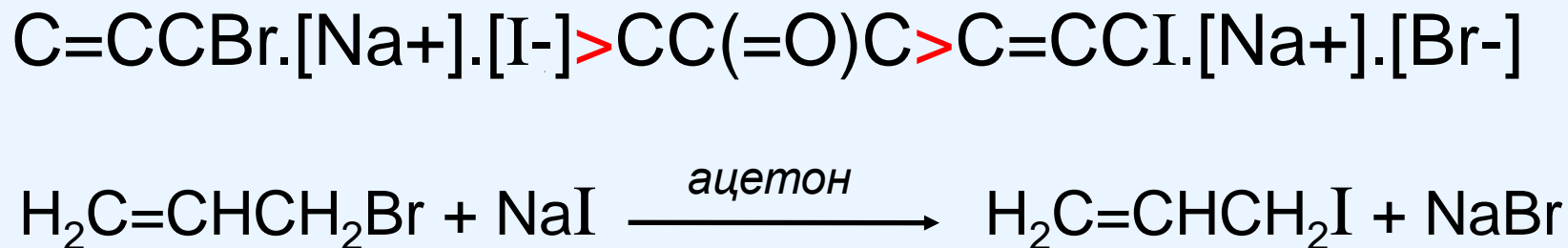


Схемы реакций

Реагент >> продукт



Реагент > агент > продукт



SMILES и Google

Код SMILES – буквенно-цифровой;
в принципе может быть компонентом запроса
для универсальной поисковой системы.



31 млн.

[DrugBank: Showing Mitomycin \(DB00305\)](#)
23 Jun 2009 ... Canonical SMILES,
COC12C3NC3CN1C1=C(C2COC(N)=O)C(=O)C(N)=C(C)C1=O. Drug Category. Alkylating
Agents; Antibiotics, Antineoplastic ...
www.drugbank.ca/drugs/DB00305 - [Cached](#) - [Similar](#)

[TR000 C1C\(C\)C1](#) [TR001 C12\(C\)C3\(C\)C\(=O\)C4\(C\)C1\(C\)C5\(C\)C\(C\)C\(C\)C1 ...](#)
... c1cc(O)c2C(=O)C3=C(O)C4(O)C(=O)C(C(N)=O)=C(O)C(N(C)C)C4CC3C(C)(O)c2c1 [TR345](#)
[\[O-\]\[N+\]\(=O\)c1cc\(ccc1O\)\[As\]\(=O\)\(O\)O](#) [TR346](#) C(C)C [TR347](#) C1(C)=CCC(CC1)C(=C)C ...
www.predictive-toxicology.org/data/ntp/corrected_smiles.txt - [Cached](#)

Поиск работает, но есть проблемы:

- а) запрос как фрагмент кода; как сумма терминов.
- б) регистр букв тоже может иметь значение.

InChI

InChI

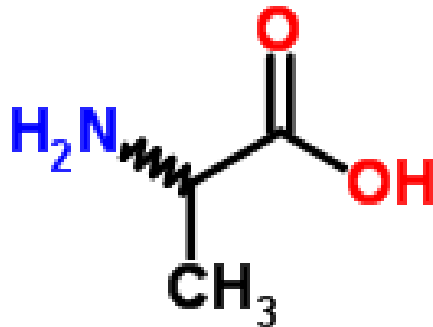
[інчи]

Systematic Name, Synonym, Trade Name,
Registry Number, SMILES or InChI

на пошковом бланке

Search

в результатах поиска



ChemSpider ID: 582
Empirical Formula: C₃H₇NO₂
Molecular Weight: 89.0932
Nominal Mass: 89 Da
Average Mass: 89.0932 Da
Monoisotopic Mass: 89.047678 Da

load save zoom

Systematic Name: 2-aminopropanoic acid

SMILES: O=C(O)C(N)C

InChI: InChI=1/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)

InChI - IUPAC International Chemical Identifier

InChI – международный текстовый идентификатор химического объекта; это компьютеризованный вариант систематического названия.

- Некоммерческий продукт.
- Использование разрешено без ограничений.
- Специалист может разобраться в строке InChI.

Пример: этанол $\text{CH}_3\text{CH}_2\text{OH}$

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

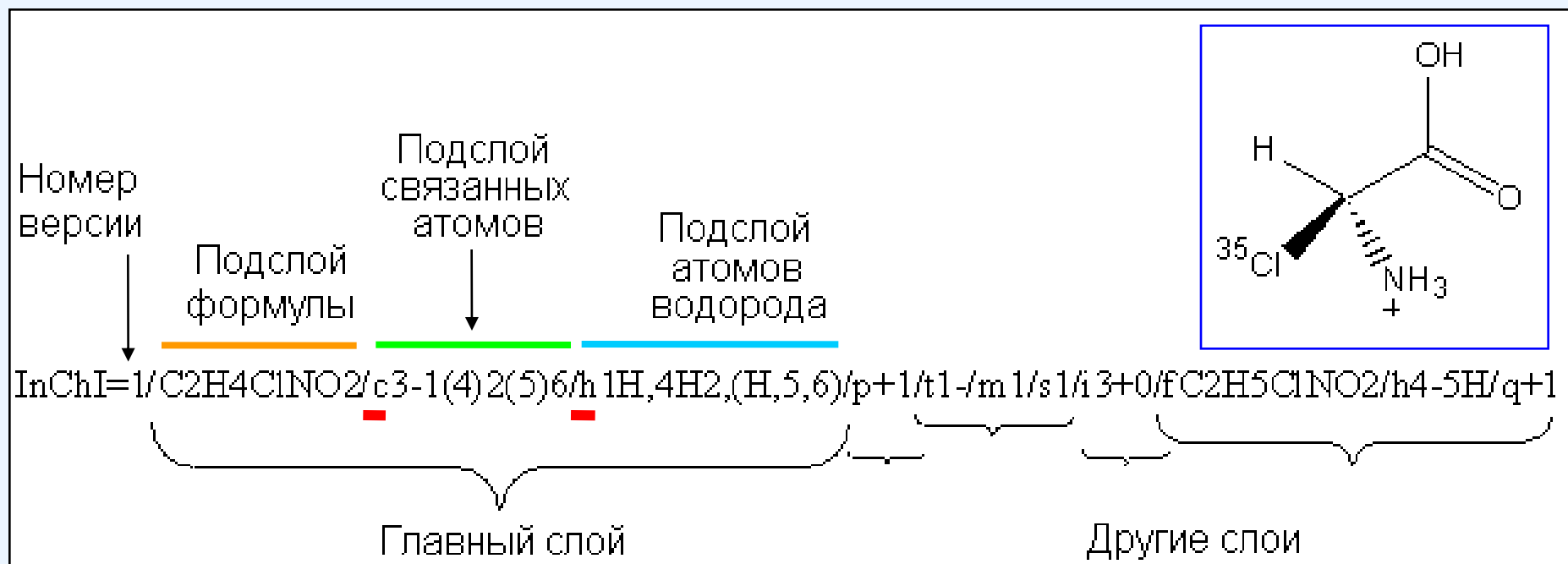
Но! Не человек, а компьютер должен генерировать код или из кода генерировать структурную формулу.

Модульная структура InChI

Код состоит из "слоев" и "подслоев".

6 слоев: главный; заряды; стереохимия; изотопный состав, ...

Подслои главного слоя: формула, перечень связанных атомов, распределение атомов водорода.

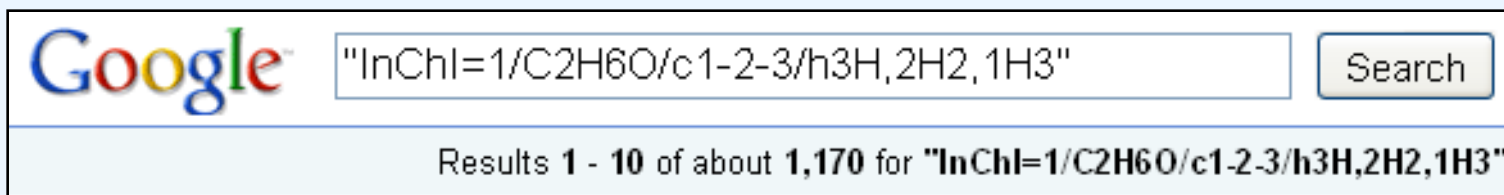


InChI

Модульная структура:

- гибко учитывает объем сведений о веществе,
- позволяет целенаправленно проводить поиск.
- Формула – единственный обязательный элемент кода.
- Новые данные о веществе дописывают в конце кода.

Текстовая строка – значит InChI можно использовать в запросе обычной поисковой системы:



InChIKey

[инчи-ки]

Код InChI громоздок, занимает много места в базе данных.
Для удобства компьютера его преобразовывают в InChIKey.

Скелет молекулы
– подслои /с

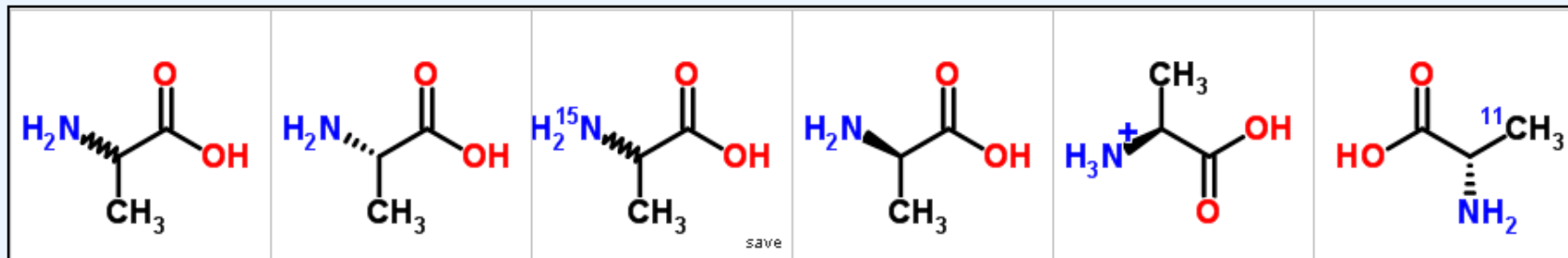
Остальные
слои

Дополнит.
индикаторы

InChIKey=QNAYBMKLOCPYGGJ-UHFFFAOYSA-N

Растет самостоятельная ценность InChIKey:

- аналог DOI, но для молекулярных структур,
- уникальный поисковый термин для Google,
- первая часть кода – для поиска структур-аналогов.



Пример: InChIKey в запросе Google



Web [+ Show options...](#) Results 1 - 10 of about 162 for **"InChIKey=QNAYBMKLOCPYGJ"**.

[L-alanine \(CHEBI:16977\)](#)
17 Oct 2009 ... InChIKey. **InChIKey=QNAYBMKLOCPYGJ-SNQCPAJUDI**.
InChIKey=QNAYBMKLOCPYGJ-SNQCPAJUDI. SMILES. C[C@H](N)C(O)=O. C[C@H](N)C(O)=O. Formula, Source ...
www.ebi.ac.uk/chebi/searchId.do?chebId=CHEBI:16977 - [Cached](#) - [Similar](#)

[alanine \(CHEBI:16449\)](#)
17 Oct 2009 ... InChIKey. **InChIKey=QNAYBMKLOCPYGJ-JSWHHWTPCH**.
InChIKey=QNAYBMKLOCPYGJ-JSWHHWTPCH. SMILES. CC(N)C(O)=O. CC(N)C(O)=O.
Formula, Source ...
www.ebi.ac.uk/chebi/searchId.do?chebId=CHEBI:16449 - [Cached](#)

[+ Show more results from www.ebi.ac.uk](#)

[Compound 3 : Natural amino acids do not require their native tRNAs ...](#)
25 Oct 2009 ... Standard **InChIKey: QNAYBMKLOCPYGJ-REOHCLBHSA-N**. SMILES:
[C@H](C(=O)[OH])([NH2])C. [Next compound](#) | [Previous compound](#) | [Compound index ...](#)
www.nature.com/nchembio/journal/v5/n12/.../nchembio.255_comp3.html