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

Wikipedia

14420 **Ethane**, 99+%
C_2H_6 *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

CH_3–CH_3 **110g** **172.00**

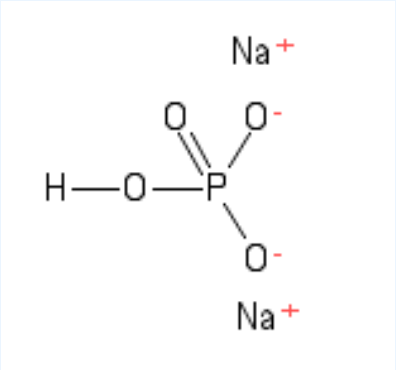
Каталог реактивов Lancaster

Na_2HPO_4

Это:

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

HNa_2O_4P ;
$Na_2O \bullet 0,5H_2O \bullet 0,5P_2O_5$;
$2Na^+ HPO_4^{2-}$
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 KH3800000

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*.

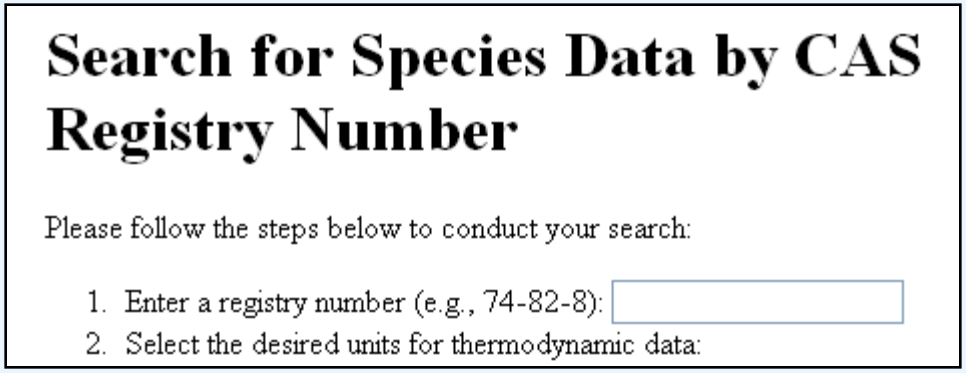
Присваивается в хронологическом порядке, химический смысл не закладывается.

Формат: **7558-79-4**.

CAS Registry Number: использование в поиске

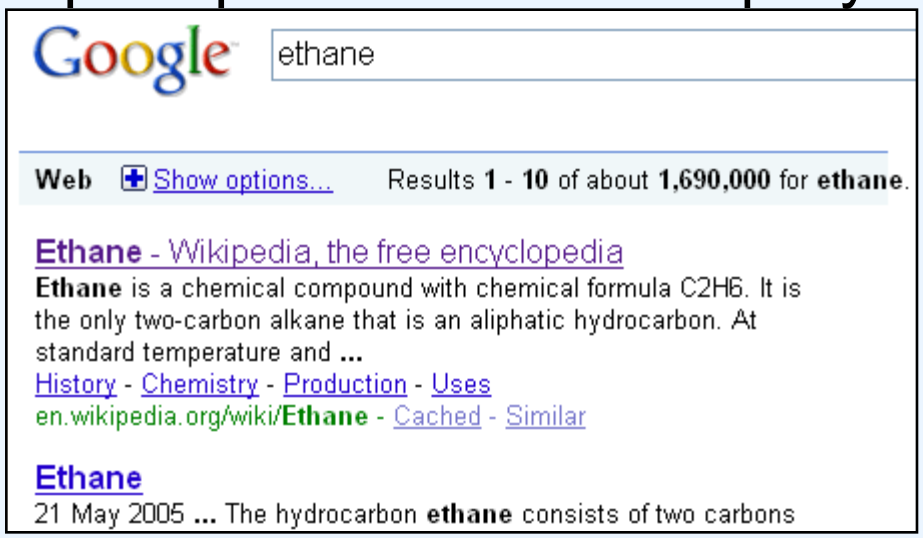
В запросе в научных, коммерческих базах данных.

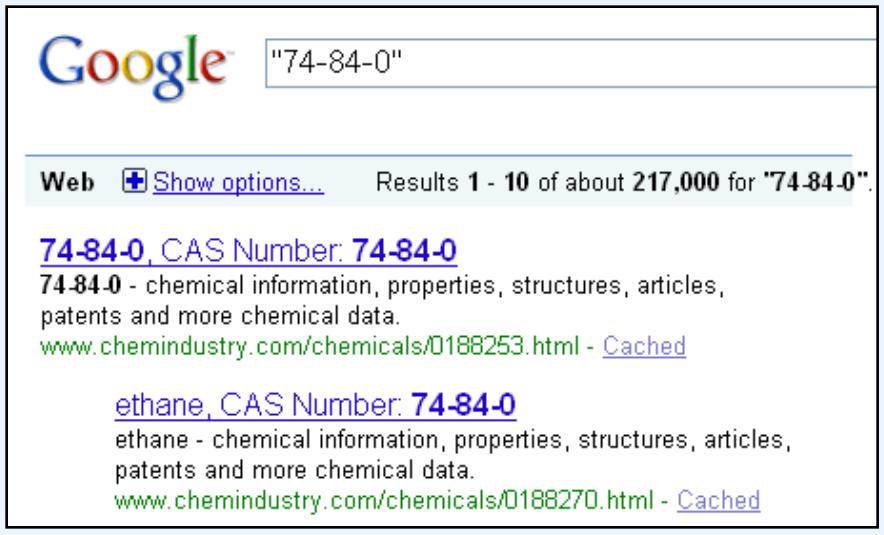
NIST ⇨



В Google – тоже

Пример. Разные списки результатов:





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-$C_2D_2Cl_2$

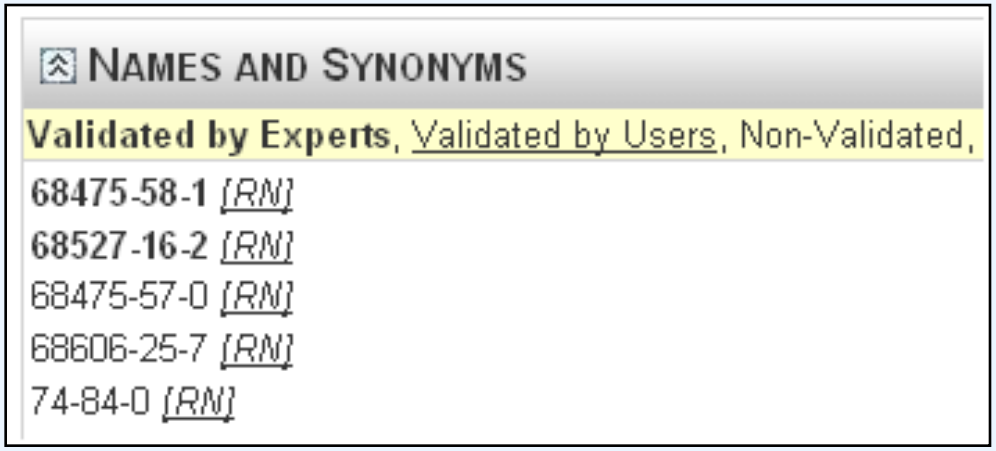
и т. д.

(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.

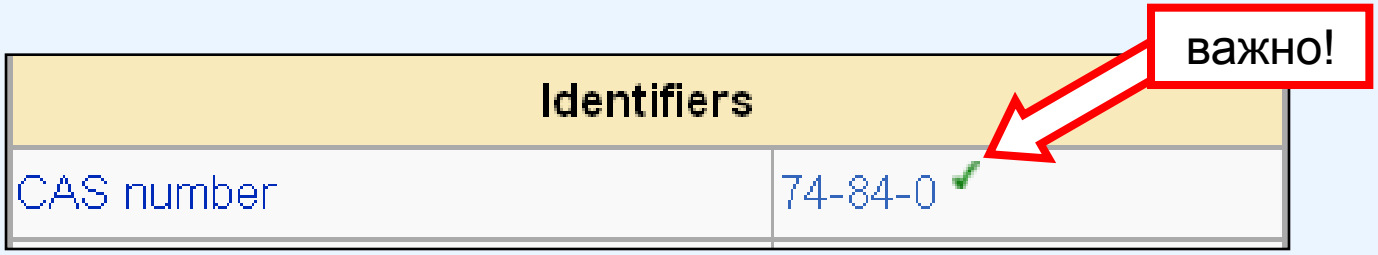


в) *CASRN* ⟺ химический объект *и*
CASRN ⟺ химическое вещество.

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).



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

Молекулярная (брутто-) формула

Na_3PO_4, NH_5CO_3, C_2H_6O

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

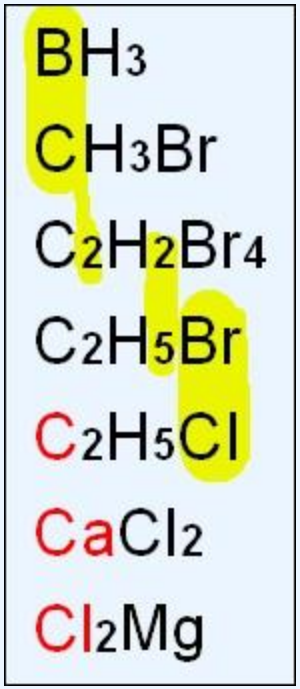
Элемент #1

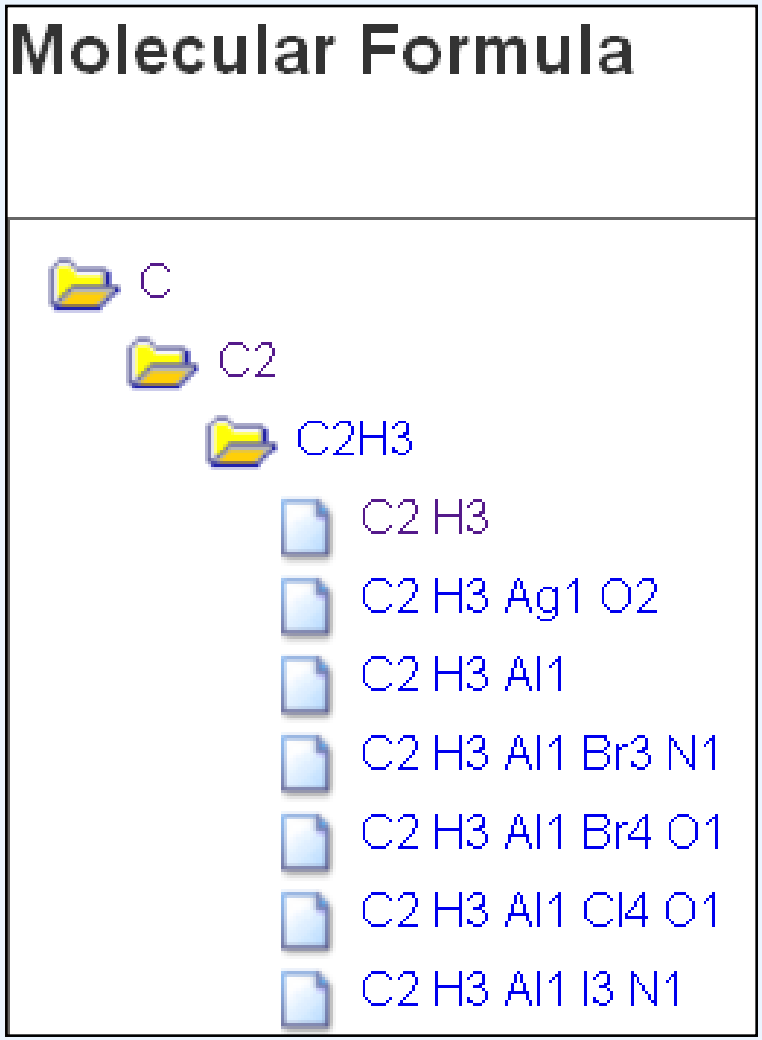
- по алфавиту

- - по числу атомов;

Элемент #2 ...

Примеры:





Двумерная графическая формула

(2D-структура)

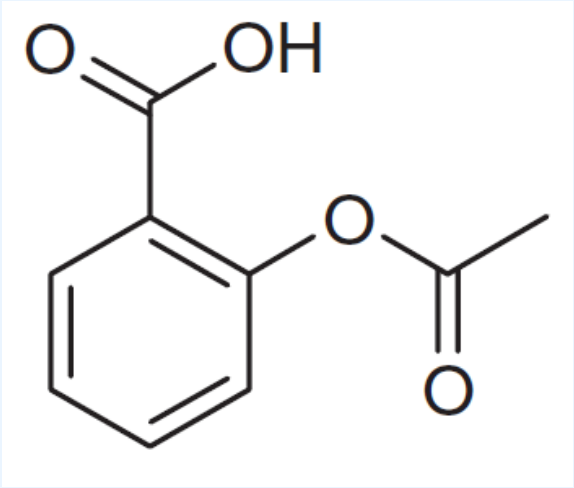
Топология структуры:

вид атомов и порядок их соединения друг с другом.

Топография структуры:

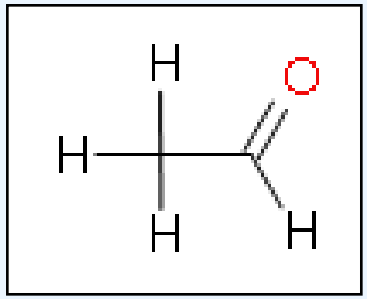
расположение атомов в пространстве.

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, ...

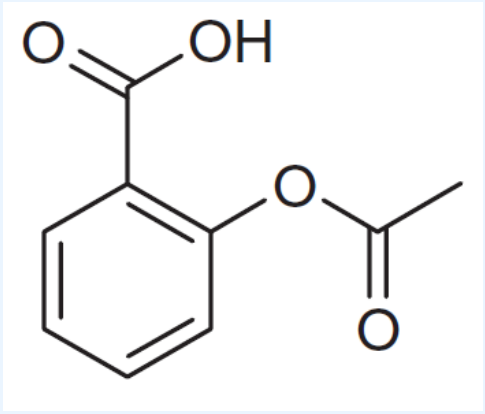
Молекулярный граф

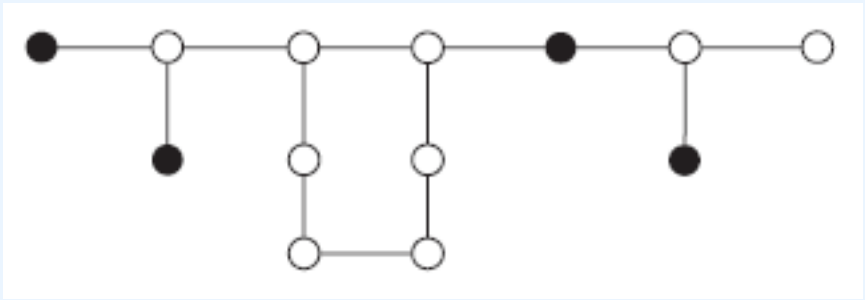
Граф – совокупность объектов со связями между ними.

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) –

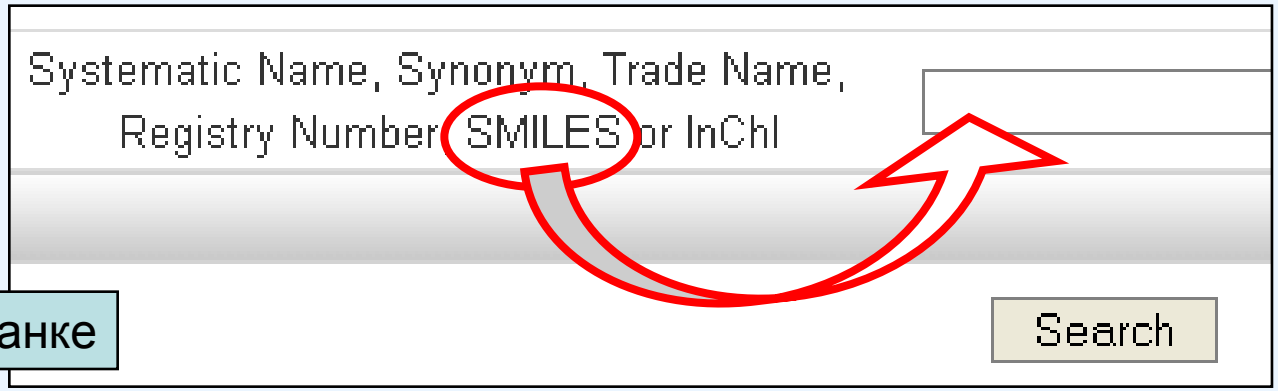
одномерная форма отображения химического объекта в виде строки буквенно-цифровых символов

Один из способов отображения молекулярного графа в форме линейной нотации:

SMILES

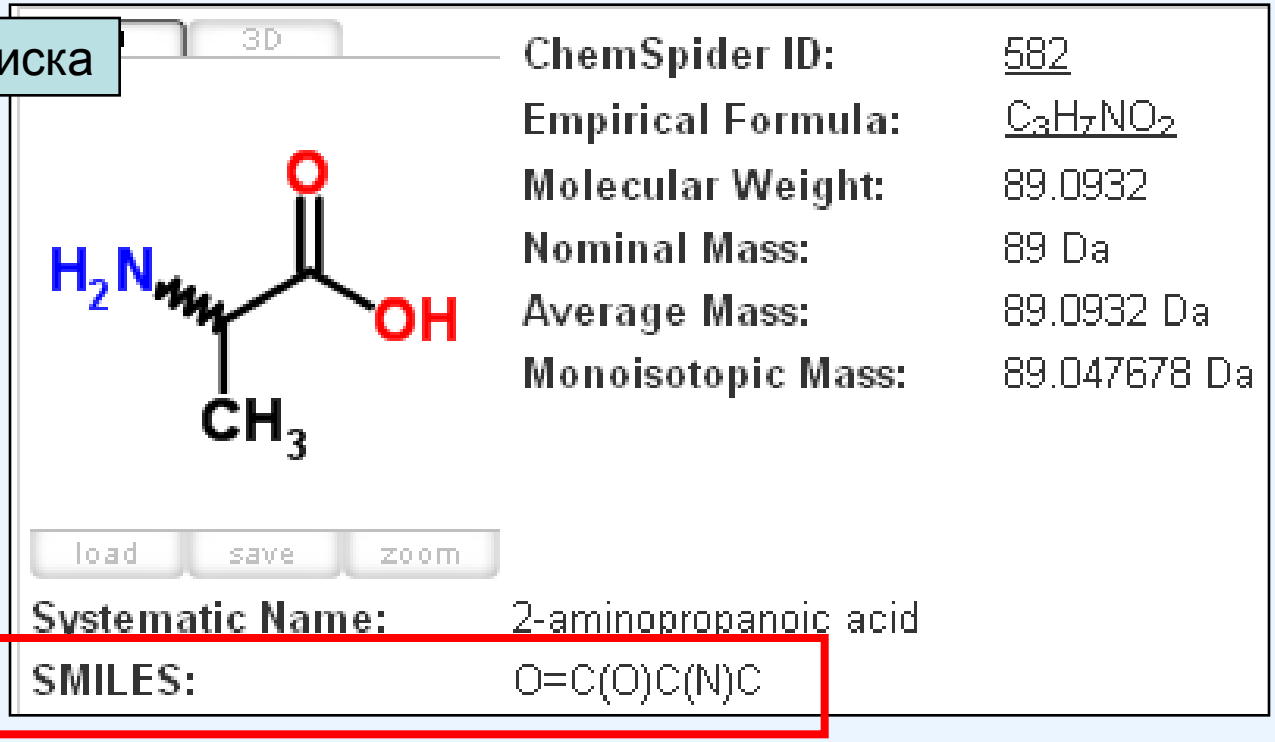
SMILES

[СМАЙЛЗ]



на поисковом бланке

в результатах поиска



Пример:

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

Объект	*Строка SMILES*
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:

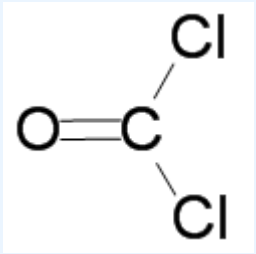
одинарная	никак (иногда -)
двойная	=
тройная	#

Например:

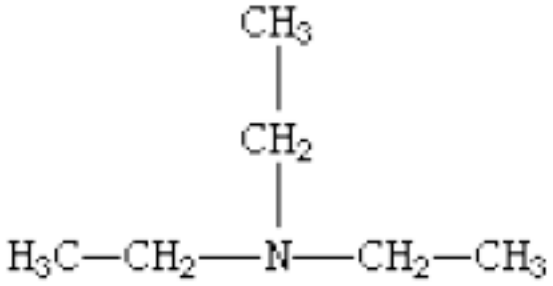
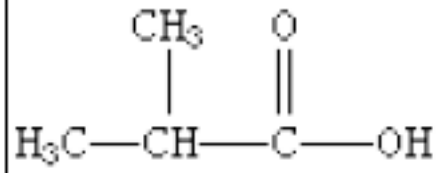
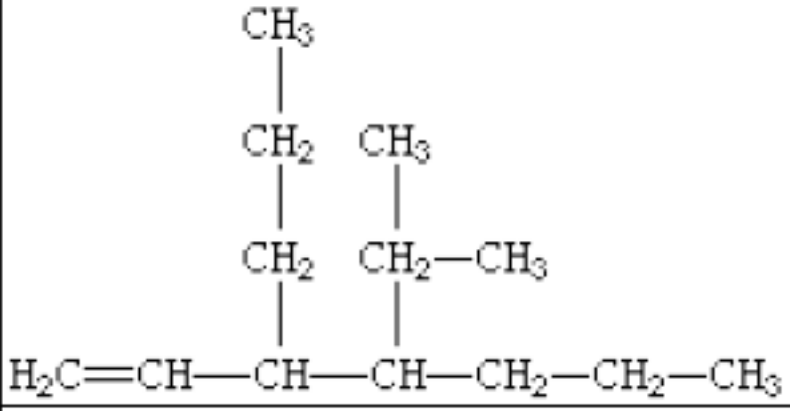
Объект	*Строка SMILES*
этан CH_3CH_3	CC
пропан $CH_3CH_2CH_3$	CCC
углекислый газ	O=C=O
синильная кислота	C#N

Боковые цепи (заместители)

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



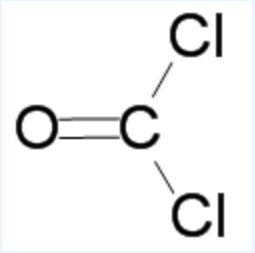
O=C(Cl)Cl

		
CCN(CC)CC	CC(C)C(=O)O	C=CC(CCC)C(C(C)C)CCC

Daylight Theory Manual v. 4.9, 2008.

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.

Объект	*Строка SMILES*
Fe^{2+}	[Fe+2]
H_3O^+	[OH3+]
NH_4^+	[NH4+]

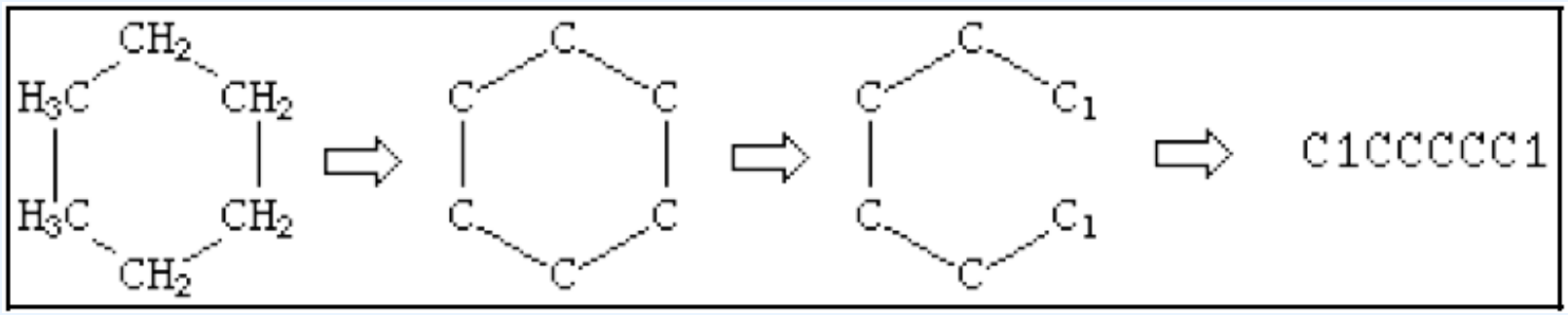
Обратим внимание на порядок записи цифр и знака "плюс"

Точкой отделяют автономные частицы.
Например, катион и анион.

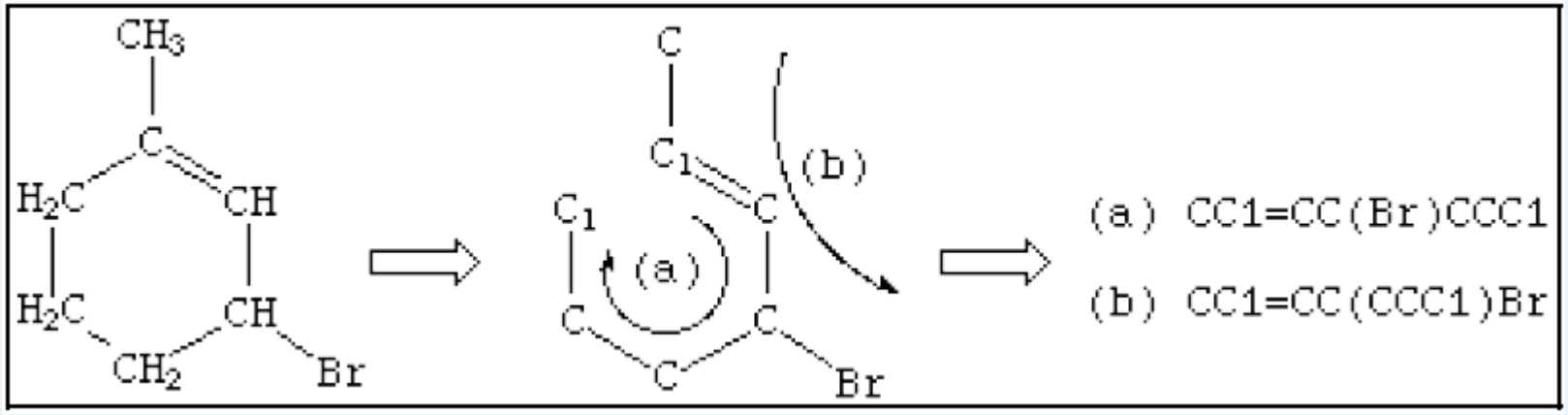
Объект	*Строка SMILES*
NaOH	[Na+].[OH-]

Циклы

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



Допускаются варианты выбора основной цепи:



Daylight Theory Manual v. 4.9, 2008.

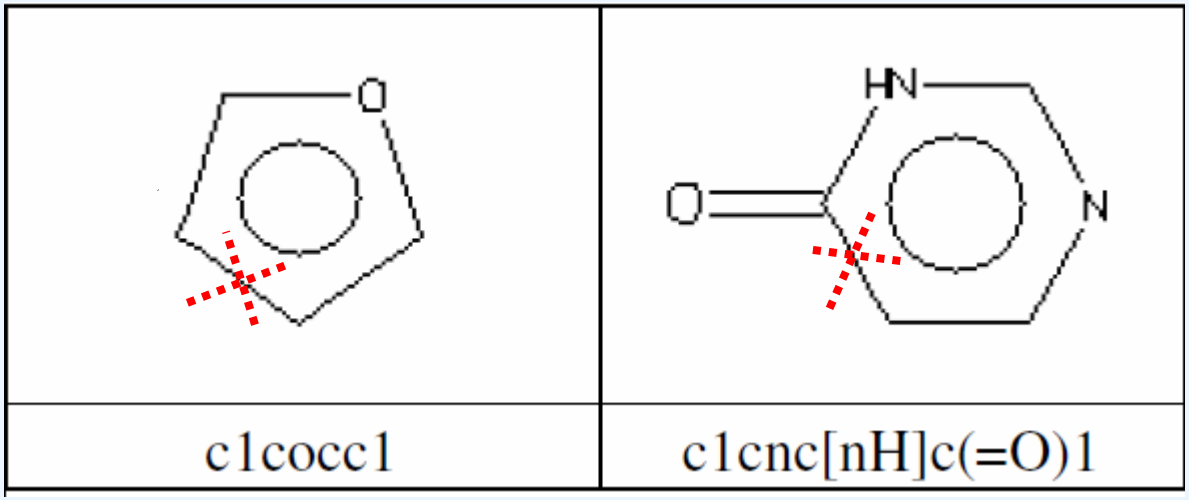
Ароматические соединения

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

Пример 1.

циклогексан	C1CCCCC1
бензол	c1ccccc1
фенол	Oc1ccccc1

Пример 2.



c1cocc1	c1cnc[nH]c(=O)1

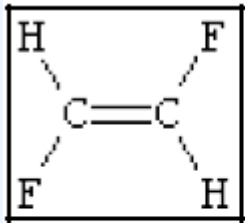
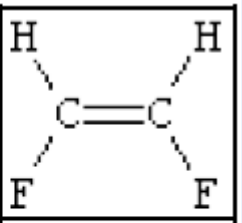
Daylight Theory Manual v. 4.9, 2008.

Пространственные изомеры

При двойной связи:

транс- \ = \ или / = /

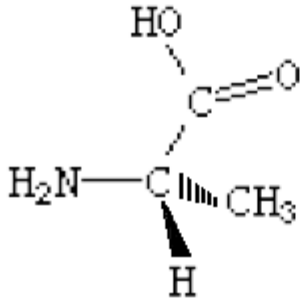
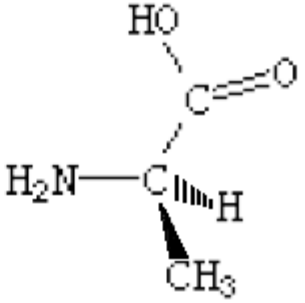
цис- \ = / или / = \

	
F/C=C/F	F/C=C\F
F\C=C\F	F\C=C/F

У тетраэдрического атома С:

@ против часовой стрелки

@@ по часовой стрелке

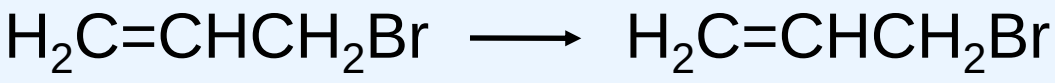
	
N[C@@]([H])(C)C(=O)O	N[C@]([H])(C)C(=O)O

Daylight Theory Manual v. 4.9, 2008.

Схемы реакций

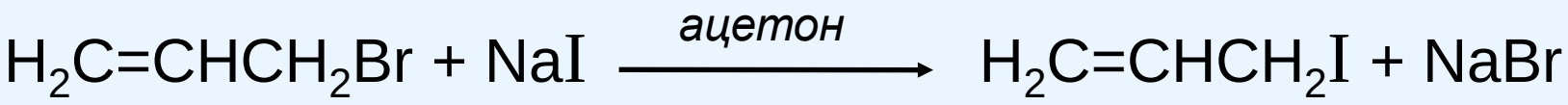
Реагент >> продукт

C=CCBr>>C=CCl



Реагент > агент > продукт

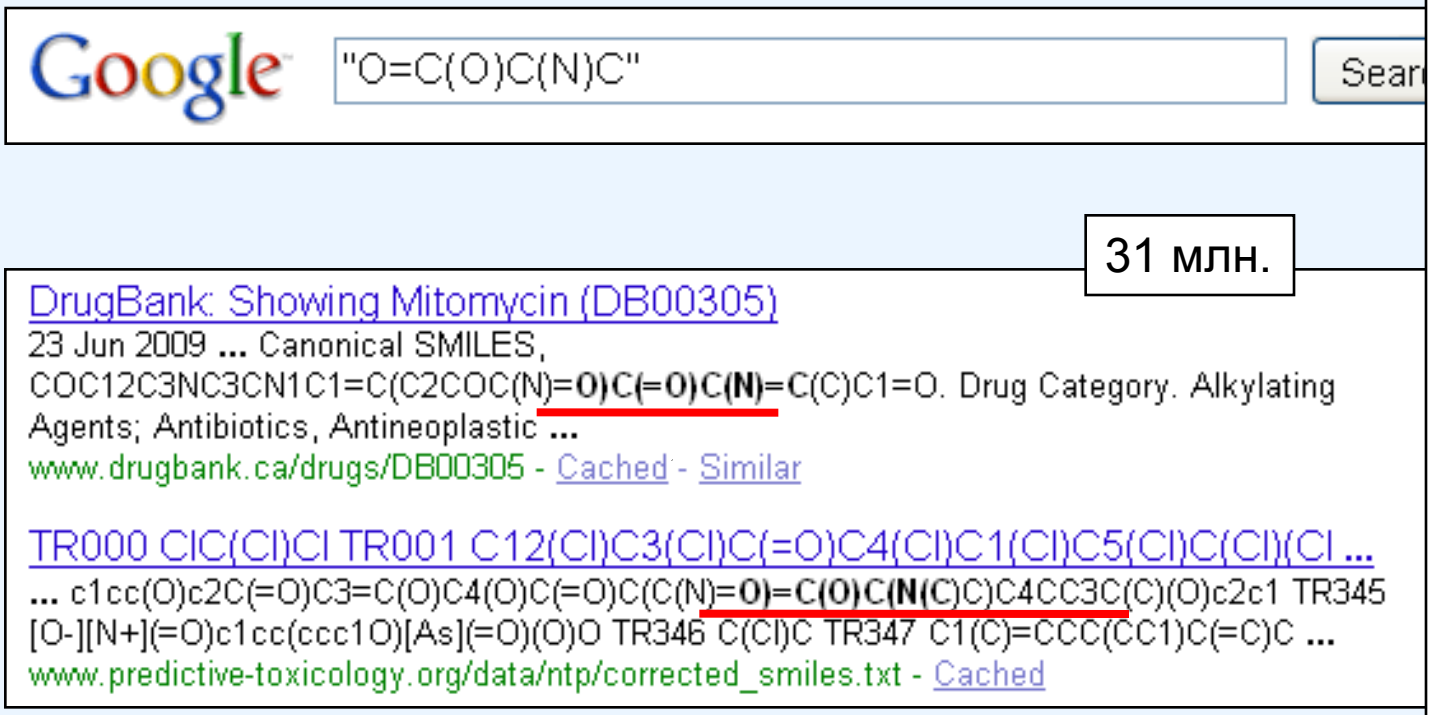
C=CCBr.[Na+].[I-]>CC(=O)C>C=CCI.[Na+].[Br-]



Daylight Theory Manual v. 4.9, 2008.

SMILES и Google

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



Поиск работает, но есть проблемы:

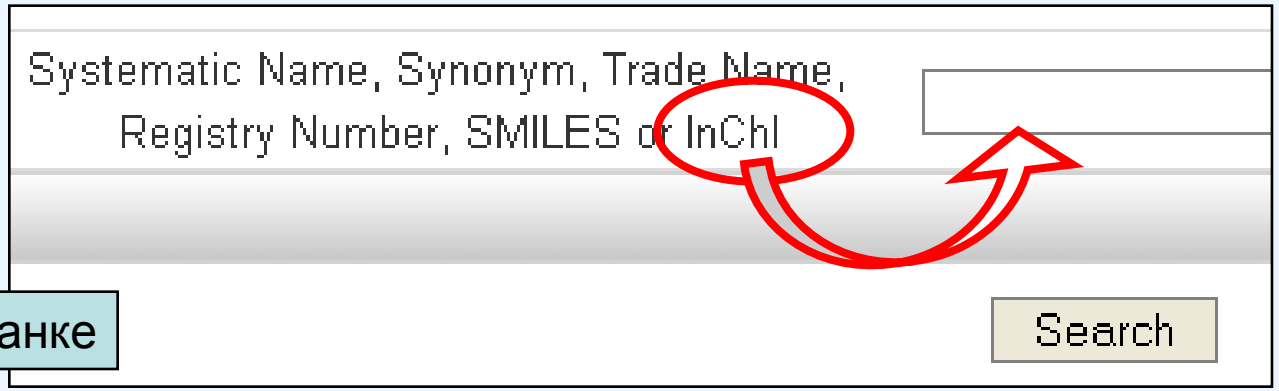
а) запрос как фрагмент кода; как сумма терминов.

б) регистр букв тоже может иметь значение.

InChI

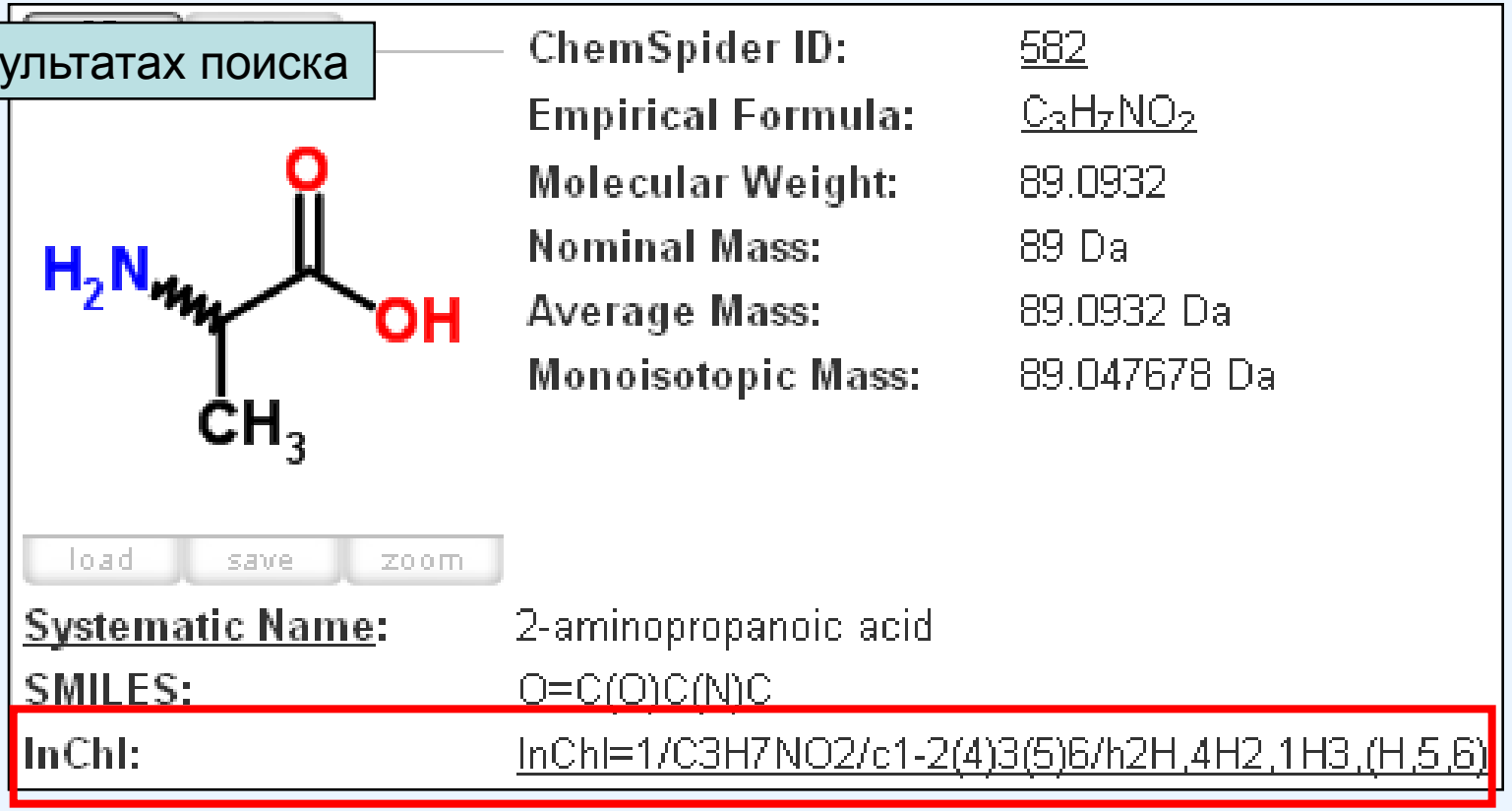
InChI

[и́нчи]



на поисковом бланке

в результатах поиска



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 CH_3CH_2OH

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

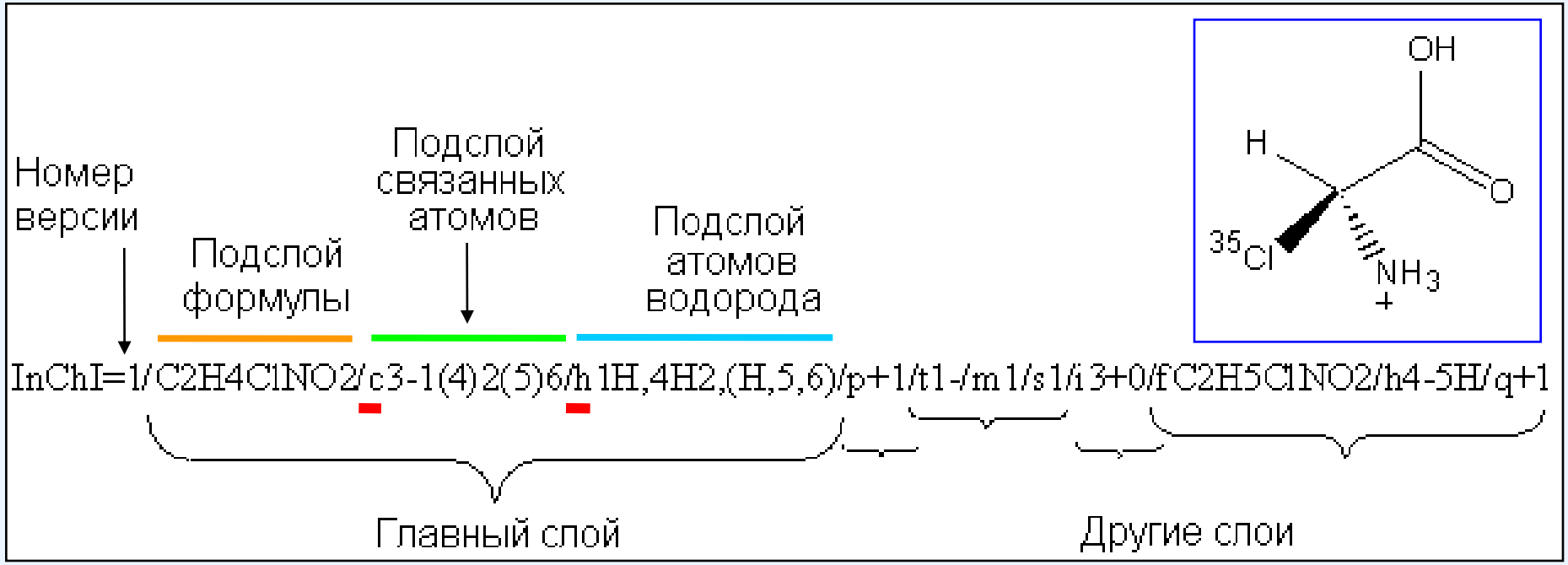
Но! Не человек, а компьютер должен генерировать код или из кода генерировать структурную формулу.

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 FAQ, 2009.

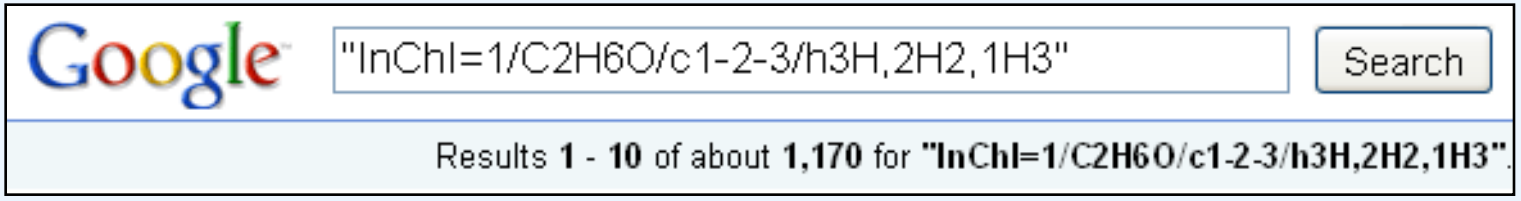
InChI

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.

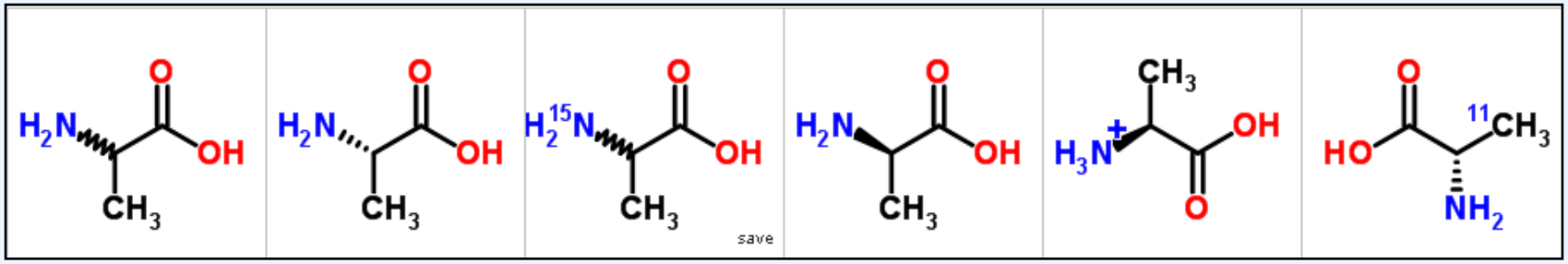
Для удобства компьютера его преобразовывают в InChIKey.

Скелет молекулы – подслой /c	Остальные слои	Дополнит. индикаторы

InChIKey=QNAYBMKLOCPYGJ-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