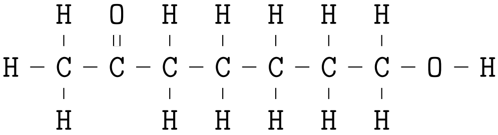
Analysing and Classifying Names of Chemical Compounds with CHEMorph

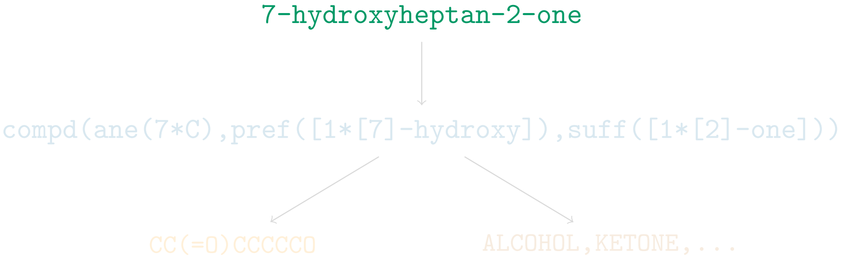
Stefanie Anstein   Gerhard Kremer

IMS, University of Stuttgart

April 11, 2006

Example Analysis





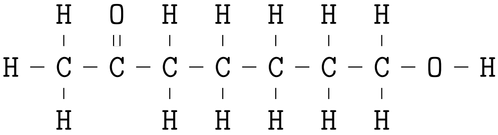
7-hydroxyheptan-2-one

compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))

CC(=O)CCCCCO

ALCOHOL,KETONE,...

Example Analysis



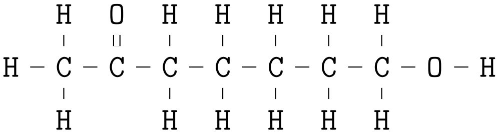
7-hydroxyheptan-2-one

compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))

CC(=O)CCCCCO

ALCOHOL,KETONE,...

Example Analysis



7-hydroxyheptan-2-one

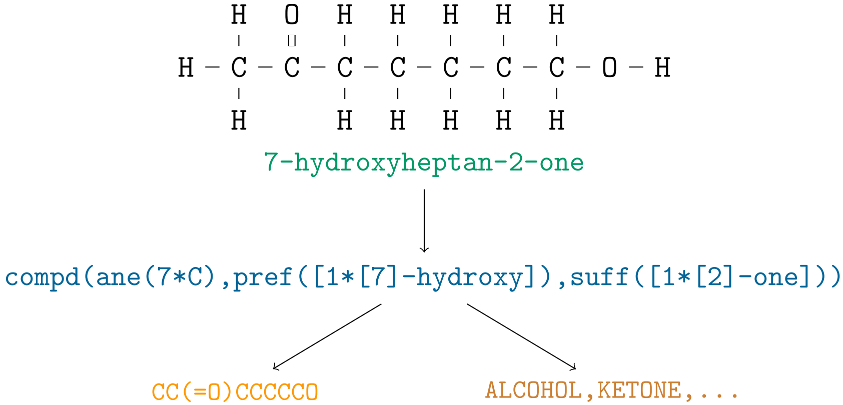
↓

compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))

CC(=O)CCCCCO

ALCOHOL,KETONE,...

Example Analysis



7-hydroxyheptan-2-one

compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))

CC(=O)CCCCCO

ALCOHOL,KETONE,...

Motivation & Background

- life sciences . . .

  and the amount of biomedical data

- terminology . . .

  and biochemical nomenclature

Motivation & Background

- life sciences . . .

  and the amount of biomedical data

- terminology . . .

  and biochemical nomenclature ▶

Challenges

- term reference ▸
- coreferences ▸

Challenges

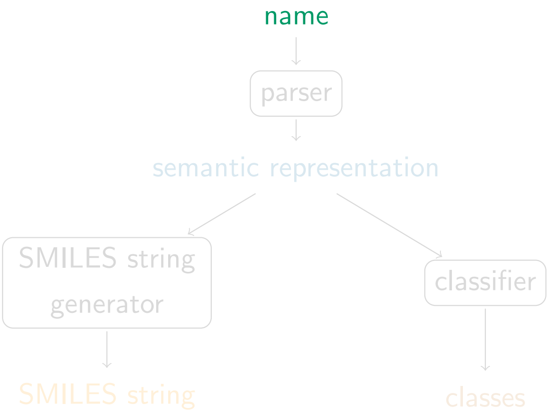
- term reference ▸
- coreferences ▸

- R-0. 1.7.3 (IUPAC nomenclature of organic compounds):

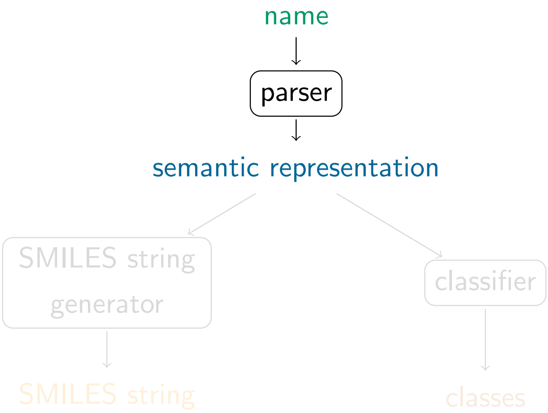
  Addition of the vowel “o”.

  For euphonic reasons, the vowel “o” is sometimes inserted between consonants.

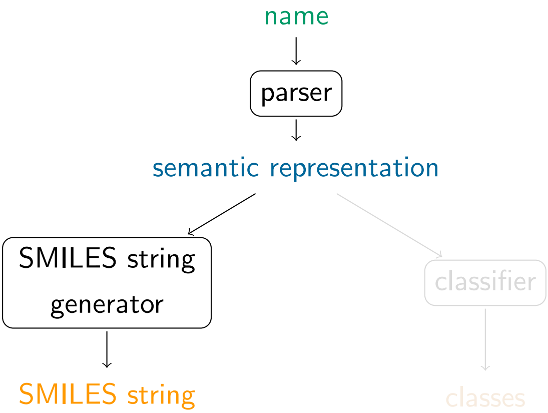
Modules Overview



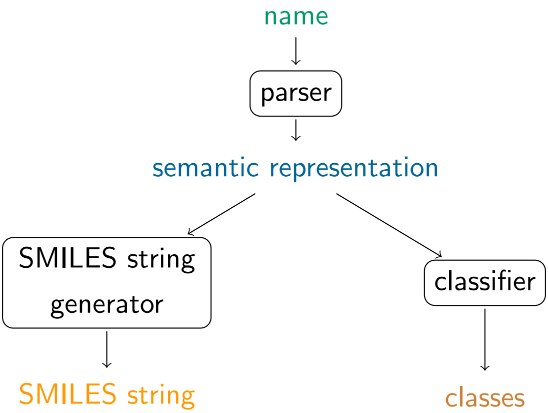
Modules Overview



Modules Overview



Modules Overview


name

parser

semantic representation

SMILES string generator

classifier

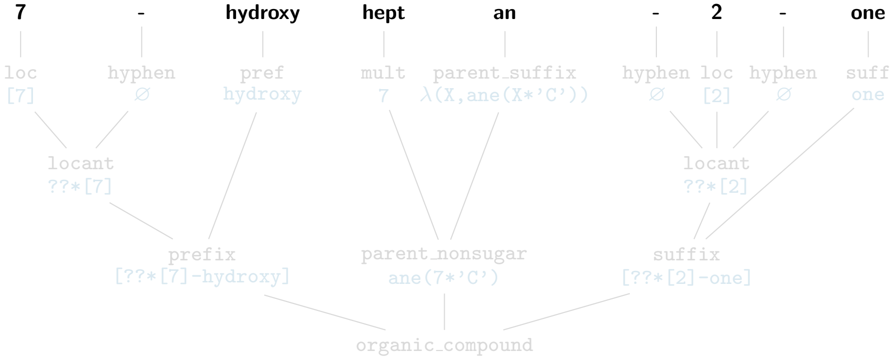
SMILES string

classes

Name Types

	fully specified	underspecified
systematic	7-hydroxyheptan-2-one	heptanone
trivial	benzene	∅
semi-systematic	benzene-1,3,5-triacetic acid	dihydrobenzene
class	∅	alcohol ▸
semi-systematic	∅	2-deoxysugar

Parser



7 **-** **hydroxy** **hept** **an** **-** **2** **-** **one**

loc [7] hyphen ∅ pref hydroxy mult 7 parent_suffix λ(X,ane(X*'C')) hyphen ∅ loc [2] hyphen ∅ suff one

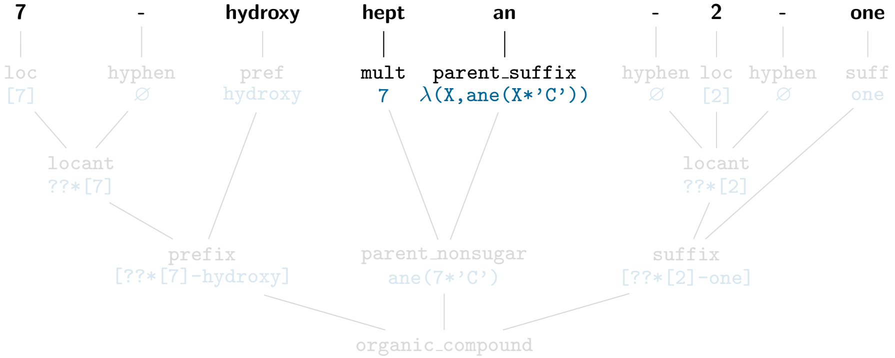
locant ??*[7] locant ??*[2]

prefix [??*[7]-hydroxy] parent_nonsugar ane(7*'C') suffix [??*[2]-one]

organic_compound

compd(ane(7*C) , pref([??*[7]-hydroxy]) ,
suff([??*[2]-one]))

Parser



7 - hydroxy hept an - 2 - one

loc [7] hyphen ∅ pref hydroxy mult 7 parent_suffix λ(X,ane(X*'C')) hyphen ∅ loc [2] hyphen ∅ suff one

locant ??*[7] locant ??*[2]

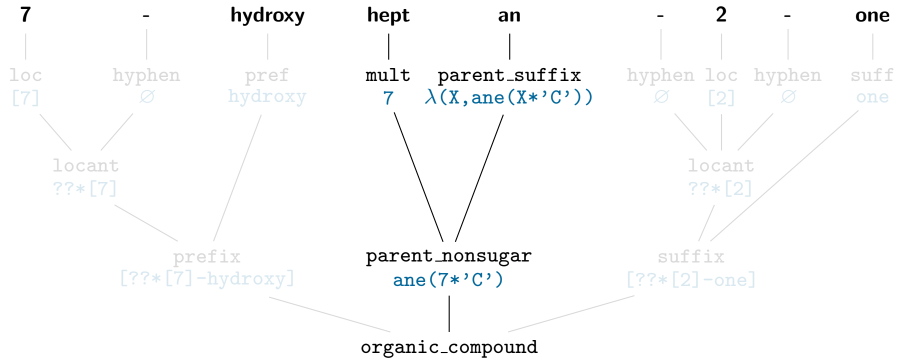
prefix [??*[7]-hydroxy] parent_nonsugar ane(7*'C') suffix [??*[2]-one]

organic_compound

compd(ane(7*C) , pref([??*[7]-hydroxy]) ,

suff([??*[2]-one]))

Parser



7 - hydroxy hept an - 2 - one

loc [7] hyphen ∅ pref hydroxy mult 7 parent_suffix λ(X,ane(X*'C')) hyphen ∅ loc [2] hyphen ∅ suff one

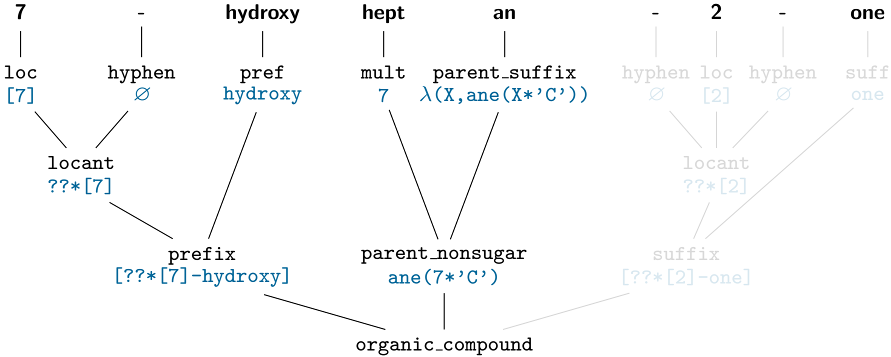
locant ??*[7] locant ??*[2]

prefix [??*[7]-hydroxy] parent_nonsugar ane(7*'C') suffix [??*[2]-one]

organic_compound

```
compd( ane(7*C) , pref( [??*[7]-hydroxy] ) ,
suff( [??*[2]-one] ) )
```

Parser



7 - hydroxy hept an - 2 - one

loc [7]
hyphen ∅
pref hydroxy
mult 7
parent_suffix λ(X,ane(X*'C'))
hyphen ∅
loc [2]
hyphen ∅
suff one

locant ??*[7]
locant ??*[2]

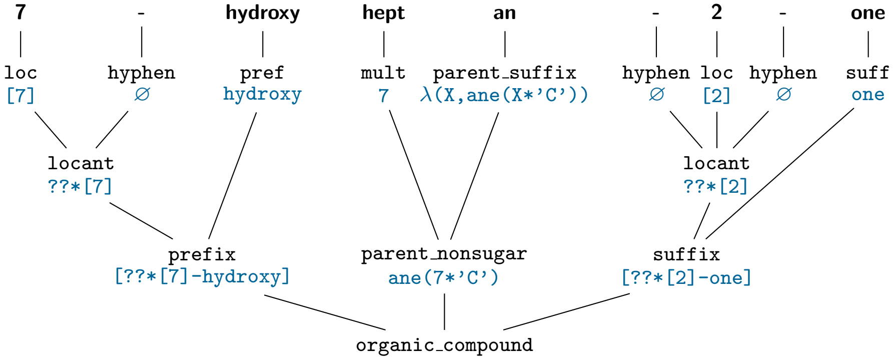
prefix [??*[7]-hydroxy]
parent_nonsugar ane(7*'C')
suffix [??*[2]-one]

organic_compound

compd(ane(7*C) , pref([??*[7]-hydroxy]) ,

suff([??*[2]-one]))

Parser



7 - hydroxy hept an - 2 - one

loc [7]

hyphen ∅

pref hydroxy

mult 7

parent_suffix λ(X,ane(X*'C'))

hyphen ∅

loc [2]

hyphen ∅

suff one

locant ??*[7]

locant ??*[2]

prefix [??*[7]-hydroxy]

parent_nonsugar ane(7*'C')

suffix [??*[2]-one]

organic_compound

compd(ane(7*C) , pref([??*[7]-hydroxy]) ,
suff([??*[2]-one]))

SMILES String Generator

- representation of single chain elements
- consistency check
- underspecification:

```
underspecified( CC(=O)CCCCC , [{1,3,4,5,6,7}-hydroxy] )
```

SMILES String Generator

- representation of single chain elements
- consistency check
- underspecification:

  `underspecified( CC(=O)CCCCC , [{1,3,4,5,6,7}-hydroxy] )`

SMILES String Generator

- representation of single chain elements
- consistency check
- underspecification:

```
underspecified( CC(=O)CCCCC , [{1,3,4,5,6,7}-hydroxy] )
```

Classifier

morpheme	class
`hydroxy- | -ol`	`ALCOHOL`
`cyclo- & -ane`	`CYCLOALKANE`

- compd(ane(7*C) , pref([1*[7]-hydroxy]) , suff([1*[2]-one]))

  → ALKANE, ALCOHOL, KETONE

- compd(ene(??*[??],ane(4*'C')) , pref([]) , suff([]))

  → ALKENE

Classifier

morpheme	class
hydroxy- | -ol	ALCOHOL
cyclo- & -ane	CYCLOALKANE

- compd(ane(7*C) , pref([1*[7]-hydroxy]) , suff([1*[2]-one]))

  → ALKANE, ALCOHOL, KETONE

- compd(ene(??*[??],ane(4*'C')) , pref([]) , suff([]))

  → ALKENE

Classifier

morpheme	class
hydroxy- | -ol	ALCOHOL
cyclo- & -ane	CYCLOALKANE

- compd(ane(7*C) , pref([1*[7]-hydroxy]) , suff([1*[2]-one]))

  → ALKANE, ALCOHOL, KETONE

- compd(ene(??*[??],ane(4*'C')) , pref([]) , suff([]))

  → ALKENE

Classifier

morpheme	class
hydroxy- | -ol	ALCOHOL
cyclo- & -ane	CYCLOALKANE

- compd(ane(7*C) , pref([1*[7]-hydroxy]) , suff([1*[2]-one]))

  → ALKANE, ALCOHOL, KETONE

- compd(ene(??*[??],ane(4*'C')) , pref([]) , suff([]))

  → ALKENE

Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution ▸
- database curation and ontology acquisition ▸

Results & Applications

- SMILES string and classification
- underspecification

- term reference
- coreference resolution ▸

- database curation and ontology acquisition ▸

Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution
- database curation and ontology acquisition

Results & Applications

- SMILES string and classification
- underspecification

- term reference
- coreference resolution ▶

- database curation and ontology acquisition ▶

Results & Applications

- SMILES string and classification
- underspecification

- term reference
- coreference resolution ▸

- database curation and ontology acquisition ▸

Conclusion & Outlook

- feasible, extendable and transferable approach

- extend grammar and lexicon
- elaborate SMILES and classification

- sophisticated linguistic analysis → database curation
- term identification → text processing applications

Conclusion & Outlook

- feasible, extendable and transferable approach

- extend grammar and lexicon
- elaborate SMILES and classification

- sophisticated linguistic analysis → database curation
- term identification → text processing applications

Conclusion & Outlook

- feasible, extendable and transferable approach

- extend grammar and lexicon
- elaborate SMILES and classification

- sophisticated linguistic analysis → database curation
- term identification → text processing applications

Acknowledgements

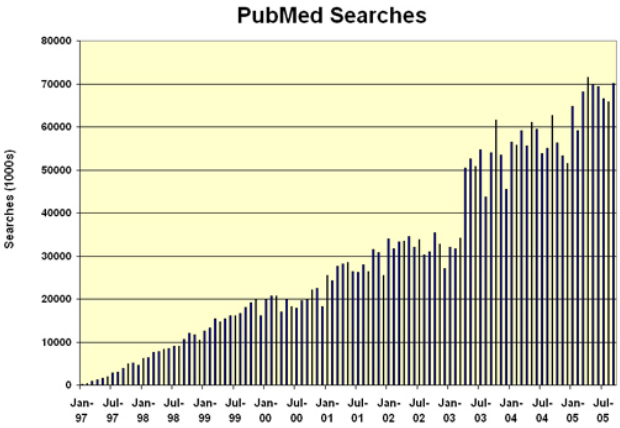
Stefanie Anstein

Uwe Reyle

Jasmin Šarić

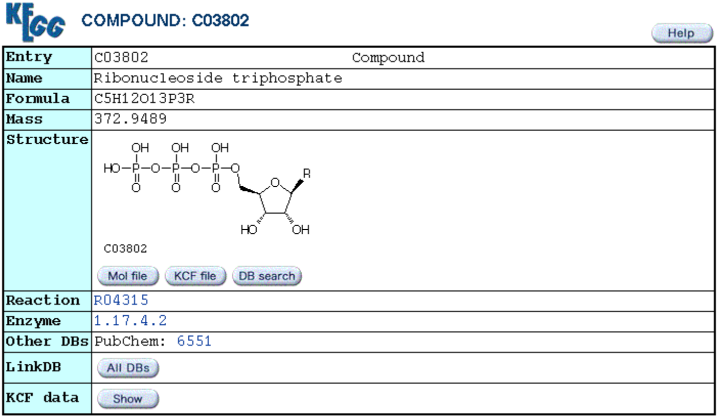
EML Research gGmbH

Schönen Dank.


PubMed Searches
Searches (1000s)
80000
70000
60000
50000
40000
30000
20000
10000
0
Jan-97
Jul-97
Jan-98
Jul-98
Jan-99
Jul-99
Jan-00
Jul-00
Jan-01
Jul-01
Jan-02
Jul-02
Jan-03
Jul-03
Jan-04
Jul-04
Jan-05
Jul-05

IUPAC Nomenclatures

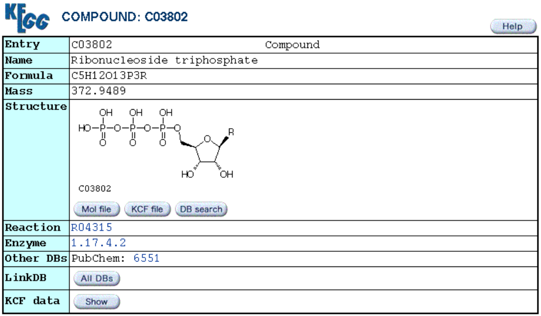
- Amino Acids and Peptides
- Biochemical thermodynamics
- Branched nucleic acids
- **Carbohydrates**
- Carotenoids
- Corrinoids (vitamin B12)
- Cyclitols
- Electron transport proteins
- Enzyme kinetics
- Enzyme nomenclature
  - EC 1 Oxidoreductases
  - EC 2 Transferases
  - EC 3 Hydrolases
  - EC 4 Lyases
  - EC 5 Isomerases
  - EC 6 Ligases
- Folic acid
- Glycolipids
- Glycoproteins
- myo-Inositol numbering
- Lignan Nomenclature
- Lipid Nomenclature
- Multienzymes
- Multiple forms of enzymes
- Nucleic acid constituents
- Nucleic acid sequence
- **Organic Chemistry**
- Peptide hormones
- Phosphorus containing compds
- Polymerized amino acids
- Polypeptide conformation
- Polynucleotide conformation
- Polysaccharide conformation
- Prenol nomenclature
- Pyridoxal (vitamin B6)
- Quinones w. an Isoprenoid Chain
- Retinoids
- Steroids
- Tetrapyrroles
- Tocopherols (vitamin E)
- Translation Factors
- Vitamin D



COMPOUND: C03802

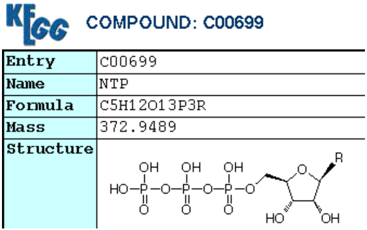
Entry	C03802 Compound
Name	Ribonucleoside triphosphate
Formula	C5H12O13P3R
Mass	372.9489
Structure	C03802
Reaction	R04315
Enzyme	1.17.4.2
Other DBs	PubChem: 6551
LinkDB	All DBs
KCF data	Show

KEGG: Kyoto Encyclopedia of Genes and Genomes



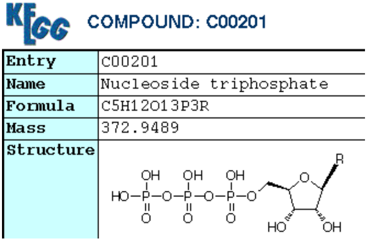
COMPOUND: C03802

Entry	C03802	Compound
Name	Ribonucleoside triphosphate	
Formula	C5H12O13P3R	
Mass	372.9489	
Structure		
Reaction	R04315	
Enzyme	1.17.4.2	
Other DBs	PubChem: 6551	
LinkDB		
KCF data		



COMPOUND: C00699

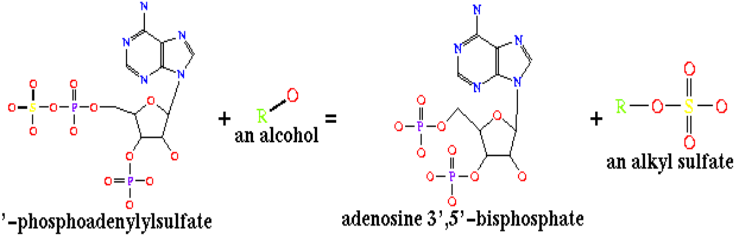
Entry	C00699
Name	NTP
Formula	C5H12O13P3R
Mass	372.9489
Structure	



COMPOUND: C00201

Entry	C00201
Name	Nucleoside triphosphate
Formula	C5H12O13P3R
Mass	372.9489
Structure	

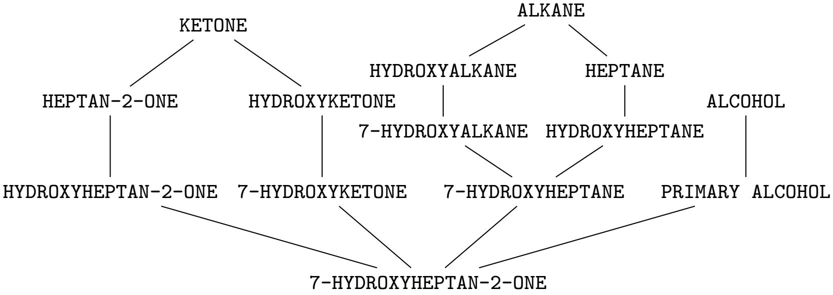
Reaction: *3′-phosphoadenylyl sulfate + an alcohol = adenosine 3′,5′-bisphosphate + an alkyl sulfate.*



Molecule diagrams generated from .mol files obtained from the KEGG ftp site.

Other name(s): *Hydroxysteroid sulfotransferase.*

Comments: *Primary and secondary alcohols, including aliphatic alcohols, ascorbate, chloramphenicol, ephedrine and hydroxysteroids, but not phenolic steroids, can act as acceptors (cf. Ec 2.8.2.15).*



ALKANE

KETONE

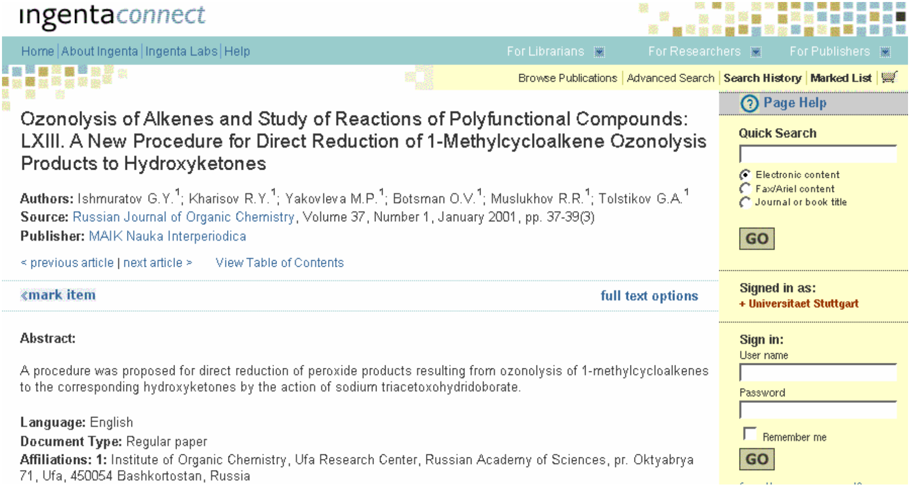
HYDROXYALKANE HEPTANE

HEPTAN-2-ONE HYDROXYKETONE ALCOHOL

7-HYDROXYALKANE HYDROXYHEPTANE

HYDROXYHEPTAN-2-ONE 7-HYDROXYKETONE 7-HYDROXYHEPTANE PRIMARY ALCOHOL

7-HYDROXYHEPTAN-2-ONE



Ozonolysis of Alkenes and Study of Reactions of Polyfunctional Compounds: LXIII. A New Procedure for Direct Reduction of 1-Methylcycloalkene Ozonolysis Products to Hydroxyketones

Authors: Ishmuratov G.Y.[1]; Kharisov R.Y.[1]; Yakovleva M.P.[1]; Botsman O.V.[1]; Muslukhov R.R.[1]; Tolstikov G.A.[1]

Source: Russian Journal of Organic Chemistry, Volume 37, Number 1, January 2001, pp. 37-39(3)

Publisher: MAIK Nauka Interperiodica

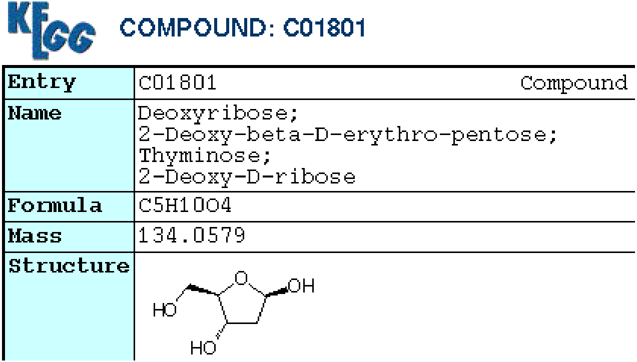
Abstract:

A procedure was proposed for direct reduction of peroxide products resulting from ozonolysis of 1-methylcycloalkenes to the corresponding hydroxyketones by the action of sodium triacetoxohydridoborate.

Language: English

Document Type: Regular paper

Affiliations: 1: Institute of Organic Chemistry, Ufa Research Center, Russian Academy of Sciences, pr. Oktyabrya 71, Ufa, 450054 Bashkortostan, Russia



COMPOUND: C01801

Entry	C01801	Compound

Name	Deoxyribose;
	2-Deoxy-beta-D-erythro-pentose;
	Thyminose;
	2-Deoxy-D-ribose	

Formula	C5H10O4	
Mass	134.0579	
Structure		