

Chemical Structure Representation

Randy Julian

Lilly Research Laboratories

Representing Structures

◆ Coordinate-based: Connection Tables

- Give x,y,z positions of atom and type
- Give formal charges
- Give bond information between each atom
- Representation by one or more tables

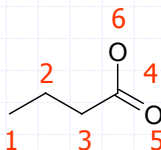
◆ Graph-based: Line notations

- Give atom type
- Describe connection and branching between atoms
- Representation by a string of characters

Example Formats

- ◆ MDL "MOL" File (Connection table)
 - Developed by MDL, Inc.
 - Format published and used by many applications
 - Two main tables: atoms, bonds
- ◆ Simplified Molecular Input Line Entry System
 - SMILES Developed by Daylight, Inc.
 - Format published in JChICS 1987
 - Single string of printable characters

MOL Files



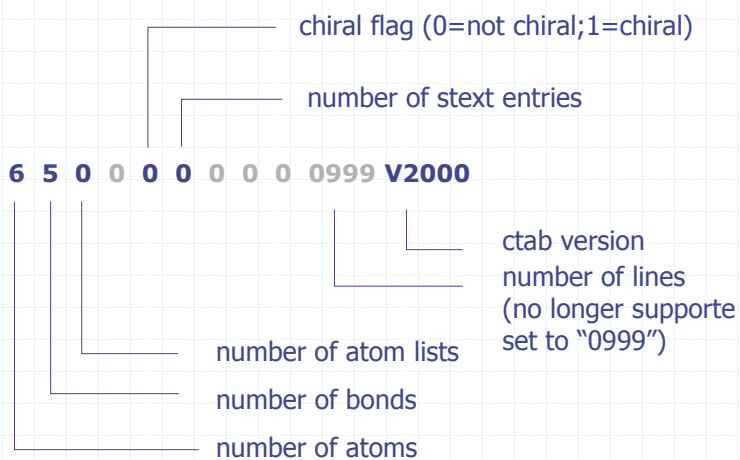
6 5 0 0 0 0 0 0 0 0999 V2000											counts line									
-3.0000	-1.2167	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-2.2875	-0.8042	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-1.5750	-1.2167	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.8625	-0.8042	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.1475	-1.2158	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.8667	0.0208	0.0000	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0														
3	4	1	0	0	0	0														
4	5	2	0	0	0	0														
2	3	1	0	0	0	0														
4	6	1	0	0	0	0														

M END

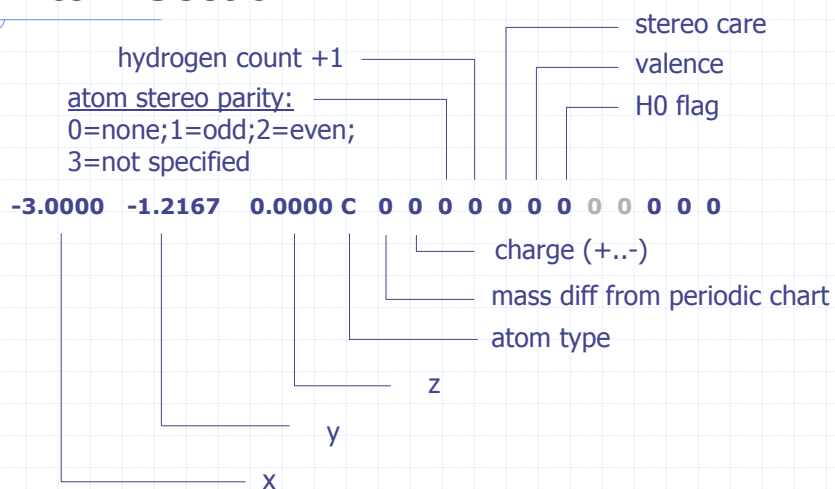
bond section

atom section

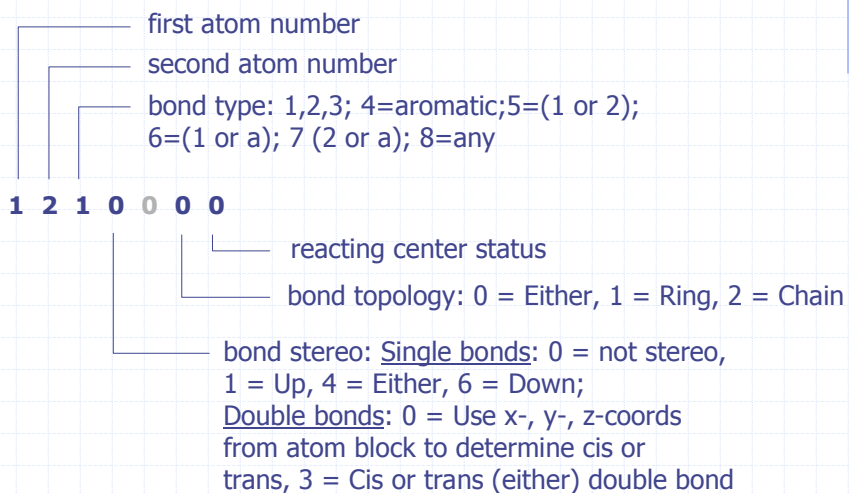
MOL file format: counts line



Atom Section



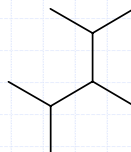
Bond Section



Branched structures

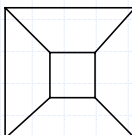
```

9 8 0 0 0 0 0 0 0 0 0 0999 V2000
  5.5583 -2.5292 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  6.2728 -2.1167 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  4.8439 -2.1167 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  5.5583 -3.3542 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  6.9833 -2.5250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  6.2720 -1.2917 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  6.9860 -0.8785 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  5.5571 -0.8799 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  6.9792 -3.3500 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  2 5 1 0 0 0 0
  1 3 1 0 0 0 0
  2 6 1 0 0 0 0
  1 2 1 0 0 0 0
  6 7 1 0 0 0 0
  1 4 1 0 0 0 0
  6 8 1 0 0 0 0
  5 9 1 0 0 0 0
M END
  
```



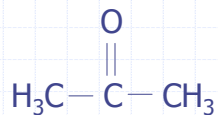
Cubane

```
8 12 0 0 0 0 0 0 0 0 0999 V2000
0.9754 -1.6212 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
0.9629 -4.0087 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
3.3545 -4.0212 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
3.3670 -1.6337 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
1.8000 -2.4458 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
1.8000 -3.2708 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
2.6250 -3.2708 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
2.6250 -2.4458 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
4 1 1 0 0 0 0
5 6 1 0 0 0 0
1 2 1 0 0 0 0
2 3 1 0 0 0 0
3 4 1 0 0 0 0
6 7 1 0 0 0 0
7 8 1 0 0 0 0
8 5 1 0 0 0 0
1 5 1 0 0 0 0
8 4 1 0 0 0 0
7 3 1 0 0 0 0
6 2 1 0 0 0 0
M END
```

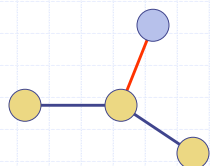


Moving to a structure proxy

```
-ISIS- 04060323172D
4 3 0 0 0 0 0 0 0 0999 V2000
-0.3458 -2.9667 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.3667 -2.5500 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.3621 -1.7250 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.0834 -2.9585 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 3 2 0 0 0 0
1 2 1 0 0 0 0
2 4 1 0 0 0 0
M END
```



Graph approach:



2-D non-directed graph:

N nodes

E edges

SMILES

- ◆ The graph of a chemical structure was to be uniquely described, including but not limiting it to the molecular graph comprising nodes (atoms) and edges (bonds).
- ◆ A user-friendly structure specification was to be provided, so that all input rules could be learned quickly and naturally.
- ◆ A machine-friendly and machine-independent system was to be designed for interpretation and generation of a unique notation.
- ◆ Unlike other chemical notation systems, brevity of notation and economy of alphabet were not primary objectives.

Atoms

C	methane (CH ₄)
N	ammonia (NH ₃)
O	water (H ₂ O)
P	phosphine (PH ₃)
S	hydrogen sulfide (H ₂ S)
Cl	hydrogen chloride (HCl)

Elements not in the organic subset must be described in brackets, e.g.

[Au] elemental gold

Attached Hydrogen & Charges

Attached hydrogens and formal charges are always specified inside brackets. The number of attached hydrogens is shown by the symbol H followed by an optional digit. Similarly, a formal charge is shown by one of the symbols + or -, followed by an optional digit. If unspecified, the number of attached hydrogens and charges is assumed to be zero for an atom inside the bracket. Examples are

[H+]	proton
[OH-]	hydroxyl anion
[OH3+]	hydronium cation
[Fe+2]	iron(II) cation
[NH4+]	ammonium cation

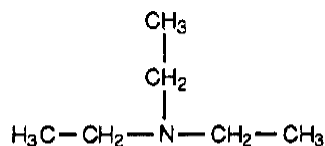
The SMILES program also recognizes constructions of the form [Fe+++] as being synonymous with the form [Fe+3].

Bonds

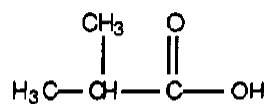
-	single (or implied)
=	double
#	triple
:	aromatic
CC	ethane (CH ₃ CH ₃)
C=C	ethylene (CH ₂ =CH ₂)
COC	dimethyl ether (CH ₃ OCH ₃)
CCO	ethanol (CH ₃ CH ₂ OH)
C=O	formaldehyde (CH ₂ O)
O=C=O	carbon dioxide (CO ₂)
O=CO	formic acid (HCOOH)
C#N	hydrogen cyanide (HCN)
[H][H]	molecular hydrogen (H ₂)

Branches

Branches are indicated by parenthesis: ()

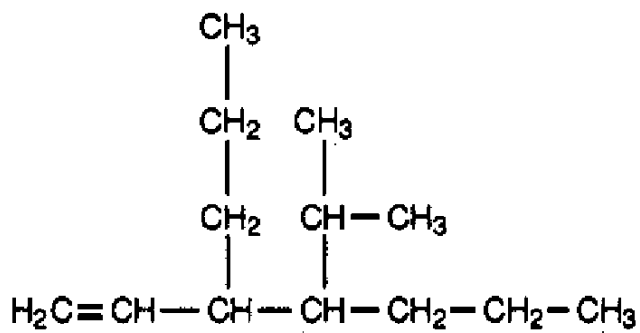


CCN(CC)CC
Triethylamine



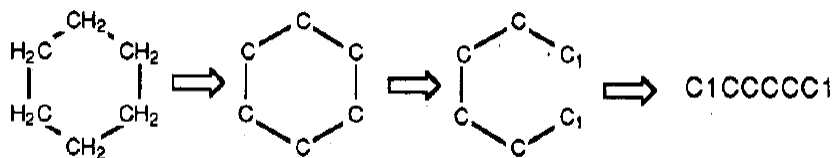
CC(C)C(=O)O
Isobutyric acid

Nested Branches

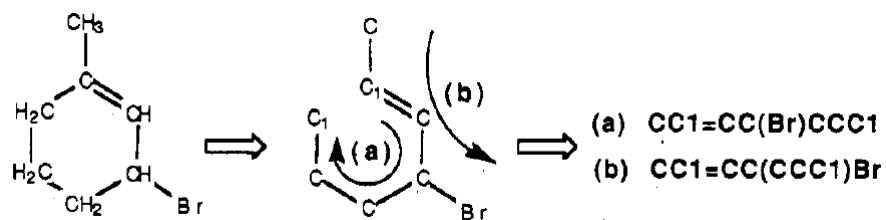


Cyclic Structures

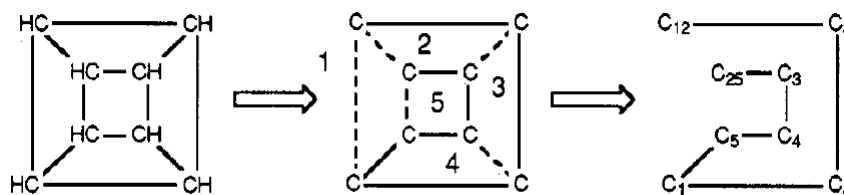
Cyclic structures are represented by breaking one single (or aromatic) bond in each ring. The bonds are numbered in any order, designating ring-opening (or ring-closure) bonds by a digit immediately following the atomic symbol at each ring closure.



Six chemists, seven SMILES...



Multiple Ring Closures



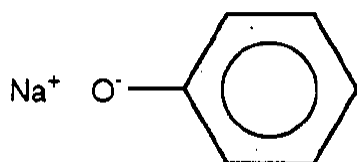
Generation of SMILES for cubane: C12C3C4C1C5C4C3C25

Higher numbered connections can be used by using % before the number:

C%12CCCCC%12 (cyclohexane)

Disconnected Structures

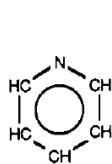
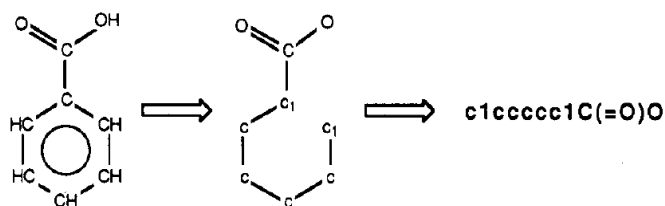
Uses "." to indicate disconnection



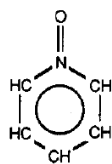
[Na+].[O-]c1ccccc1
or
c1cc([O-].[Na+])ccc1

Aromaticity

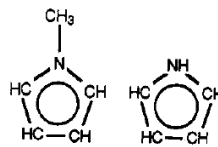
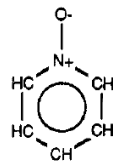
Use lower case letters for aromatic connections



n1ccccc1
Pyridine



O=n1ccccc1 **[O-][N+]c1ccccc1**
Pyridine-N-oxide



Cn1ccccc1 **[nH]1ccccc1**
methyl and 1H-pyrrole

Unique SMILES

- ◆ Problem comparing strings to compare structures
- ◆ Need a unique SMILES string for a given structure
- ◆ Fundamentally a classification problem...
- ◆ Use node invariants to establish node order

Atom Invariants

1. number of connections
2. number of non-hydrogen bonds
3. atomic number
4. sign of charge
5. absolute charge
6. number of attached hydrogens

Invariants for Pentane: CCCCC

<u>Atom type</u>	<u>individual invariant</u>	<u>combined invariant</u>
methyl carbon	1, 1, 6, 0, 0, 3	10106003
methylene carbon	2, 2, 6, 0, 0, 2	20206002

Rank Equivalence

Invariants for Pentane: CCCCC

<u>Atom type</u>	<u>individual invariant</u>	<u>combined invariant</u>
methyl carbon	1, 1, 6, 0, 0, 3	10106003
methylene carbon	2, 2, 6, 0, 0, 2	20206002

10106003 - 20206002 - 20206002 - 20206002 - 10106003

1-2-2-2-1

Simple Extended Connectivity

Sum over nodes in one direction:

1-2-2-2-1 *becomes*

2-3-4-3-2 *which can be replaced by*

1-2-3-2-1

Repeat until the set no longer changes

Any function can be used however replacing rank with 'Corresponding Prime' and then with product of neighbor, ensures unique numbering via the prime factorization theorem.

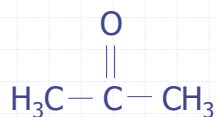
"CANON" Algorithm Summary

1. Set atomic vector to initial invariants. Go to step 3.
2. Set vector to product of primes corresponding to neighbors' ranks
3. Sort vector, maintaining stability over previous ranks.
4. Rank atomic vector.
5. If not invariant partitioning, go to step 2.
6. On first pass, save partitioning as symmetry classes.
7. If highest rank is smaller than number of nodes, break ties, go to step 2.
8. ... else done.

Generating Unique String: "GENES"

- ◆ With atoms properly ranked...
- ◆ Do a depth-first-search on the structure as if it were a tree
- ◆ Need to decide where to 'start'
 - Lowest numbered atom - this will be a terminal atom in the structure
- ◆ Need a branch decision
 - Branch toward lowest ranked atom
 - In cyclic structures, branch to a double or triple bond if it exists

Acetone Example



C - C (=O)- C CC(=O)C
10106003-30406000-(10208000)-10106003

1-4-(3)-2

C - C (C)- =O
10106003-30406000-(10106003)- 10208000

1-4-(2)-3

CC(C)=O

Tools

- ◆ MDL ISIS/Draw (Free for Academic Use)
 - Draws structures, exports MDL molfiles
- ◆ BABEL (various open source projects)
 - C/C++
 - Imports/exports large number of formats
 - Some versions do not generate unique SMILES
- ◆ Perl
 - Can be used to compute information, rearrange, etc.

Position Sensitive Text Files

- ◆ Information context is given by position in file or line
- ◆ Validity of file cannot be tested prior to reading
- ◆ This is a problem when automating things (like the www).
- ◆ What might help: context from grammar

XML - Extensible Markup Language

- ◆ Transition from position sensitive to grammar sensitive files
- ◆ Use 'tags' to define meaning
- ◆ Use grammatical tests to determine 'well-formedness' (Readability)
- ◆ Use templates to test validity

Example: Chemical Markup Language

- ◆ CML is an XML-based Connection Table

```
<document>
  <!-- CML document - acetone - - -->
  <!-- file converted from: MDL .mol -->
  <cml title="acetone" id="cml_acetone_" xmlns="x-schema:cml_schema_ie_02.xml">
    <molecule title="acetone" id="mol_acetone_">
      <atomArray>
        <atom id="acetone__a_1">
          <float builtin="x3" units="A">-0.3458</float>
          <float builtin="y3" units="A">-2.9667</float>
          <float builtin="z3" units="A">0.0000</float>
          <string builtin="elementType">C</string>
        </atom>
        <atom id="acetone__a_2">
          <float builtin="x3" units="A">0.3667</float>
          <float builtin="y3" units="A">-2.5500</float>
          <float builtin="z3" units="A">0.0000</float>
          <string builtin="elementType">C</string>
        </atom>
        <atom id="acetone__a_3">
          <float builtin="x3" units="A">0.3621</float>
          <float builtin="y3" units="A">-1.7250</float>
          <float builtin="z3" units="A">0.0000</float>
          <string builtin="elementType">O</string>
        </atom>
        <atom id="acetone__a_4">
          <float builtin="x3" units="A">1.0834</float>
          <float builtin="y3" units="A">-2.9585</float>
          <float builtin="z3" units="A">0.0000</float>
          <string builtin="elementType">C</string>
        </atom>
      </atomArray>
    </molecule>
  </cml>
</document>
```

```

<bondArray>
  <bond id="acetone__b_1">
    <string builtin="atomRef">acetone__a_2</string>
    <string builtin="atomRef">acetone__a_3</string>
    <string builtin="order" convention="MDL">2</string>
  </bond>
  <bond id="acetone__b_2">
    <string builtin="atomRef">acetone__a_1</string>
    <string builtin="atomRef">acetone__a_2</string>
    <string builtin="order" convention="MDL">1</string>
  </bond>
  <bond id="acetone__b_3">
    <string builtin="atomRef">acetone__a_2</string>
    <string builtin="atomRef">acetone__a_4</string>
    <string builtin="order" convention="MDL">1</string>
  </bond>
</bondArray>
</molecule>
</cml>
</document>

```

References

- ◆ CTfile Formats, MDL Inc., *August 2002* (www.mdl.com)
- ◆ SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, DAVID WEININGER, *J. Chem. Inf. Comput. Sci.* 1988, 28, 31-36.
- ◆ SMILES. 2. Algorithm for Generation of Unique SMILES Notation, DAVID WEININGER, ARTHUR WEININGER, and JOSEPH L. WEININGER, *J. Chem. Inf. Comput. Sci.* 1989, 29, 97-101.
- ◆ SMILES. 3. Depict. Graphical Depiction of Chemical Structures, DAVID WEININGER, *J. Chem. Inf. Comput. Sci.* 1990, 30, 231-243.
- ◆ Daylight Inc.: www.daylight.com
- ◆ BABEL: <http://smog.com/chem/babel/> (among others)
- ◆ CML: <http://www.xml-cml.org/>