Application of Computer in Chemistry SSC 3533

COMPUTER REPRESENTATION OF CHEMICAL STRUCTURE

Prof. Mohamed Noor Hasan
Dr. Hasmerya Maarof
Department of Chemistry

Outline

- Representation of Chemical Structures
- Linear Notation
- SMILES
- Connection Table
- Exercises

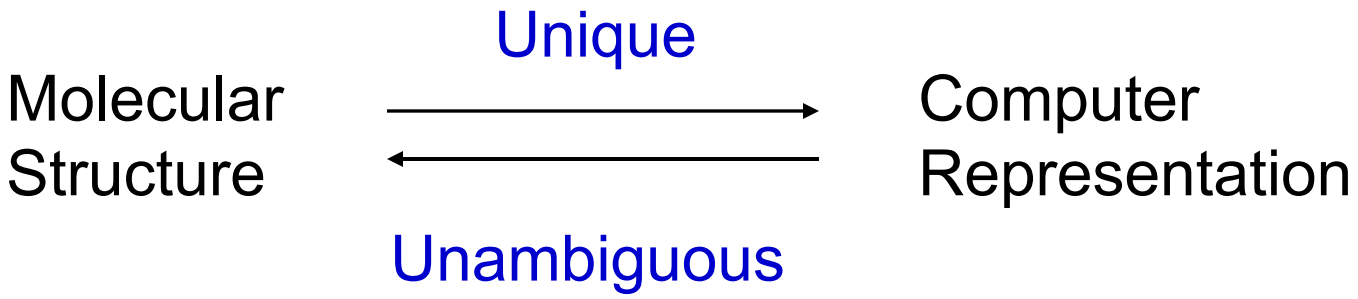
Introduction

- Applications of computer in chemistry is not limited to numerical calculations - can also be used to assist chemists in handling molecular structures
- There is a need for computer representation of chemical structure
- Store structures in a database
- To search structures with certain functional groups (substructure)
- To search compounds with similar structures
- To estimate properties/activities from structures

Representation of Chemical Structures

- Chemists have long been using molecular structures to represent a chemical compound
- Although the structure does not show the actual shape of the molecule, the information it conveys is adequate to describe the compound
- Will discuss various methods that have been used to represent and handle molecular structure information using computer

Computer Representation



- Unique means only one computer representation can be derived for a structure
- Unambiguous means only one structure can be produced from one computer representation

Methods for Computer Representation

Some of the commonly used methods for representing chemical structures:

- Linear Notations
  - Wiswesser Line Notation (WLN)
  - SMILES Coding*
  - ROSDAL
  - Sybyl Line Notation (SLN)
- Connection Table*
- Fragment Coding

Wiswesser Line Notation

- First introduced by W. J. Wiswesser in the early ‘50s
- A structural formula is translated into symbols which then arranged according to certain rules.
- The resulted notation is compact and simple because chemical bonds between atoms are already included in the symbols.
- Large structures such as aromatic rings can be represented by one symbol only

WLN Symbols

- Originally the number of symbols were only 40
- Consist of all capital letters (ABC .... XYZ), numbers (012 .... 89) and some special characters “&”, “-”, “/”, “ ”
- Because of widespread use, number of symbols has been increased to include lower case letters (abc .... xyz) and other special characters
- WLN differentiate the environments of an atom. An element might be represented by different symbols, depending on elements surrounding it and how they are bonded

Disadvantages of WLN

- Although WLN is compact, simple and space saving, there are a number of weaknesses that cause the system no longer widely used.
- First: the rules are a bit complicated because many symbols have to memorized
- Second: the rule itself is changing from time to time
- Therefore its usage now is limited only to those industries that have used this system in their database.

SMILES Notation

- Simplified Molecular Input Line Entry System (SMILES)
- Widely used and very efficient
- Consists of symbols for atoms, bonds and branching; and a set of intuitive rules
- Hydrogen is not shown
- Reference: D. Weininger, *J.Chem. Inf. Comput. Sci*. 1988, 28, 31-36.

Atom

Atoms are represented by their own symbols

C	Carbon	F	Fluorine
O	Oxygen	Cl	Chlorine
N	Nitrogen	Br	Bromine
S	Sulfur	I	Iodine
P	Phosphorus	He	Helium

Bond

Single❖	-
Double	=
Triple	#
Aromatic⌘	:

❖ Usually, single bond is not shown. If no bond is indicated between two atoms, the default is single

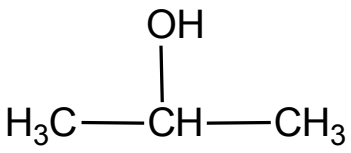
⌘ Aromatic compounds are identified by using lower case letters for the atomic symbols. There is no need to show the bonds explicitly.

Examples

Methane	CH_4	C
Propane	CH_3-CH_2-CH_3	CCCC
Chloroethane	CH_3-CH_2-Cl	CCCl
Ethene	CH_2=CH_2	C=C
Bromoethene	CHBr=CH_2	BrC=C
Acethelyne	CH≡CH	C#C
Acetonitrile	CH_3C≡N	CC#N

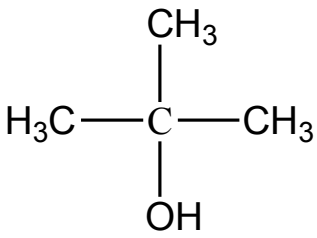
Branches

- Branches are identified by enclosing them in parentheses and they can be stacked or nested.



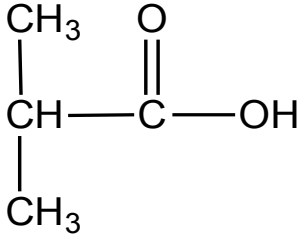
2-propanol

CC(O)C



Tert-butanol

CC(C)(O)C

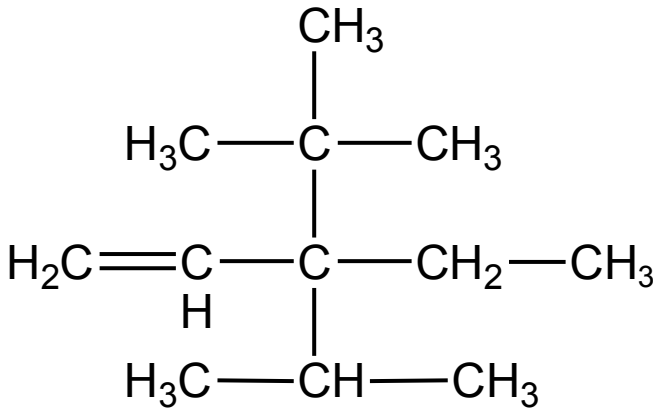


Isobutyric acid

CC(C)C(=O)O

Branches

- Branches can be nested



3-isopropyl-3tert-butyl-1-pentena

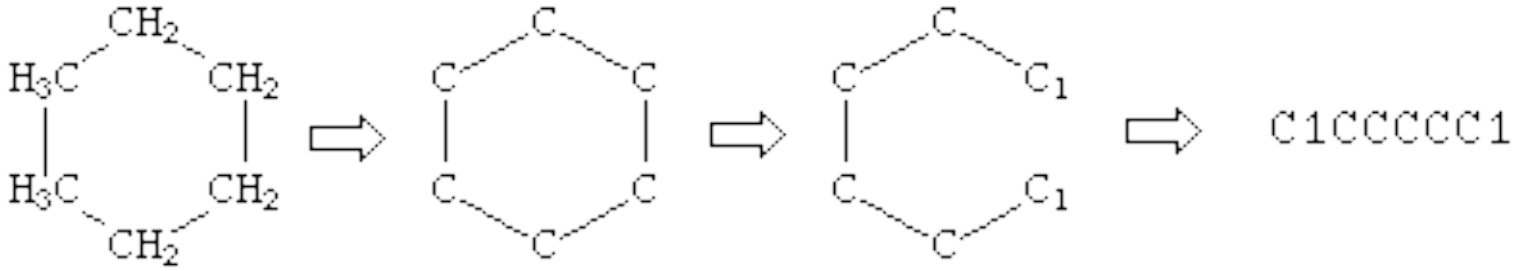
C=CC(C(C)C)(C(C)(C)C)CC

Some Conventions

- If possible, avoid two consecutive left parentheses
- Strive for the fewest number of possible branches
- A branch cannot begin a SMILES notation
- A branch cannot immediately follow a double- or triple-bond symbol

Cyclic Structures

- Rings are represented by breaking one 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

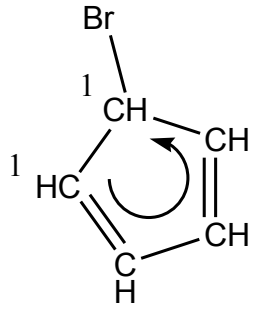


Aromatic Compounds

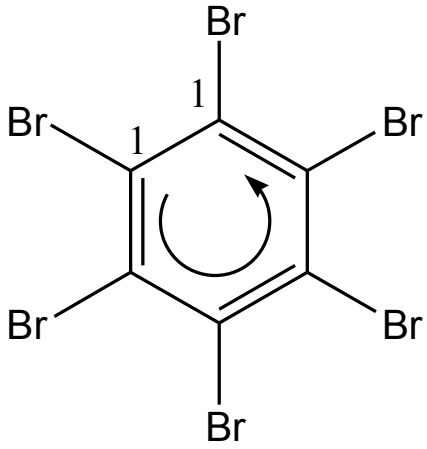
- Atoms in an aromatic ring are indicated by lower case letters – aliphatic atoms are given upper case letters.
- For ring closure, use matching digit
- Example:

c1ccccc1	Benzene
C1CCCCC1	Cyclohexane

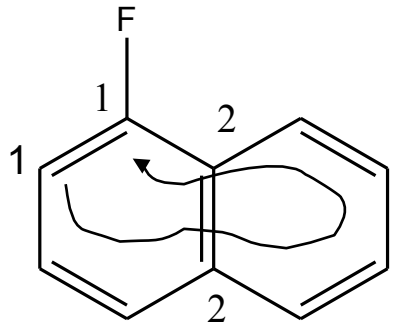
More Examples of Cyclic Structures



C1=CC=CC1Br



Brc1c(Br)c(Br)c(Br)c(Br)c1Br



c1ccc2ccccc2c1F

Charges

- Square brackets are used to indicate hydrogen and charges
- Number of Hydrogen and charges indicated with a number (digit).

Example Charges

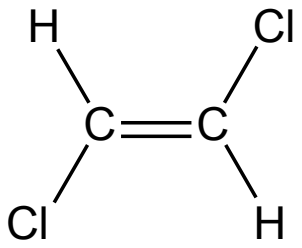
$[H^+]$	Proton
[OH-]	Hydroxyl ion
[OH3+]	Hydronium ion
[Al+++]	Aluminum (III)
[NH4+]	Ammonium ion

Metals

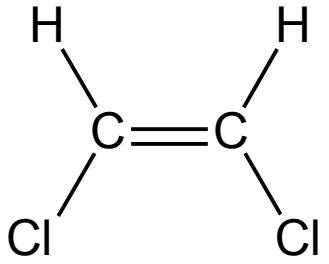
Aluminum	[Al]	Arsenic	[As]
Gold	[Au]	Berrylium	[Be]
Bismuth	[Bi]	Cadmium	[Cd]
Calcium	[Ca]	Iron	[Fe]
Mercury	[Hg]	Potassium	[K]
Sodium	[Na]	Nickel	[Ni]
Platinum	[Pt]	Tin	[Sn]
Zinc	[Zn]	Zirconium	[Zr]

Isomer

- Isomer around double bond is indicated by forward and back slash: / \
- Examples:



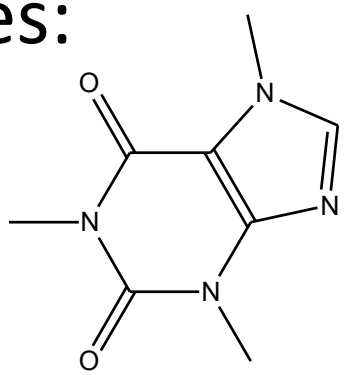
trans-1,2-dichloroethene
Cl/C=C/Cl



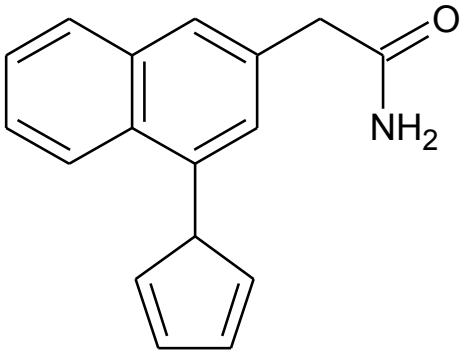
cis-1,2-dichloroethene
Cl/C=C\Cl

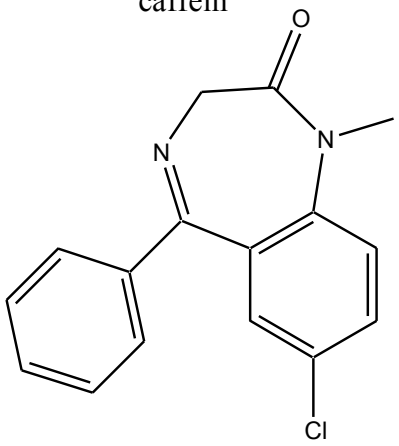
Exercise

- Write down smiles code for the following structures:

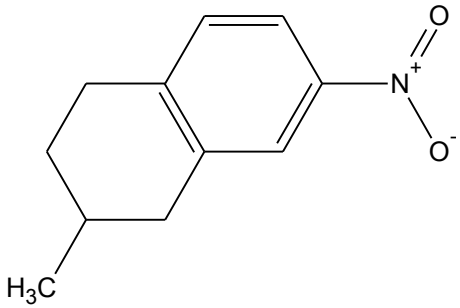


caffein





valium



Exercise

Convert the following SMILES codes into structure

CC(C)(N)Cc1ccccc1

CN1C(=O)CN=C(c2ccccc2)c3cc(Cl)ccc13

CN1CCC[C@H]1c2cccnc2

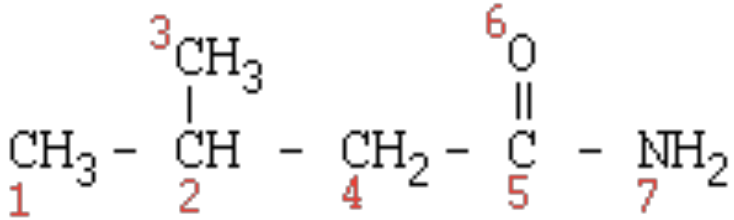
Cn1cnc2n(C)c(=O)n(C)c(=O)c12

*Refer www.daylight.com website for more exercises

Connection Table

- In a connection table, all atoms (except hydrogen), bonds, and how they are connected to each other are shown explicitly.
- Although it occupies more space than linear notation, the rules for creating a connection table are much simpler.
- Connection tables are recognised by most structure drawing software and databases of chemical structures.
- We can a a connection table in *matrix* or *table* formats.

Connection Table (Matrix format)



	1	2	3	4	5	6	7
1	C	1					
2	1	C	1	1			
3		1	C				
4		1		C	1		
5				1	C	2	1
6					2	O	
7					1		N

Connection Table (Table format)

- A connection table can also be presented in table form
- The first two columns on the left indicate the atom numbers and types
- In the pairs of columns on the right are shown to which atom they are connected and the type of bonds involved
- The table however can be simplified by showing the connection to lower numbered atoms. The resulted table is called compact connection table.

Connection Table (Table format)

Atom No	Atom Type	Connection					
		Atom No	Bond type	Atom No	Bond type	Atom No	Bond type
1	C	2	1	-	-	-	-
2	C	1	1	3	1	4	1
3	C	2	1	-	-	-	-
4	C	2	1	5	1	-	-
5	C	4	1	6	2	7	1
6	O	5	2	-	-	-	-
7	N	5	1	-	-	-	-

Compact Connection Table

No. Atom	Atom type	Connection	
		Atom no.	Bond type
1	C	-	-
2	C	1	1
3	C	2	1
4	C	2	1
5	C	4	1
6	O	5	2
7	N	5	1

Morgan Algorithm

- Produce a unique numbering system based on graph theory
- Atom is represented as a point and a bond (regardless of type) is represented as an edge.
- Number of atoms bonded to an atom is called the connectivity for that atom and this concept is used to show the degree of branching in a graph.
- Larger connectivity value means greater degree of branching in the structure and an atom that has the largest value is considered the innermost atom in the structure.
- The numbering begins from the innermost atom. The next numbers are given to atoms connected to it.

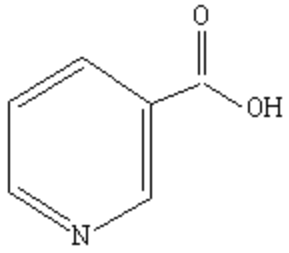
Procedure (see example)

1. Determine the 1st stage connectivity for each atom by counting the number of atoms connected to it.
2. Determine the 2nd stage connectivity for each atom by summing up the 1st stage connectivity values for all atoms connected to it,
3. Determine the (i+1)th stage connectivity value for each atom by summing up the connectivity values at ith stage for all atoms connected to it.
4. At each stage, count the number of unique connectivity values, *k*, in the structure.
5. As long as *k* increases, repeat step 3 and 4.
6. When reaching the stage where the number of connectivity values is constant or decreases, stop the procedure. Use the connectivity values of the previous stage.

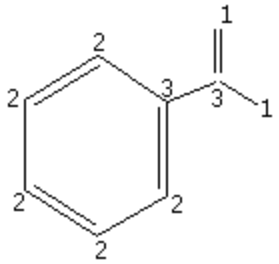
Procedure (Cont.)

7. Atom that has the largest connectivity value is given number 1.
8. Give number 2, 3, and so on to atoms connected to atom number 1 based on the connectivity values, from largest to smallest.
9. If two atoms have similar connectivity values, give the next number to the atom which is earlier in the alphabetical order. Carbon is given lower number that Oxygen. If the types of atom is the same, look at bond; atom with single bond are given lower number than double bonded atoms.
10. Atoms bonded to atom number 2 are numbered according to the sequence of their connectivity values. If two atoms have the same connectivity values give numbers based on the rules given in step 8.
11. Follow the procedure until all atoms are numbered.

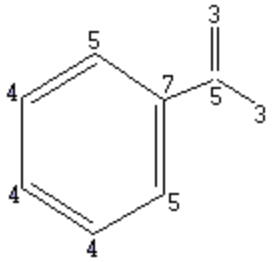
Example



Nicotinic Acid

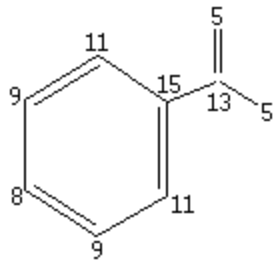


Stage 1:
k = 3 [1,2,3]

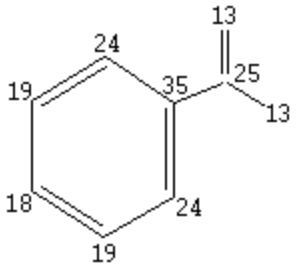


Stage 2:
k = 4 [3,4,5,7]

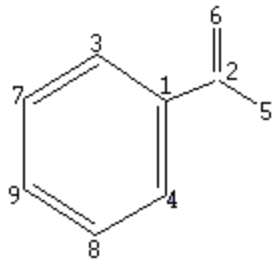
Example (Cont.)



Stage 3:
k = 6 [5,8,9,11,13,15]



Stage 4:
k = 6 [13,18,19,24,25,35]



A unique numbering system is achieved