L[0]: Representation of Chemical Compounds

Khalimat A. Murtazalieva

Fundamentals of Cheminformatics

@khalimat

September 21, 2021

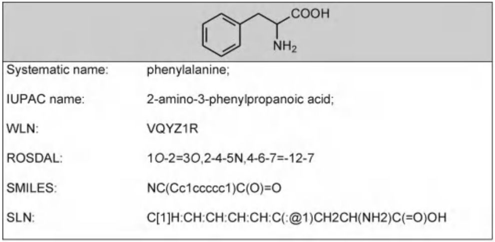
Overview

1. Line Notations
   - SMILES
2. 2D Notations
   - Graph Notations
     - Matrix Representations
     - Connection Table
   - Standard Structure Exchange Formats
     - Molfile
     - SDF format
3. Canonicalization
     - Morgan Algorithm
4. Special Notations of Chemical Structures
   - Markush Structures
   - Fragment Coding
   - Fingerprints
   - Hash table Hashed Fingerprints
5. References

Line notations

Definition

Line notations represent the structure of chemical compounds as a linear sequence of letters and numbers. The IUPAC nomenclature represents such a kind of line notation. There is a table below from [1] with examples of line notations.



Systematic name:	phenylalanine;
IUPAC name:	2-amino-3-phenylpropanoic acid;
WLN:	VQYZ1R
ROSDAL:	1O-2=3O,2-4-5N,4-6-7=-12-7
SMILES:	NC(Cc1ccccc1)C(O)=O
SLN:	C[1]H:CH:CH:CH:CH:C(:@1)CH2CH(NH2)C(=O)OH

Figure: Different line notations for the structure diagram of phenylalanine[1].

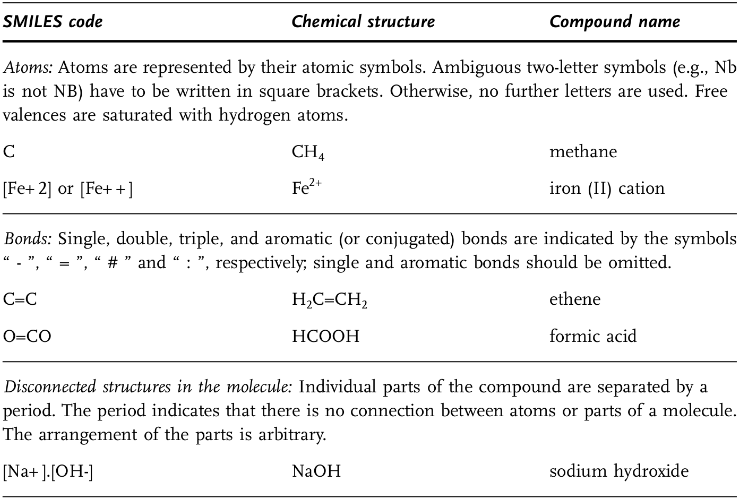
SMILES

The SMILES (Simplified Molecular Input Line Entry System) Coding was developed by David Weininger in 1986 for chemical data processing. The advantages of this system are that it enables to compress and simplify chemical information. The basic SMILES rules are:

- Atoms are represented by their atomic symbols.
- Hydrogen atoms automatically saturate free valences and are omitted (simple hydrogen connection).
- Neighboring atoms stand next to each other.
- Double and triple bonds are characterized by $=$ and $\#$, respectively.
- Branches are represented by parentheses.
- Rings are described by allocating digits to the two "connecting" ring atoms.

Further information can be found here.

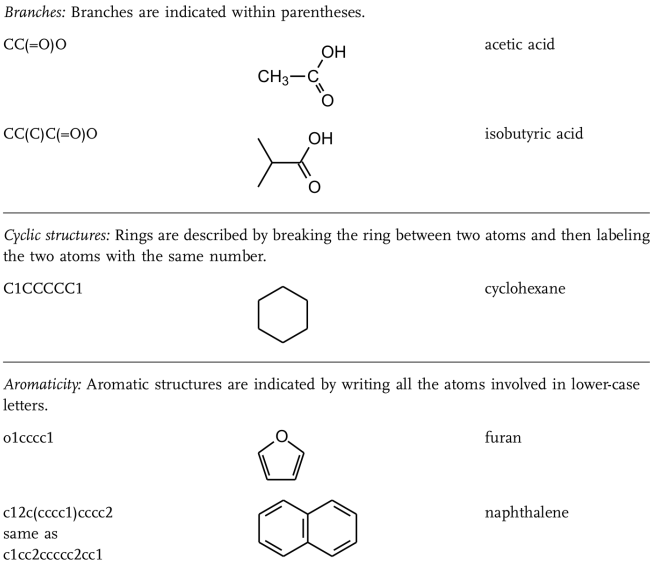
SMILES syntax, 1



SMILES code	Chemical structure	Compound name
Atoms: Atoms are represented by their atomic symbols. Ambiguous two-letter symbols (e.g., Nb is not NB) have to be written in square brackets. Otherwise, no further letters are used. Free valences are saturated with hydrogen atoms.		
C	CH_4	methane
[Fe+ 2] or [Fe+ +]	Fe^{2+}	iron (II) cation
Bonds: Single, double, triple, and aromatic (or conjugated) bonds are indicated by the symbols “ - ”, “ = ”, “ # ” and “ : ”, respectively; single and aromatic bonds should be omitted.		
C=C	$H_2C{=}CH_2$	ethene
O=CO	HCOOH	formic acid
Disconnected structures in the molecule: Individual parts of the compound are separated by a period. The period indicates that there is no connection between atoms or parts of a molecule. The arrangement of the parts is arbitrary.		
[Na+].[OH-]	NaOH	sodium hydroxide

Figure: SMILES syntax, from [1].

SMILES syntax, 2



Branches: Branches are indicated within parentheses.

CC(=O)O	acetic acid
CC(C)C(=O)O	isobutyric acid

OH
CH_3—C
O
OH
O

Cyclic structures: Rings are described by breaking the ring between two atoms and then labeling the two atoms with the same number.

C1CCCCC1	cyclohexane

Aromaticity: Aromatic structures are indicated by writing all the atoms involved in lower-case letters.

o1cccc1	furan
c12c(cccc1)cccc2 same as c1cc2ccccc2cc1	naphthalene

Intro

The analogy of structural diagrams to graphs is the reason why graph notations are widely used in cheminformatics. Hence, as a first step, we have to get acquainted with some basic definitions of graph theory. If you wish to achieve a deeper understanding, please pass this course. For our course, please memorize everything from this short intro. Moreover, please read this article. You must be able to reproduce and apply algorithms from the article during our course.

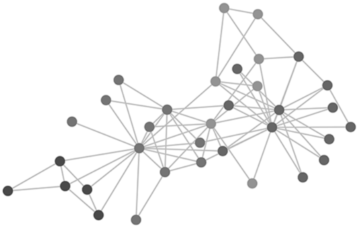
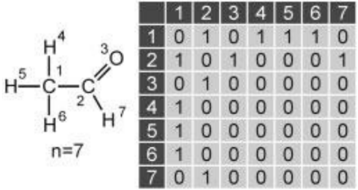


Figure: A graph

Matrix Representations

A graph can be represented as a matrix. Thus, quite early on, matrix representations of molecular structures were explored. Their major advantage is that the calculation of paths and cycles can be performed easily by well-known matrix operations.

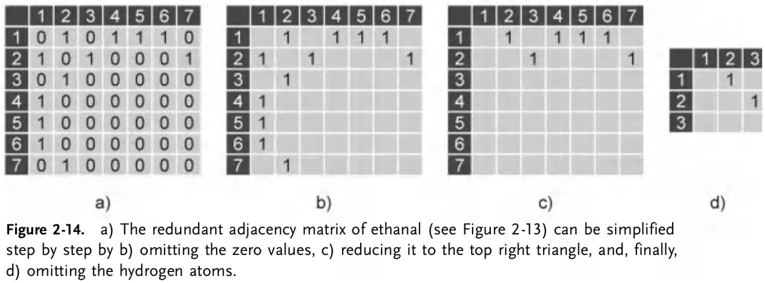


	1	2	3	4	5	6	7
1	0	1	0	1	1	1	0
2	1	0	1	0	0	0	1
3	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0

Figure: Adjacency matrix of ethanal[1].

Storage efficiency

The matrix representation is redundant. Hence, it could be simplified by omitting the zero 0, reducing to the top triangle, and omitting the hydrogen atoms.



	1	2	3	4	5	6	7
1	0	1	0	1	1	1	0
2	1	0	1	0	0	0	1
3	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0

a)

	1	2	3	4	5	6	7
1		1		1	1	1	
2	1		1				1
3		1					
4	1						
5	1						
6	1						
7		1					

b)

	1	2	3	4	5	6	7
1		1		1	1	1	
2			1				1
3							
4							
5							
6							
7							

c)

	1	2	3
1		1	
2			1
3			

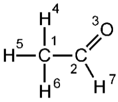
d)

Figure 2-14. a) The redundant adjacency matrix of ethanal (see Figure 2-13) can be simplified step by step by b) omitting the zero values, c) reducing it to the top right triangle, and, finally, d) omitting the hydrogen atoms.

Figure: A simplification procedure [1].

Connection Table

A major disadvantage of a matrix representation for a molecular graph is that the number of entries increases with the square of the number of atoms in the molecule. What is needed is a representation of a molecular graph where the number of entries increases only as a linear function of the number of atoms in the molecule. Such a representation can be obtained by listing, in tabular form only the atoms and the bonds of a molecular structure.

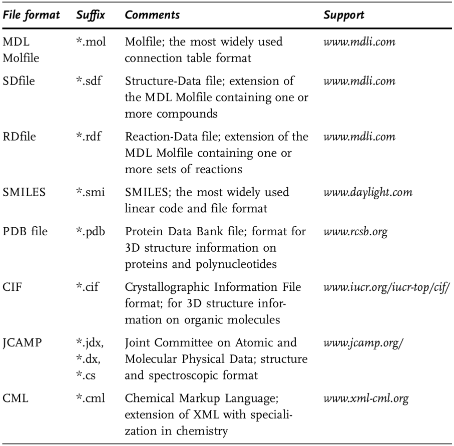


Atom list	
1	C
2	C
3	O
4	H
5	H
6	H
7	H

Bond list		
1st atom	2nd atom	bond order
1	2	1
2	3	2
2	7	1
1	4	1
1	5	1
1	6	1

Figure 2-20. A connection table: the structure diagram of ethanal, with the atoms arbitrarily labeled, is defined by a list of atoms and a list of bonds.

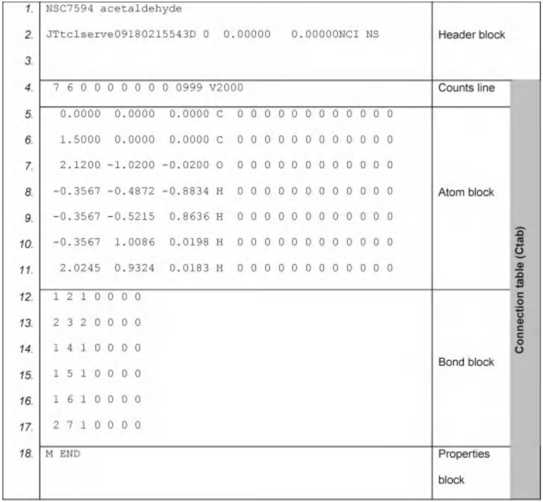
Overview



File format	*Suffix*	*Comments*	*Support*
MDL Molfile	*.mol	Molfile; the most widely used connection table format	*www.mdli.com*
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds	*www.mdli.com*
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions	*www.mdli.com*
SMILES	*.smi	SMILES; the most widely used linear code and file format	*www.daylight.com*
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides	*www.rcsb.org*
CIF	*.cif	Crystallographic Information File format; for 3D structure information on organic molecules	*www.iucr.org/iucr-top/cif/*
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format	*www.jcamp.org/*
CML	*.cml	Chemical Markup Language; extension of XML with specialization in chemistry	*www.xml-cml.org*

Figure: The most important file formats for exchange of chemical information [1].

An example of Molfile

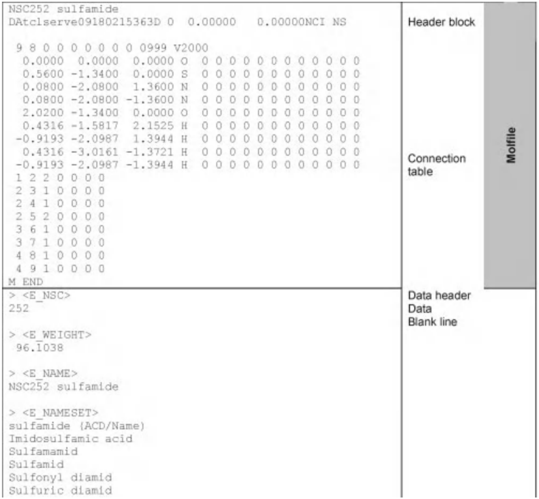


```
NSC7594 acetaldehyde
JTtclserve09180215543D 0   0.00000     0.00000NCI NS

  7  6  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1200   -1.0200   -0.0200 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3567   -0.4872   -0.8834 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3567   -0.5215    0.8636 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3567    1.0086    0.0198 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.0245    0.9324    0.0183 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  2  0  0  0  0
  1  4  1  0  0  0  0
  1  5  1  0  0  0  0
  1  6  1  0  0  0  0
  2  7  1  0  0  0  0
M  END
```

Figure: Molfile representing the ethanal structure [1].

An example of SDF-file



```
NSC252 sulfamide
DAtclserve09180215363D 0   0.00000    0.00000NCI NS

 9 8 0 0 0 0 0 0 0 0999 V2000
  0.0000  0.0000  0.0000 O   0 0 0 0 0 0 0 0 0 0 0 0
  0.5600 -1.3400  0.0000 S   0 0 0 0 0 0 0 0 0 0 0 0
  0.0800 -2.0800  1.3600 N   0 0 0 0 0 0 0 0 0 0 0 0
  0.0800 -2.0800 -1.3600 N   0 0 0 0 0 0 0 0 0 0 0 0
  2.0200 -1.3400  0.0000 O   0 0 0 0 0 0 0 0 0 0 0 0
  0.4316 -1.5817  2.1525 H   0 0 0 0 0 0 0 0 0 0 0 0
 -0.9193 -2.0987  1.3944 H   0 0 0 0 0 0 0 0 0 0 0 0
  0.4316 -3.0161 -1.3721 H   0 0 0 0 0 0 0 0 0 0 0 0
 -0.9193 -2.0987 -1.3944 H   0 0 0 0 0 0 0 0 0 0 0 0
 1 2 2 0 0 0 0
 2 3 1 0 0 0 0
 2 4 1 0 0 0 0
 2 5 2 0 0 0 0
 3 6 1 0 0 0 0
 3 7 1 0 0 0 0
 4 8 1 0 0 0 0
 4 9 1 0 0 0 0
M END
> <E_NSC>
252

> <E_WEIGHT>
 96.1038

> <E_NAME>
NSC252 sulfamide

> <E_NAMESET>
sulfamide (ACD/Name)
Imidosulfamic acid
Sulfamamid
Sulfamid
Sulfonyl diamid
Sulfuric diamid
```

Figure: SDF-file representing the ethanal figure [1].

Canonicalization

A structure with n atoms can be numbered in n! different manners and thus has up to n! A process of converting input representation to canonical form is called “canonicalisation” or “canonisation”. Morgan’s algorithm is commonly used for this purpose. It is based on extended connectivity (EC) - considering the degree of neighbours of an atom. The next slide illustrates the principle.

Morgan’s algorithm

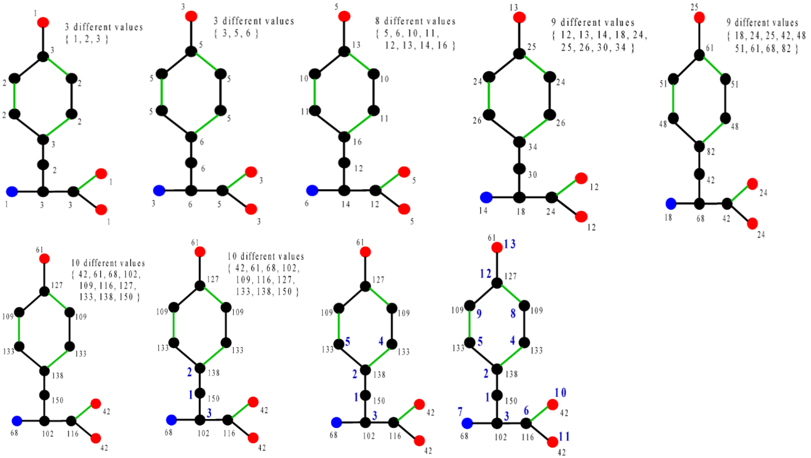
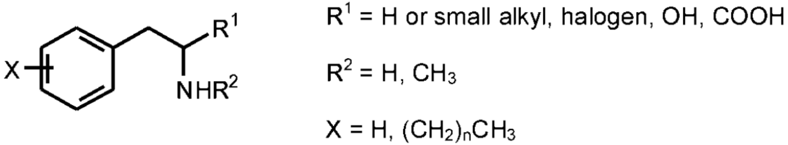


Figure: The Morgan’s algorithm

Markush Structures

Markush structures are mainly used in patents, for protecting compounds related to an invention. The Markush structure diagram is a specific type of representation of a series of chemical compounds. This diagram can describe not only a specific molecule but also various compound families, which is why it is also called a generic structure diagram.



R^1 = H or small alkyl, halogen, OH, COOH

R^2 = H, CH_3

X = H, $(CH_2)_nCH_3$

Figure: An example of a Markush Structure [1]

Fragment Coding

Fragment codes have always played an important role in chemical information systems. Basically, they are indexing expressions of chemical structures. Usually, these are small assemblies of atoms, functional groups, ring systems, etc., which can be specified beforehand.

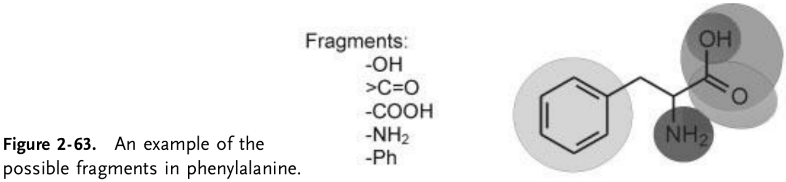


Figure 2-63. An example of the possible fragments in phenylalanine.

Fragments:
- -OH
- >C=O
- -COOH
- $-NH_2$
- -Ph

Figure: An example of the possible fragments in phenylalanine [1]

Fingerprints

A fingerprint is a characteristic property used to describe molecules. So a molecule, for example, can be described by the structure or structural keys. These keys indicate whether or not a specific substructure or fragment exists in the molecule. The fragments of chemical structures can be coded in binary keys.

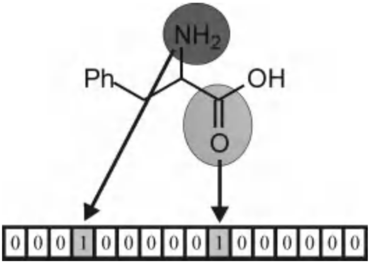


Figure 2-64. How an excerpt from a binary code could appear, if only –NH$_2$ and C=O are available in the fragment library.

Figure: An example of fingerprints coding [1]

Hash table

A hash table (hash map) is a data structure that implements an associative array abstract data type, a structure that can map keys to values. A hash table uses a hash function to compute an index, also called a hash code, into an array of buckets or slots, from which the desired value can be found. During lookup, the key is hashed and the resulting hash indicates where the corresponding value is stored.

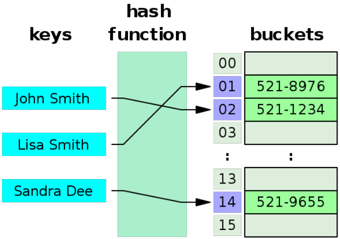


Figure: Hashing, from Wikipedia [1]

Hashed Fingerprints

In this procedure, all bonds in the molecule are traversed, starting at an atom and proceeding through several (e.g., seven) bond lengths. Thereby, one receives information about the substructures of the molecule and also about its internal relationships.

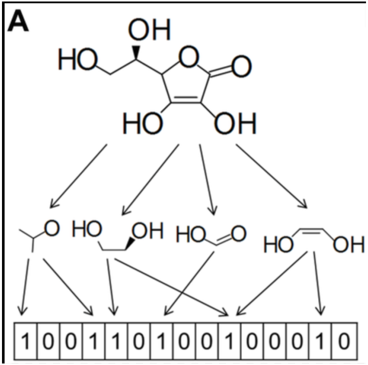


Figure: Hashed Fingerprints [2]

References

Johann Gasteiger and Thomas Engel. *Chemoinformatics: a textbook*. John Wiley & Sons, 2006.

Marek Śmieja and Dawid Warszycki. "Average information content maximization—a new approach for fingerprint hybridization and reduction". In: *PloS one* 11.1 (2016), e0146666.

TO DO

- Install RDKit
- Go through the RDKit's tutorial, commit a notebook with your code to GitHub.
- Read chapter 2 (Representation of Chemical Compounds, 15-157 pp) from [1]