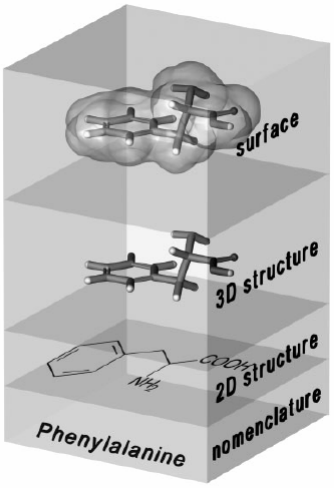
Cheminformatica – visualizzatori, editor

Programmi di visualizzazione delle strutture molecolari

p.es. *ChemSketch* (http://www.acdlabs.com/resources/freeware/)

Disegno di composti lineari e ciclici
Misura delle distanze di legame
Misura degli angoli e degli angoli diedri
Utilizzo degli strumenti di rappresentazione con linee tratteggiate e a cuneo
Generazione della nomenclatura IUPAC
Visualizzazione 2D e 3D

Rappresentazioni grafiche di molecole



Cheminformatica – catalogazione e nomenclature

Catalogazione

CAS Registry (http://www.cas.org/)
Informazione chimica
Indicizzazione

PubChem (http://pubchem.ncbi.nlm.nih.gov/)
Informazione chimica e biologica
Ricerca di composti

Notazioni lineari

Smiles (http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html)
Regole di nomenclatura

InChI (http://www.iupac.org/inchi/)
Formato

Formati tabulari

Informazioni topologiche
- Matrici di adiacenza
- Matrici di incidenza
- Matrici di distanza

Informazioni spaziali
- Coordinate cartesiane
- Coordinate interne

MDL MOL format (http://www.symyx.com/index.jsp)

Cheminformatica – formati

Table 2-5. The most important file formats for exchange of chemical structure information.

File format	*Suffix*	*Comments*	*Support*	*Ref.*
MDL Molfile	*.mol	Molfile; the most widely used connection table format	*www.mdli.com*	50
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds	*www.mdli.com*	50
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions	*www.mdli.com*	50
SMILES	*.smi	SMILES; the most widely used linear code and file format	*www.daylight.com*	20, 21
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides	*www.rcsb.org*	53
CIF	*.cif	Crystallographic Information File format; for 3D structure information on organic molecules	*www.iucr.org/iucr-top/cif/*	55
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format	*www.jcamp.org/*	56
CML	*.cml	Chemical Markup Language; extension of XML with specialization in chemistry	*www.xml-cml.org*	57–59

SMILES

- *Simplified Molecular Input Line Entry Specification*
- A string of letters, numbers and other characters that specify the atoms, their connectivity, bond orders, & chirality
- http://www.daylight.com/smiles/f_smiles.html

Depiction	SMILES	Name
H_2O	O	water
CH_4	C	methane
	CC(=O)O	acetic acid
	C1CCCCC1	cyclohexane
	c1ccccc1	benzene
	c1ccccc1[N+](=O)[O-]	nitrobenzene

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

Cheminformatica – notazioni lineari

Branches: Branches are indicated within parentheses.

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

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

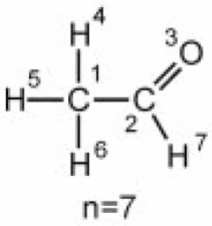
C1CCCCC1		cyclohexane

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

o1cccc1		furan
c12c(cccc1)cccc2 same as c1cc2ccccc2cc1		naphthalene

Cheminformatica – notazioni tabulari

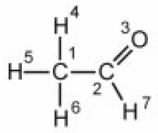
The matrix of a structure with n atoms consists of an array of $n \times n$ entries. A molecule with its different atoms and bond types can be represented in matrix form in different ways depending on what kind of entries are chosen for the atoms and bonds. Thus, a variety of matrices has been proposed: adjacency, distance, incidence, bond, and bond–electron matrices.



	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

Adiacency matrix of ethanal

Cheminformatica – notazioni tabulari



a)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1.400	2.190	1.022	1.023	1.022	2.106
C2	1.400	0	1.123	1.999	1.982	1.999	1.022
O3	2.190	1.123	0	2.349	2.708	2.995	1.859
H4	1.022	1.999	2.349	0	1.668	1.661	2.895
H5	1.023	1.982	2.708	1.668	0	1.668	2.562
H6	1.022	1.999	2.955	1.661	1.668	0	2.336
H7	2.106	1.022	1.859	2.895	2.566	2.336	0

b)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1	2	1	1	1	2
C2	1	0	1	2	2	2	1
O3	2	1	0	3	3	3	2
H4	1	2	3	0	2	2	3
H5	1	2	3	2	0	2	3
H6	1	2	3	2	2	0	3
H7	2	1	2	3	3	3	0

Distance matrices of ethanal with a) geometric distances in Å and b) topological distances. The matrix elements of b) result from counting the number of bonds along the shortest walk between the chosen atoms.

Cheminformatica – notazioni tabulari

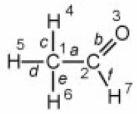
The incidence matrix is an $n \times m$ matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m). An entry obtains the value of 1 if the corresponding edge ends in this particular node

	C1	C2	O3	H4	H5	H6	H7
a	1	1	0	0	0	0	0
b	0	1	1	0	0	0	0
c	1	0	0	1	0	0	0
d	1	0	0	0	1	0	0
e	1	0	0	0	0	1	0
f	0	1	0	0	0	0	1

a)

	C1	C2	O3	H4	H5	H6	H7
a	1	1					
b		1	1				
c	1			1			
d	1				1		
e	1					1	
f		1					1

b)



n=7; m=6

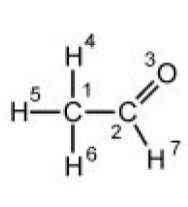
	C1	C2	O3
a	1	1	
b		1	1

c)

a) The redundant incidence matrix of ethanal can be compressed by b) omitting the zero values and c) omitting the hydrogen atoms. In the non-square matrix, the atoms are listed in columns and the bonds in rows.

Cheminformatica – notazioni tabulari

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.



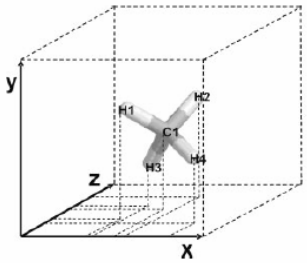
atom index	element	1st index of atom	bond order	2nd index of atom	bond order
1	C	2	1		
2	C			3	2
3	O				

Cheminformatica – rappresentazioni tridimensionali

Clearly, the next step is the handling of a molecule as a real object with a spatial extension in 3D space. Quite often this is also a mandatory step, because in most cases the 3D structure of a molecule is closely related to a large variety of physical, chemical, and biological properties. In addition, the fundamental importance of an unambiguous definition of stereochemistry becomes obvious, if the 3D structure of a molecule needs to be derived from its chemical graph. The molecules of stereoisomeric compounds differ in their spatial features and often exhibit quite different properties. Therefore, stereochemical information should always be taken into account if chiral atom centers are present in a chemical structure.

Basically, two different methods are commonly used for representing a chemical structure in 3D space. Both methods utilize different coordinate systems to describe the spatial arrangement of the atoms of a molecule under consideration.

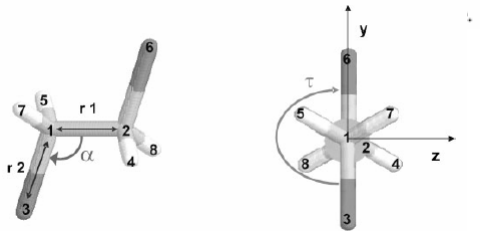
The most common way is to choose a Cartesian coordinate system, i.e., to code the *x*-, *y*-, and *z*-coordinates of each atom, usually as floating point numbers. For each atom the Cartesian coordinates can be listed in a single row, giving consecutively the *x*-, *y*-, and *z*-values.



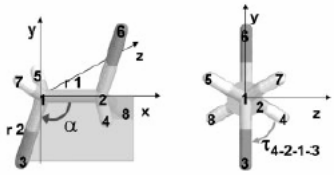
	x	y	z
C1	-0.0127	1.0858	0.0080
H1	0.0021	-0.0041	0.0020
H2	1.0099	1.4631	0.0003
H3	-0.5399	1.4469	-0.8751
H4	-0.5229	1.4373	0.9048

The second method for representing a molecule in 3D space is to use internal coordinates such as bond lengths, bond angles, and torsion angles. Internal coordinates describe the spatial arrangement of the atoms relative to each other.

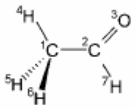
The most common way to describe a molecule by its internal coordinates is the so-called *Z-matrix*.



C1						
C2	1.5	1				
C13	1.7	1	109	2		
H4	1.1	2	109	1	-60	3
H5	1.1	1	109	2	180	4
C16	1.7	2	109	1	60	5
H7	1.1	1	109	2	-60	6
H8	1.1	2	109	1	180	7



Cheminformatica – notazioni tabulari tridimensionali– MDL files



1.	NSC7594 acetaldehyde	Header block	
2.	JTtcl1serve09180215543D 0 0.00000 0.00000CNCI NS		
3.			
4.	7 6 0 0 0 0 0 0 0 0999 V2000	Counts line	Connection table (Ctab)
5.	0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0	Atom block	
6.	1.5000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0		
7.	2.1200 -1.0200 -0.0200 O 0 0 0 0 0 0 0 0 0 0 0 0		
8.	-0.3567 -0.4872 -0.8834 H 0 0 0 0 0 0 0 0 0 0 0 0		
9.	-0.3567 -0.5215 0.8636 H 0 0 0 0 0 0 0 0 0 0 0 0		
10.	-0.3567 1.0086 0.0198 H 0 0 0 0 0 0 0 0 0 0 0 0		
11.	2.0245 0.9324 0.0183 H 0 0 0 0 0 0 0 0 0 0 0 0		
12.	1 2 1 0 0 0 0	Bond block	
13.	2 3 2 0 0 0 0		
14.	1 4 1 0 0 0 0		
15.	1 5 1 0 0 0 0		
16.	1 6 1 0 0 0 0		
17.	2 7 1 0 0 0 0		
18.	M END	Properties block	

Cheminformatica – notazioni tabulari tridimensionali – MDL files

Description	Number of atoms	Number of bonds	Number of atom lists	(obsolete)	Chiral flag	Other properties ignored for Molfiles	Number of additional properties	Current Ctab version
Column	123	456	789	1 012	345	2 3 678901234567890	123	456789
Data	7	6	0	0	0	0 0 0 0 0	999	V2000

Description	Cartesian coordinates (x, y, z)			(space)	Atom symbol	Mass difference	Charge	9 miscellaneous properties
Column	1 1234567890	2 1234567890	3 1234567890	1	234	56	789	4 012...
Data	0.0000	0.0000	0.0000		C	0	0	0...
	1.5000	0.0000	0.0000		C	0	0	0...
	2.1200	-1.0200	-0.0200		O	0	0	0...