Chemical table file

Chemical table file (CT File) is a family of text-based chemical file formats that describe molecules and chemical reactions. One format, for example, lists each atom in a molecule, the x-y-z coordinates of that atom, and the bonds among the atoms.

Contents

File formats

There are several file formats in the family.

The formats were created by MDL Information Systems (MDL), which was acquired by Symyx Technologies then merged with Accelrys Corp., and now called BIOVIA, a subsidiary of Dassault Systemes of Dassault Group.[1]

CT File is an open format, BIOVIA publishes its specification.[2] BIOVIA requires users to register to download the CTFile format specifications.[3]

Molfile

ctab

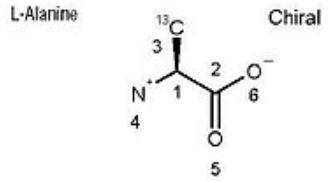
Filename extension	`.mol`
Internet media type	`chemical/x-mdl-molfile`
Type of format	chemical file format

An **MDL Molfile** is a file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.

The molfile consists of some header information, the Connection Table (CT) containing atom info, then bond connections and types, followed by sections for more complex information.

The molfile is sufficiently common that most, if not all, cheminformatics software systems/applications are able to read the format, though not always to the same degree. It is also supported by some computational software such as Mathematica.

The current *de facto* standard version is molfile V2000, although, more recently, the V3000 format has been circulating widely enough to present a potential compatibility issue for those applications that are not yet V3000-capable.



The contents of a Molfile of L-Alanine

`L-Alanine`	**Title line** (can be blank but line must exist)	**Header Block** **(3 lines)**
`ABCDEFGH09071717443D`	**Program / file timestamp line** (Name of source program and a file timestamp)	
`Exported`	**Comment line** (can be blank but line must exist)	
`6 5 0 0 1 0 3 V2000`	**Counts line**	**Connection table**

`-0.6622 0.5342 0.0000 C 0 0 2 0 0 0`
`0.6622 -0.3000 0.0000 C 0 0 0 0 0 0`
`-0.7207 2.0817 0.0000 C 1 0 0 0 0 0`
`-1.8622 -0.3695 0.0000 N 0 3 0 0 0 0`
`0.6220 -1.8037 0.0000 O 0 0 0 0 0 0`
`1.9464 0.4244 0.0000 O 0 5 0 0 0 0`
	Atom block (1 line for each atom): x, y, z (in angstroms), element, etc.	

`1 2 1 0 0 0`
`1 3 1 1 0 0`
`1 4 1 0 0 0`
`2 5 2 0 0 0`
`2 6 1 0 0 0`
	Bond block (1 line for each bond): 1st atom, 2nd atom, type, etc.	

`M CHG 2 4 1 6 -1` `M ISO 1 3 13`	**Properties block**	
`M END`	**END line** (NOTE: some programs don't like a blank line before M END)	**END**

Counts line block specification

Value	6	6	0	0	0	1	V2000
Description	number of atoms	number of bonds	number of atom list	Chiral flag, 1 = chiral; 0 = not chiral	number of stext entries	number of lines of additional properties	mol version
Type	[Generic]	[Generic]	[Query]	[Generic]	[ISIS/Desktop]	[Generic]	

Bond block specification

The Bond Block is made up of bond lines, one line per bond, with the following format:

111 222 ttt sss xxx rrr ccc

where the values are described in the following table:

Field	Meaning	Values
111	first atom number	
222	second atom number	
ttt	bond type	1= Single, 2 = Double, 3 = Triple, 4 = Aromatic,5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any
sss	bond stereo	For single bonds: 0 = not stereo; 1= up; 4=either, 6= down For double bonds: 0= Use x-, y-, z-coords from atom block to determine cis or trans; 3=Cis or trans (either) double bond
xxx	not used	
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible

Extended Connection Table (V3000)

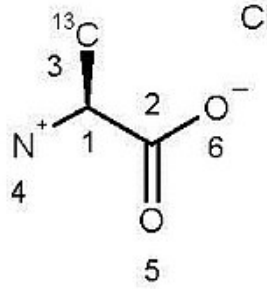
The extended (V3000) molfile consists of a regular molfile “no structure” followed by a single molfile appendix that contains the body of the connection table (Ctab). The following figure shows both an alanine structure and the extended molfile corresponding to it.

Note that the “no structure” is flagged with the “V3000” instead of the “V2000” version stamp. There are two other changes to the header in addition to the version:

- The number of appendix lines is always written as 999, regardless of how many there actually are. (All current readers will disregard the count and stop at M END.)
- The “dimensional code” is maintained more explicitly. Thus “3D” really means 3D, although “2D” will be interpreted as 3D if any non-zero Z-coordinates are found.

Unlike the V2000 molfile, the V3000 extended Rgroup molfile has the same header format as a non-Rgroup molfile.

L-Alanine Chiral



L-Alanine	**Description**	**Header block**
GSMACCS-II07189510252D 1 0.00366 0.00000 0	**Header with timestamp**	
Figure 1, J. Chem. Inf. Comput. Sci., Vol 32, No. 3., 1992	**Comment line**	
0 0 0 0 0 999 V3000	**V2000-compatibility line**	
M V30 BEGIN CTAB		**Connection table**
M V30 COUNTS 6 5 0 0 1	**Counts line**	
M V30 BEGIN ATOM M V30 1 C -0.6622 0.5342 0 0 CFG=2 M V30 2 C 0.6622 -0.3 0 0 M V30 3 C -0.7207 2.0817 0 0 MASS=13 M V30 4 N -1.8622 -0.3695 0 0 CHG=1 M V30 5 O 0.622 -1.8037 0 0 M V30 6 O 1.9464 0.4244 0 0 CHG=-1 M V30 END ATOM	**Atom block**	
M V30 BEGIN BOND M V30 1 1 1 2 M V30 2 1 1 3 CFG=1 M V30 3 1 1 4 M V30 4 2 2 5 M V30 5 1 2 6 M V30 END BOND	**Bond block**	
M V30 END CTAB M END		

Counts line

A counts line is required, and must be first. It specifies the number of atoms, bonds, 3D objects, and Sgroups. It also specifies whether or not the CHIRAL flag is set. Optionally, the counts line can specify molregno. This is only used when the regno exceeds 999999 (the limit of the format in the molfile header line). The format of the counts line is:

```
         M V30 COUNTS na nb nsg n3d chiral
```

```
M V30 COUNTS na  nb  nsg n3d chiral [REGNO=regno]
M V30 COUNTS 6   5   0   0   1
```

number of atoms
number of bonds
number of Sgroups
number of 3D constrains
if 1 = molecule is chiral
molecule or model regno

SDF

SDF is one of a family of chemical-data file formats developed by MDL; it is intended especially for structural information. "SDF" stands for structure-data file, and SDF files actually wrap the molfile (MDL Molfile) format. Multiple compounds are delimited by lines consisting of four dollar signs ($$$$). A feature of the SDF format is its ability to include associated data.

ctab

Filename extension	`.sd, .sdf`
Internet media type	`chemical/x-mdl-sdfile`
Type of format	chemical file format

Associated data items are denoted as follows:

```
>  <Unique_ID>
XCA3464366

>  <ClogP>
5.825

>  <Vendor>
Sigma

>  <Molecular Weight>
499.611
```

Multiple-line data items are also supported. The MDL SDF-format specification requires that a hard-carriage-return character be inserted if a single line of any text field exceeds 200 characters. This requirement is frequently violated in practice, as many SMILES and InChI strings exceed that length.

Other formats of the family

There are other, less commonly used formats of the family:

- **RXNFile** - for representing a single chemical reaction;
- **RDFile** - for representing a list of records with associated data. Each record can contain chemical structures, reactions, textual and tabular data;
- **RGFile** - for representing the Markush structures (deprecated, Molfile V3000 can represent Markush structures);
- **XDFile** - for representing chemical information in XML format.

See also

- Chemical file format#Converting Between Formats

References

1. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. (1992). "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited". *Journal of Chemical Information and Modeling*. **32** (3): 244. doi:10.1021/ci00007a012 (https://doi.org/10.1021%2Fci00007a012).
2. "CT File Formats" (https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf) (PDF). Biovia. August 2020. Archived (https://web.archive.org/web/20210219065450/https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf) (PDF) from the original on 2021-02-19. Retrieved 2021-02-19.
3. "Registration form" (https://discover.3ds.com/ctfile-documentation-request-form). Biovia. Archived (https://web.archive.org/web/20201001232143/https://discover.3ds.com/ctfile-documentation-request-form) from the original on 2020-10-01. Retrieved 2021-02-19.

External links

- SDF Toolkit (http://cactus.nci.nih.gov/SDF_toolkit/) free software to process SD files (SDF).
- NCI/CADD Chemical Identifier Resolver (http://cactus.nci.nih.gov/chemical/structure) generates SD files (SDF) from chemical names, CAS Registry Numbers, SMILES, InChI, InChIKey, ...
- KNIME (http://www.knime.org/) free software to manipulate data and do datamining, can also read and write SD files (SDF).
- Comparative Toxicology Dashboard (https://comptox.epa.gov/dashboard) service provided by the Environmental Protection Agency (EPA) which generates SD files (SDF) from chemical names, CAS Registry Numbers, SMILES, InChI, InChIKey, ...