Chemical File Format Conversion Tools : An Overview

Kavitha C. R
Research Scholar, Bharathiyar University
Dept of Computer Applications
SNGIST
Cochin, India

Dr. T Mahalekshmi
Principal
Sree Narayana Institute of Technology
Kollam, India

***Abstract*— There are a lot of chemical data stored in large databases, repositories and other resources. These data are used by different researchers in different applications in various areas of chemistry. Since these data are stored in several standard chemical file formats, there is a need for the inter-conversion of chemical structures between different formats because all the formats are not supported by various software and tools used by the researchers. Therefore it becomes essential to convert one file format to another. This paper reviews some of the chemical file formats and also presents a few inter-conversion tools such as Open Babel [1], Mol converter [2] and CncTranslate [3].**

Keywords*— File format Conversion, Open Babel, mol converter, CncTranslate, inter- conversion tools.

I. INTRODUCTION

Cheminformatics deals with the process of storing and retrieving information about chemical compounds. These chemical compounds are available in different formats in the chemical databases [4, 5] and other repositories. These representations focus on specific atomic or molecular information and may not attempt to store all possible chemical data. For example, coordinate information is available in line notations like Simplified molecular-input line-entry system (*S*MILES) [6], chemical bonding data is not present in crystallographic or quantum mechanical formats and Hydrogen atoms are omitted in x-ray crystallography format. There is some loss of information in other types of representations also. In order to avoid these problems, a standard format has been used for storing chemical data, which also include the recently developed Chemical Markup Language (CML) [7] format. Using such standard formats for storing chemical data makes everything much easier i.e. the output of one module can be the input for another. But a difficulty that is faced by the researchers in the field of cheminformatics area is the interconversion of molecular structures between different formats which is a process of extracting and interpreting the chemical data and semantics.

The main aim of this paper is to present the readers with the review of inter-conversion of molecular structures between different formats. The remaining paper is organized into 4 sections. Section II gives an overview of the five different chemical file formats. Three types of file format conversion tools are discussed in section III. And the conclusion is given in section IV followed by the references.

II. CHEMICAL FILE FORMATS

A chemical is a collection of atoms bonded together in space. The structure of a chemical makes it unique and gives it its physical and biological characteristics. This structure is represented in a variety of chemical file formats. These formats are used to represent chemical structure records and its associated data fields. Some of the file formats are CML (Chemical Markup Language), SDF (Structural Data format), PDB (Protein Data Bank), SMILES (Simplified Molecular Input Line Entry Specifications) and XYZ file format. Using different chemical file formats, molecules can be represented and can be saved with their corresponding extension like ".pdb", ".sdf", ".xyz" etc. The following paragraphs discuss some of the existing file formats.

A. Chemical Markup Language (CML)

Chemical Markup Language (ChemML or CML) [7] was developed by Peter Murray-Rust and Henry Rzepa in 1995. Chemical markup language (CML) was developed for containing chemical information components within documents. CML supports a wide range of chemical concepts such as molecules, reactions, spectra and analytical data, computational chemistry and chemical crystallography and materials. As an example, the CML for methanol [8] is given below:

```
<molecule id="METHANOL">
atomArray>
    <stringArray builtin="elementType">C O H H
H H</stringArray>
  <floatArray builtin="x3" units="pm">
    -0.748 0.558 -1.293 -1.263 -0.699 0.716
  </floatArray>
atomArray>
</molecule>
```

B. Structure Data Format (SDF)

Structure Data Format (SDF) [9] also known as SD Files is a common chemical file formats used to represent

multiple chemical structure records and associated data fields. SDF was developed and published by Molecular Design Limited (MDL) and became the most widely used standard for importing and exporting information on chemicals. The general format of an SDF file consists of blocks of information, with a single compound record format represented as shown in the figure1 below.

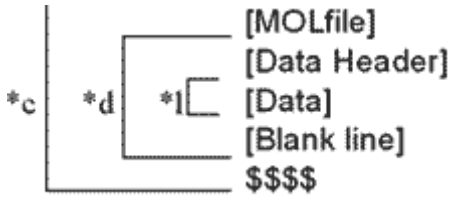

Fig. 1 SDF file general format [10]

The figure 1 represents the SDF file format where ***c** = compound record format is repeated for the length of the SDF file, ***d** = data item format is repeated for each data item associated with a compound record, ***l** = a separate line is used for each data value. MOL file format is the MDL format for storage of chemical structure information. A sample SDF file for 1, 2-trans-dichloroethene containing 4 data fields each is shown in figure 2 below.

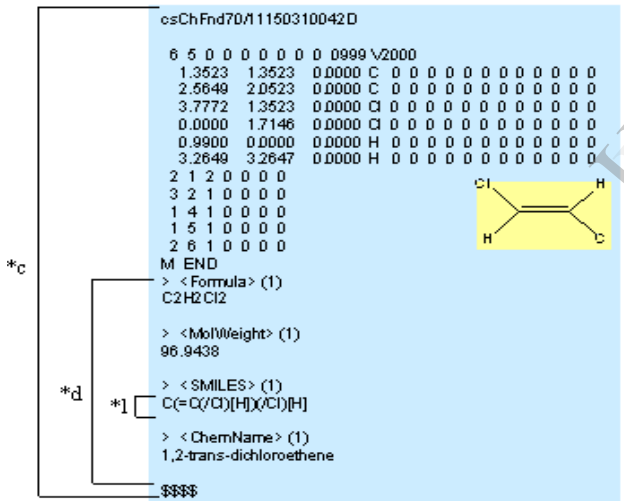

Fig. 2 sample SDF file 1, 2-trans-dichloroethene [11]

C. Protein Data Bank (PDB)

The Protein Data Bank (PDB) file format [12] is a textual file format describing the three dimensional structures of molecules held in the Protein Data Bank. The PDB format provides protein description and annotation, nucleic acid structures including atomic coordinates, observed side chain rotamers, secondary structure assignments and also the atomic connectivity. Structures which are deposited with other molecules such as water, ions, nucleic acids, ligands etc. can also be described in the pdb format.

A PDB file describing a protein can contain hundreds to thousands of lines. The figure 3 shows a sample PDB file which describes the structure of a synthetic collagen-like peptide.

```
HEADER    EXTRACELLULAR MATRIX                  22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-
LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
```

Fig. 3 PDB file describing a protein [12]

HEADER (1st line in Fig 3), TITLE (2nd line in Fig 3) and AUTHOR (6th line in Fig 3) records gives the researcher's information who defined the structure. REMARK (9th line in Fig 3) records contains free-form annotation, but can also include standardized information; for example, the REMARK 350 BIOMT records describe how to compute the coordinates of the experimentally observed multimer from those of the explicitly specified ones of a single repeating unit. SEQRES (13th line in Fig 3) records give the sequences of the three peptide chains (named A, B and C), which are very short in this example but usually span multiple lines. ATOM (16th line in Fig 3) records describe the coordinates of the atoms that are part of the protein. For example, the first ATOM line above describes the alpha-N atom of the first residue of peptide chain A, which is a proline residue; the first three floating point numbers are its x, y and z coordinates and are in units of Angstroms. The next three columns are the occupancy, temperature factor, and the element name, respectively. HETATM (21st) records describe coordinates of hetero-atoms, i.e. those atoms which are not part of the protein molecule. [12]

D. SMILES

SMILES (Simplified Molecular Input Line Entry Specification) notation was developed in 1988 by Weininger. [6] A SMILE is a string notation used to describe the nature and topology of molecular structures. It is a specification in the form of a line notation for describing the structure of chemical molecules using American Standard Code for Information Interchange (ASCII) strings. SMILES strings include connectivity but do not include 2D or 3D coordinates.

Hydrogen atoms are not represented. Other atoms are represented by their element symbols B (boron), C (carbon), N (nitrogen), O (oxygen), F (Fluorine) , P (Phosphorous) , S (Sulphur), Cl (chlorine), Br (bromine), and I (Iodine). The symbol "=" represents double bonds and "#" represents triple bonds. Branching is indicated by (). Rings are indicated by pairs of digits. Figure 4.a shows a traditional two-dimensional structure diagram of the chemical structure of aspirin and Figure 4.b shows its SMILES notation, both without explicit hydrogen.

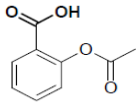

Fig.4.a: molecular structure of aspirin [5]

```
O=C(O)c1ccccc1OC(=O)C
```

Fig. 4.b: SMILES notation of aspirin [5]

E. XYZ

The XYZ file format [13] is a chemical file format which specifies the molecule geometry by giving the number of atoms with Cartesian coordinates that will be read on the first line, a comment on the second, and the lines of atomic coordinates in the following lines. The XYZ file format as shown in the figure 5 is used in computational chemistry programs for importing and exporting geometries. The units are generally in Ångströms. Some variations include using atomic numbers instead of atomic symbols, or skipping the comment line. Files using the XYZ format conventionally have the `.xyz` extension.

```
<number of atoms>
comment line
atom_symbol1 x-coord1 y-coord1 z-coord1
atom_symbol2 x-coord2 y-coord2 z-coord2
...
atom_symboln x-coordn y-coordn z-coordn
```

Fig. 5 .xyz file format [13]

A sample XYZ format of the methane molecule is shown in the figure 6 given below.

```
5
methane molecule (in Ångströms)
C        0.000000        0.000000        0.000000
H        0.000000        0.000000        1.089000
H        1.026719        0.000000       -0.363000
H       -0.513360       -0.889165       -0.363000
H       -0.513360        0.889165       -0.363000
```

Fig. 6 methane molecule in xyz format

III. FILE FORMAT CONVERSION TOOLS

The Chemical data is stored in different file formats in different application areas like databases, visualizing programs and modeling in the real world. Each of these programs represents their molecular data in their own file formats including 2D, 3D or symmetric representations which may not be acceptable to all software. In such cases one may need to convert files to their respective format for further processing.

Chemical File format conversion is the process of converting the chemical data files in one format (input format) to another format (output format) so that these files can be effectively used in different platforms. There is a need to interconvert formats of chemical data because there is a huge number of an application in the different areas of chemistry. The chemical data may be stored in different formats (2D, 3D, SMILES for example). There are different tools available for performing file format conversion. Few tools such as Open Babel, mol converter and CncTranslate are discussed here.

A. Open Babel

OpenBabel [1] was developed by OELib which is opensource software by OpenEye Scientific under the GPL (General Public License). OpenBabel is available as free to download, which runs on Windows, Linux, Mac OSX as a cross platform. The two important components of OpenBabel are a command line utility which is used to translate between different file formats and a C++ library which contains file conversion codes and other utilities for various open source scientific software. It offers an “extensible plugin interface” for file formats, fingerprints, charge models, descriptors, and molecular mechanics force fields.

OpenBabel is a free open source tool designed to convert different molecular file formats. It helps to search, convert, analyze or store data which has a wide range of applications in the different fields of molecular modeling, computational chemistry etc. It converts file formats used in cheminformatics including SMILES, MOL, MOL2 as input and computational chemistry packages like GAMESS, Gaussian, MOPAC etc. as output. It also converts crystallographic file formats (CIF, ShelX), reaction formats (MDLRXN), molecular dynamics and docking (AutoDock, Amber), 2D drawing packages (ChemDraw), 3D viewers (Chem3D, Molden) and chemical kinetics and thermodynamics (ChemKin, Thermo). [14]

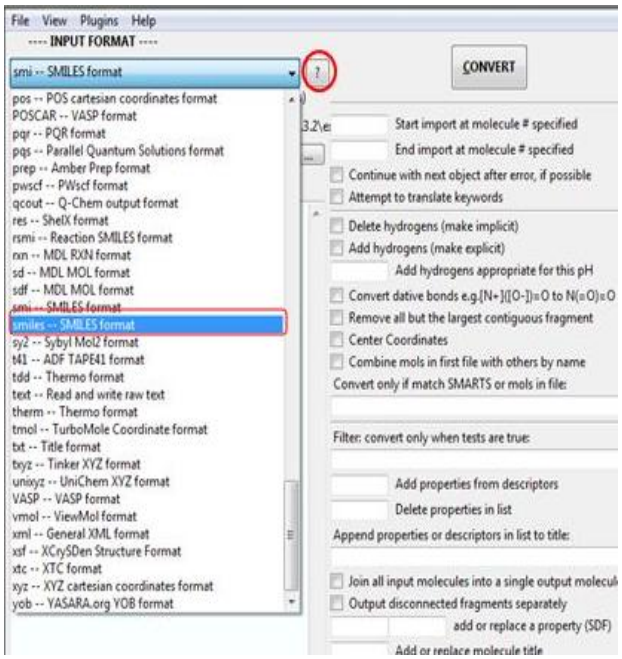

Fig. 7 Selecting a SMILES format from the list [14]

Select the format of interest and click on the "?" button on the left side of selected format to know more about the file formats. Figure 7 shows an example with SMILES format.

Open Babel manages around 110 chemical file formats with which one can read 82 formats and write 85 formats. It provide ready to use programs and a software package necessary for cheminformatics lab including file formats like SMILES, MOL, MOL2 as input and computational chemistry packages like GAMESS, Gaussian, MOPAC etc. as output. The OpenBabel software package includes various other tools which include the following: [14]

1. **Babel:** To interconvert between many file formats used in molecular modeling and computational chemistry.

2. **Obchiral:** It prints the chirality (molecules) information for the molecules. A chiral molecule is a type of molecule that has a non-superposable mirror image.

3. **Obconformer:** It is used to generate low-energy conformers, a form of stereoisomerism in which the isomers can be interconverted exclusively by rotations about formally single bonds Such isomers also referred as conformational isomers or conformers and, specifically, as rotamers.

4. **Obenergy:** It is used to calculate energy of the molecule.

5. **Obfit:** It is used to superimpose two molecules by using SMARTS.

6. **Obgen:** It is used to generate 3D coordinates for the molecule.

7. **Obgrep:** It is used to find molecules inside multi-molecule database files using SMARTS.

8. **Obminimize:** It is used to optimize, minimize and calculate the energy of the molecule.

9. **Obprobe:** It is used in docking experiments to calculate MMFF94 energy by creating electrostatic probe grid.

10. **Obprop:** It is used to print the standard properties of molecules in a file.

11. **Obrotamer:** It is used to create random rotational isomers based on their rotating dihedral angles.

12. **Obrotate:** It rotates the dihedral angle of molecules with their matching SMARTS pattern.

The most common use of Open Babel is to convert chemical file formats. [15] This is illustrated by an example that converts a PDB file to MOL format. First a folder called 'work' is created on the Desktop. The PDB file for insulin (4ins) is downloaded from the Protein Data Bank and it is saved in the 'work' folder. Now set the input file format to PDB, the input filename to the downloaded PDB file, set the output file format to MOL, the output filename to file:*4ins.mol* in the 'work' folder and click convert as shown in the figure 8. [15]

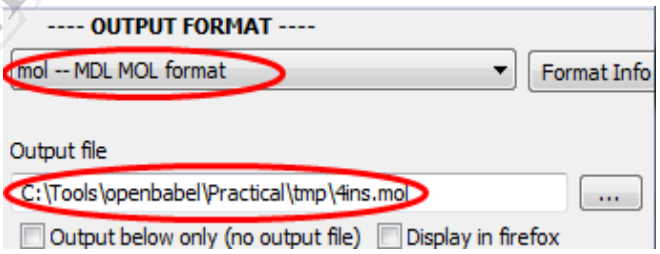

Fig. 8 conversion of PDB to MOL [15]

It is also possible to paste the contents of a chemical file format into the input box as shown in the figure 9. The results of the conversion will be obtained in the output box. In this case there is no need to use input and output files. This is illustrated with the SMILES format i.e. how stereochemistry is handled by SMILES. Choose the SMILES format as the input format and Tick the box Input below to ignore input file and copy and paste the SMILES strings (and molecule titles) shown in the figure 10 into the input box. Choose the SMILES format as the output format. And also tick the box for Output below only and Display in Firefox as shown in the figure 11 and click convert. [14]

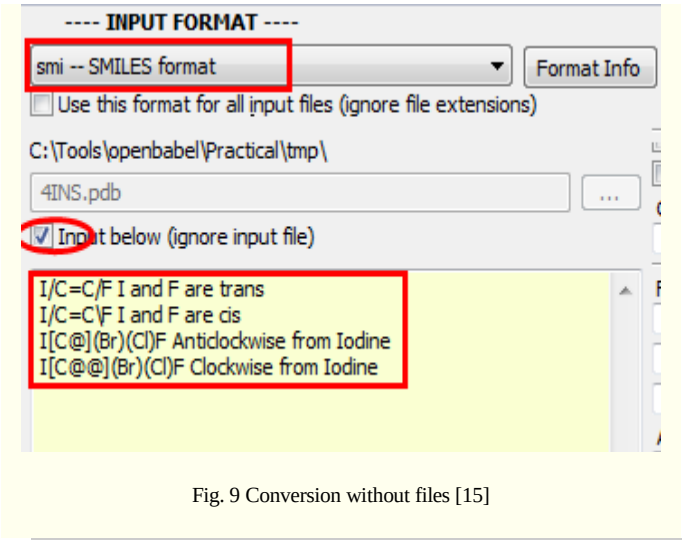

Fig. 9 Conversion without files [15]

- I/C=C/F I and F are trans
- I/C=C\F I and F are cis
- I[C@](Br)(Cl)F Anticlockwise from Iodine

I[C@@](Br)(Cl)F Clockwise from Iodine

Fig. 10 SMILE strings [14]

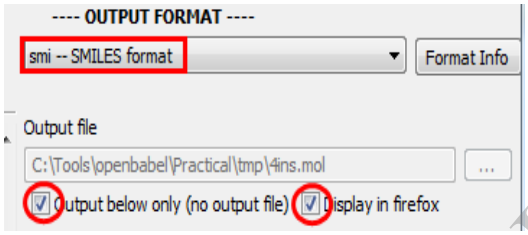

Fig. 11 Output format [15]

Few examples of file interconversion using Babel (command line utility) are given below:

1. Convert SDF file to MOL2 file.
   prompt> babel −in foo . sdf −out bar . mol2

2. Convert gzipped SDF file to SMILES.
   prompt> babel −in foo . sdf . gz −out bar . smi

3. Convert SDF file to MOL2 file using shortcut"keyless" syntax.
   prompt> babel foo . sdf . gz bar . smi

4. Convert gzipped SDF stream from stdin to SMILES.
   prompt> cat foo . sdf . gz | babel −in . sdf . gz −out bar . smi

5. Convert gzipped SDF file to OEBinary multiconformer file, where consecutive molecules may be interpreted as conformers of the same molecule.
   prompt> babel −in mongodb . sdf . gz −out bar . oeb . gz –mc

6. Convert gzipped SDF file to OEBinary multiconformer file, where consecutive molecules may be interpreted as conformers of the same molecule, but different stereo isomers are considered different molecules.
   prompt> babel −in mongodb . sdf . gz −out bar . oeb . gz −mc_isomer

B. *Molconverter*

MolConverter [2] is a command line program in JChem [16] and Marvin Beans [17] that converts between various file types.

molconvert [options] outformat[:exportoptions] [files...]

The out format argument must be one of the following strings:

TABLE 1: OUTFORMAT ARGUMENT

(document formats)	mrv
(molecule file formats)	mol, rgf, sdf, rdf, csmol, csrgf, cssdf, csrdf, cml, smiles, cxsmiles, abbrevgroup, peptide, sybyl, mol2, pdb, xyz, inchi, name, cdx, cdxml, skc
(graphics formats)	jpeg, msbmp, png, pov, ppm, svg, emf
(compression& encoding)	gzip, base64

molconvert [options] query-encoding [files...] can also be used to query the automatically detected encodings of the specified molecule files. Molconvert is able to recognize the name of compounds from files having doc, docx, ppt, pptx, xls, odt, pdf, xml, html or txt format and convert it to any of the above mentioned output formats.

Few of the Options in the command line argument are given below:

-o file Write output to specified file instead of standard output

-m Produce multiple output files

-e charset Set the input character encoding. The encoding must be supported by Java.

-e [in]..[out] Set the input (in) and/or output (out) character encodings. Examples: UTF-8, ASCII, Cp1250 (Windows Eastern European), Cp1252 (Windows Latin 1), ms932 (Windows Japanese).

-s string Read molecule from specified SMILES, SMARTS or peptide string (try to recognize its format)

-s string{format:options} Read molecule from the string in the specified format (can be omitted), using the specified import options (can be omitted)

--smiles string Read molecule from specified SMILES string

--smarts string Read molecule from specified SMARTS string

--peptide string Read molecule from specified peptide string
-g Continue with next molecule on error (default: exit on error)

-Y Remove explicit H atoms

-I <range> process input molecules with molecule index (1-based) falling into the specified range (e.g. 5-8,15 refers to molecules 5,6,7,8,15)

-U fuse input molecules and output the union

-R <file>[:<range>] fuse fragments to input molecule(s) from file with specified mol index range range syntax: "-5,10-20,25,26,38-" (e.g. -R frags.mrv:20-)

-R<i> <file>[:<range>] fuse R<i> definitionmembers to input molecule(s) from file in specified index range (e.g. -R1 rdef1.mrv:5-8,19)

-R<i>:<1|2> <file>[:<range>] fuse R<i> definition members to input molecule(s) from file in specified index range, filter molecules having 1 (2, resp.) attachment points (e.g. -R1:2 rdef1.mrv:-3,8-10)

Few examples are given below:

1. Printing the SMILES string of a molecule in a molfile:
   molconvert smiles caffeine.mol
2. Dearomatizing an aromatic molecule:
   molconvert smiles:-a -s "c1ccccc1"
3. Aromatizing a molecule:
   molconvert smiles:a -s "C1=CC=CC=C1"
4. Aromatizing a molecule using the basic algorithm:
   Molconvertsmiles:a_bas -s "CN1C=NC2=C1C(=O)N(C)C(=O)N2C"
5. Converting a SMILES file to MDL Molfile:
   molconvert mol caffeine.smiles -o caffeine.mol

C. CnCTranslate

CncTranslate [3] is a central component of ChemNavigator's software tool kit for cheminformatics data management and analysis. It is used to perform file conversion among many chemical structure file formats. CncTranslate offers a full command line interface and may be used as a cheminformatics utility program within other applications. Different options that are available in CncTranslate are described below.

CncTranslate Basic

Basic includes tools for file conversion between standard chemical structure file formats and graphical display formats common to the cheminformatics industry.

CncTranslate Advanced

The Advanced option includes rules based structure normalization and validation of structure representation.

CncTranslate 3D

The 3D option allows approximate 3D coordinates to be calculated for 2D structures.

TABLE2: CNCTRANSLATE BASIC SUPPORTED FILE FORMATS [3]

Format Type	Input	Output
MDL SD Files	Yes	Yes
SMILES	Yes	Yes
Tripos SYBYL Line Notation (SLN)	Yes	Yes
JME Format (Novartis Java Editor)	Yes	Yes
Tripos MOL2	Yes	Yes
PNG Graphics	N/A	Yes
2D Coordinate Generation	N/A	Yes
2D bitmap fingerprint generation	N/A	Yes
Protein Databank Format (PDB)	Yes	Yes

CncTranslate Drug-Like

The Drug-Like option permits the calculation of commonly used molecular descriptors including: molecular weight, counts of hydrogen-bond donors & acceptors, the number or rotatable bonds. This tool also permits the definition and counting of custom molecular fragments.

CncTranslate Search

The Search option enables searching of chemical structure lists in SLN format. Search enables the searching functionality for data items and substructure chemical searching. Search may be applied to large files and is commonly used as a flat-file based cheminformatics system.

CncTranslate RGroups

The RGroups option enables the R-group decomposition functions. These functions locate a core substructure within each chemical structure in an input series, and determine the R-groups that are attached to it for each structure in the input set. The input is a core substructure pattern and a structure file containing the series of chemical structures for analysis. The output includes each input structure with the identified core marked and any R-groups as molecular fragments.

IV. CONCLUSION

File format Conversion tools are very essential for the researchers to conduct several experiments in different cheminformatics applications. In this paper we presented a review of some of the chemical file formats and a few inter-

conversion tools such as Open Babel, molconvert and Cnc translate.

REFERENCES

[1] Morley, Tim Vandermeersch and Geoffrey R Hutchison, 'OpenBabel:An open chemical toolbox', O'Boyle et al. Journal of Cheminformatics 2011, 3:33

[2] 'Molconverter' Available on http://www.chemaxon.com/marvin/help/applications/molconvert.html accessed on November 19, 2013.

[3] 'Cnctranslate' available on http://www.chemnavigator.com /cnc/products/ cnctranslate.asp accessed on December 1, 2013

[4] Antony J. Williams, 'Public Chemical Compound Databases',http://www.academia.edu/3062783/Public_chemical_ compound_databases , accessed on November 19,2013.

[5] Kavitha C.R, Dr. Mahalakshmi, 'Chemical Databases: A Brief Walk', International Journal of Emerging Technology and Advanced Engineering (IJETAE), ISSN 2250-2459,ISO 9001:2008Certified Journal Volume 3 Issue 8 August 2013

[6] David Weininger, 'SMILES, a Chemical Language and Information System. 1. Introduction Methodology and Encoding Rules', J. Chem. Inj Comput. Sci. 1988, 28, 31-36

[7] Weerapong Phadungsukanan, Markus Kraft, Joe A Townsend and Peter Murray-Rust, 'The semantics of chemical Markup Language (CML) for computational chemistry: COMPCHEM', Journal of cheminformatics 2012, 4:15, http://WWW.Jcheminf.com/content/4/1/15

[8] Ápplications of XML', Available at http://cs.au.dk/~amoeller/XML/xml/applications.html accessed on December 10, 2013.

[9] 'Structure Data Format' Available at http://www.ccl.net/cca/software/PERL/SDF/ accessed on December 11, 2013.

[10] Dalby, A., J.G. Nourse, W.D. Hounshell, A.K.I. Gushurst, D.L. Grier, B.A. Leland, J. Laufer (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited, J. Chem. Inf. Comput. Sci. 32:244-255.

[11] 'A sample SDF file', Available at http://www.epa.gov/ncct/dsstox/MoreonSDF.html#Sample accessed on January 10, 2014

[12] 'Protein Data Bank', Available at http://en.wikipedia.org/wiki /Protein_Data_Bank_ (file_format) accessed on December 12, 2013.

[13] 'XYZ format' available at http://en.wikipedia.org/wiki/xyz_file_ format accessed on December 12, 2013.

[14] 'Converting chemical file formats', available at http://amrita.vlab.co.in/?sub=3&brch=275&sim=1499&cnt=1 accessed on December 7, 2013

[15] 'Converting chemical file formats', Available at https://open-babel.readthedocs.org/en/ latest/ GUITutorial/ Conversion.html accessed on December 21, 2013

[16] 'Jchem' , Available at http://www.chemaxon.com/ jchem/intro/ index.html, accessed on December 25, 2013

[17] 'Marvin Beans', Available at http://www.chemaxon. com/ marvin/help/developer/beans/beanfaq.html, accessed on December 25, 2013