

Section III

Cheminformatics - Basics: 2D and 3D Molecular Structures. Chemical Structure Drawing Packages

Dan C. Fara, Andrei Leitão and Tudor I. Oprea

1. Molecular representation

The familiar two-dimensional (2D) chemical structure diagram could be considered to be the “natural language” of the chemist, and in everyday communication chemists almost invariably resort to it as their preferred form of structure representation, whether printed in a learned journal, pasted into a word-processes document, or scribbled on a whiteboard or the back of an envelope. Whilst they have found their way into computer-based information systems, their appearance is generally restricted to the user interface, and other types of representation are used internally. Structure diagrams normally show the connectivity of a molecule, and might also show its stereochemistry by means of “up” and “down” bond symbols, chirality designators, etc. For example, the 2D representation can show if a compound of formula C_2H_6O is an alcohol or aldehyde – see Figure 1. Although ethanol and ethanal have the same 1D notation (C_2H_6O), they are different compounds from distinct chemical classes.

An obvious choice as a description of the steric aspects of a molecule is the explicit recording of a triplet of x-, y-, and z-coordinates for each atom. This method is generally called the 3D description of a molecule, and is appropriate for the storage and manipulation of crystal structures and other models which directly relate to a specific conformation of a molecule. As can be seen, for ethanol and ethanal an infinite number of conformations can be obtained just by changing the positions of the atoms.

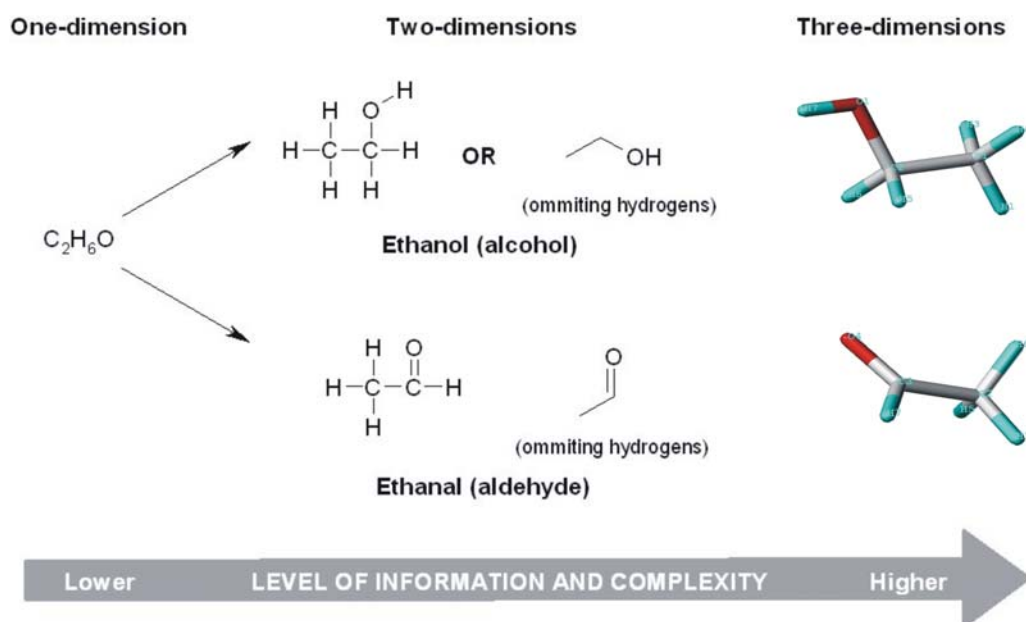


Figure 1. Different levels of molecular representation

2. Molecular editors - Drawing chemical structures

Many “molecule editor” programs are available to enable the user to draw a structure diagram on a computer screen by using a “mouse” or other pointing device; an internal (usually connection table) representation of the structure is built at the same time. The 2D coordinates of each atom on the screen can be retained for re-display of the diagram when required, although programs are also available to generate new coordinates if these are not available. Programs have also been developed, both in commercial and academic environments to perform “optical character recognition” in printed structure diagrams.

Those molecular editors are very similar, but the options to draw compounds can change from one to another. The JME (Java Molecular Editor) runs directly in the web browser, and is free for anyone to use, i.e. at the Biocomputing home page (<http://biocomp.health.unm.edu/jme/index.shtml>). To use ISIS Draw, ChemDraw, you must own a license for the software and have it installed on your computer. The ACD/ChemSketch is free for academic use. Another free molecular editor is MarvinSketch application provided by [ChemAxon](#). The picture below shows the interface of IsisDraw and ACD/ChemSketch (Figure 2).

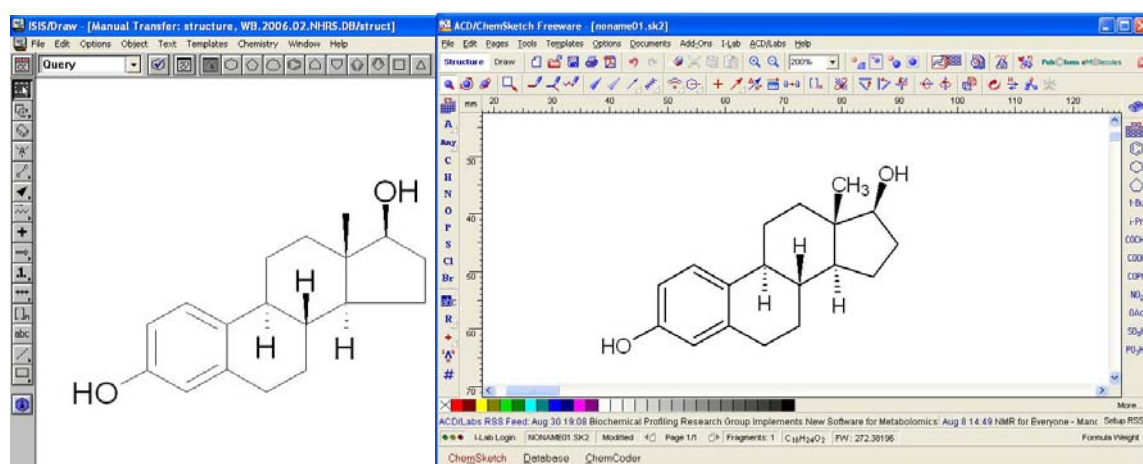


Figure 2. IsisDraw and ChemSketch interfaces with estradiol

- Go to the <http://biocomp.health.unm.edu/jme/index.shtml> and open the JME application.
- Draw the estradiol compound using the options provided around the screen (Figure 3).
- Click on *Get smiles* button and you will see a string of characters that represent the compound (you can also see the SMILES code in another window when you click on the smiling yellow face at the top left corner of the drawing sheet, see the circle in Figure 3). Other types of representation can be shown by clicking on *Get JME file* and *Get molecule* buttons.

Figure 3. JME editor at the Biocomputing home page

Estradiol is drawn and other information is provided the other fields. Please see the text.

3. SMILES and SMARTS

3.1 SMILES (Simplified Molecular Input Line Entry System)

SMILES is a *line notation* (a typographical method using printable characters) for entering and representing molecules and reactions or, in other words, a chemical nomenclature which represents valence models of molecules and reactions. It was developed at [Daylight](http://www.daylight.com) (Chemical Information Systems, Inc.) and is based on only six rules:

- ✓ Atoms are represented by atomic symbols
- ✓ Double bonds are represented by '=', triple bonds by '#'
- ✓ Branches are indicated by matching parentheses
- ✓ Ring closures are indicated by matching pairs of digits
- ✓ Disconnections are indicated by a '.' (period)
- ✓ Reactions are written as reactants>agents>products

Some examples are shown in Figure 4. One advantage of using SMILES is the possibility of storing millions of compounds using only a small fraction of the actual size of the same 3D library of compounds.

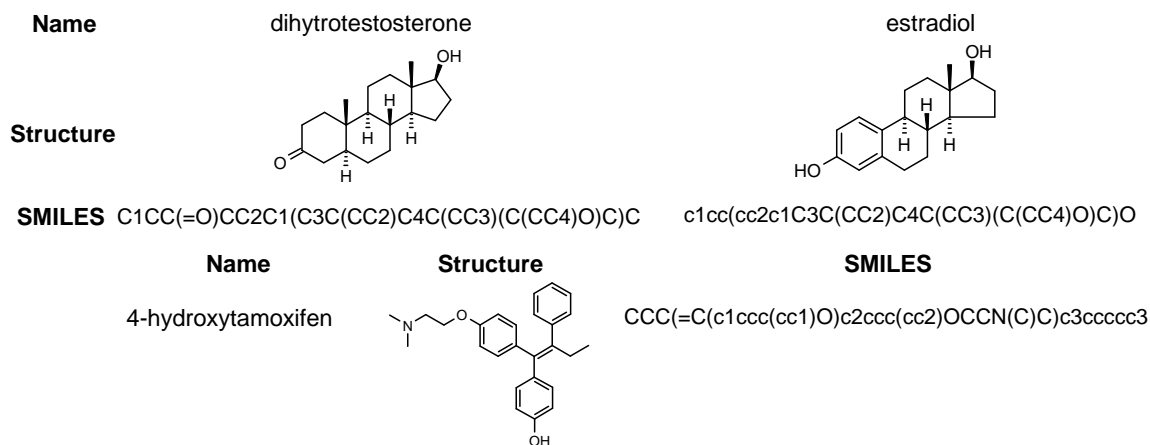


Figure 4. SMILES representation of some nuclear receptor ligands

The detailed theory for SMILES can be found [here](#). Also, a comprehensive tutorial, examples, and SMILES practice can be accessed from the Daylight [website](#).

3.2 SMARTS (SMiles ARbitrary Target Specification)

SMARTS is a language for specifying substructural patterns in molecules, which has been also developed at [Daylight](#). The SMARTS line notation is expressive and allows extremely precise and transparent substructural specification and atom typing.

The following are stated on the [Daylight Theory: SMARTS](#) webpage:

“*Substructure searching*, the process of finding a particular pattern (subgraph) in a molecule (graph), is one of the most important tasks for computers in chemistry. It is used in virtually every application that employs a digital representation of a molecule, including depiction (to highlight a particular functional group), drug design (searching a database for similar structures and activity), analytical chemistry (looking for previously-characterized structures and comparing their data to that of an unknown), and a host of other problems.

SMARTS is a language that allows you to specify substructures using rules that are straightforward extensions of SMILES. For example, to search a database for phenol-containing structures, one would use the SMARTS string **[OH]c1ccccc1**, which should be familiar to those acquainted with SMILES. In fact, almost all SMILES specifications are valid SMARTS targets. Using SMARTS, flexible and efficient substructure-search specifications can be made in terms that are meaningful to chemists.

In the SMILES language, there are two fundamental types of symbols: *atoms* and *bonds*. Using these SMILES symbols, one can specify a molecule's graph (its "nodes" and "edges") and assign "labels" to the components of the graph (that is, say what type of atom each node represents, and what type of bond each edge represents).

The same is true in SMARTS: One uses atomic and bond symbols to specify a graph. However, in SMARTS the labels for the graph's nodes and edges (its "atoms" and "bonds") are extended to include "logical operators" and special atomic and bond symbols; these allow SMARTS atoms and bonds to be more general. For example, the SMARTS atomic symbol **[C,N]** is an atom that can be aliphatic **C** or aliphatic **N**; the SMARTS bond symbol **~** (tilde) matches any bond.”

The detailed theory for SMARTS can be found [here](#). Also, a comprehensive tutorial, examples, and SMARTS practice can be accessed from the Daylight [website](#).

3.3 SubStructure Searching (SSS) using SMILES/SMARTS

[Roadrunner](#) is a chemical database application developed and supported by the Informatics Core of the [New Mexico Molecular Libraries Screening Center](#) at the University of New Mexico [Health Sciences Center](#).

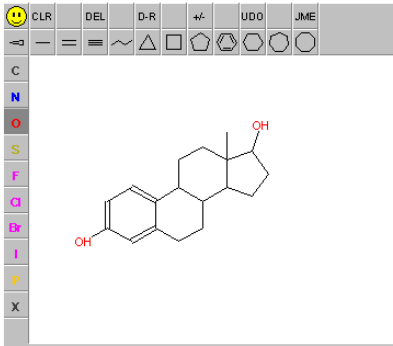
- Go to [Roadrunner](#)
- Create an account and login
- Go to *Compound Search > Substructure/Property* and draw the structure in the Java Molecular Editor, or enter the SMILES (see above Figure 4) or SMARTS string directly. Click "Smi2Sma" to convert SMILES to SMARTS. As can be seen in Figure 5 (bottom), the search will be performed on the MLSMR (222107 compounds) library
- Click *Search*

Fetch Substance By ID:

Compound Search Form

Use this form to query for compounds. All search criteria (Substructure, Properties, Name and/or Library) will be ANDed together. In other words, query results will be the intersection of all query criteria. "Search" buttons will do to submit your search.

Search by Substructure:
 Draw a structure in the JME Molecular Editor, or enter a SMILES or SMARTS string directly. Click "Smi2Sma" to convert a SMILES to a SMARTS (note however, that this will produce a very generic SMARTS that may match more than you intend).



JME Editor courtesy of Peter Ertl, Novartis

SMILES:

Smi2Sma
Clear

SMARTS:

Clear

Search

Search by Property:
 Enter a minimum and maximum value for from one to four properties (criteria will be ANDed together):

<input type="text"/>	Molecular Weight <input type="button" value="v"/>	≤	<input type="text"/>
<input type="text"/>	H-Bond Acceptors <input type="button" value="v"/>	≤	<input type="text"/>
<input type="text"/>	Polar Surface Area <input type="button" value="v"/>	≤	<input type="text"/>
<input type="text"/>	Simple Molecular Complexity Metric <input type="button" value="v"/>	≤	<input type="text"/>

Search
Clear

Search by Name:
 Enter a Search Term (Eg. trimethylpurine). Select "Include Synonyms?" to search by MLS#.

☒ Include Synonyms?

Search
Clear

Constrain Search To A Particular Library:
 Note: to retrieve an entire library, go to the "Browse" option in the navigation menu.

MLSMR (222107 Substances)

Figure 5

As can be seen from Figure 6, the substructure search for estradiol has found 17 compounds in the MLSMR library.

The New Mexico Molecular Libraries Screening Center Chemical Database Application

- ▶ Home Page
- ▶ Documentation
- ▶ Browse
- ▶ Compound Search
 - Substructure/Property
 - Similarity
 - Batch Similarity
- ▶ Assays
 - AssayDef Search
 - Create AssayDef
 - Assay Upload
 - View Uploads
 - Assay Aggregator
 - View Aggregators
- ▶ Plates
 - Plate Search
 - Plate Map List
- ▶ Post-HTS Analysis
 - New Analysis
 - View Analyses
- ▶ Admin
 - Account Settings

Fetch By ID:

Compound Search Results

Found 17 Compounds

[Export Results](#)

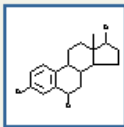
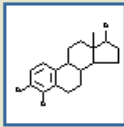
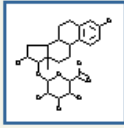
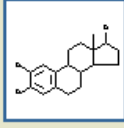
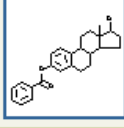
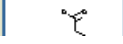
#	Structure	ID	SMILES
1		20098	<chem>CC12CCC3c4ccc(cc4C(CC3C1CCC2O)O)O</chem>
2		20111	<chem>CC12CCC3c4ccc(c(c4CCC3C1CCC2O)O)O</chem>
3		20118	<chem>CC12CCC3c4ccc(cc4CCC3C1CC(C2OC5C(C(C(C(O5)C(=O)O)O)O)O)O)O</chem>
4		63882	<chem>CC12CCC3c4cc(c(c4CCC3C1CCC2O)O)O</chem>
5		88025	<chem>CC12CCC3c4ccc(cc4CCC3C1CCC2O)OC(=O)c5ccccc5</chem>
			

Figure 6

These results can be exported and saved on your computer. By clicking either on “structure” or “ID” the “Compound Report” will be opened. This report contains various data about the selected compound, i.e. ID, name, SMILES, molecular weight, 31 calculated molecular descriptors (properties), and links to PubChem and bioactivity – see Figure 7.

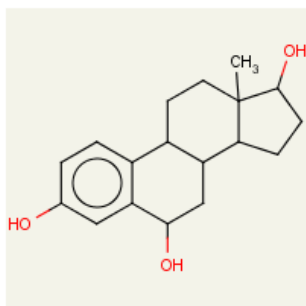


- ▶ Home Page
- ▶ Documentation
- ▶ Browse
- ▶ Compound Search
 - Substructure/Property
 - Similarity
 - Batch Similarity
- ▶ Assays
 - AssayDef Search
 - Create AssayDef
 - Assay Upload
 - View Uploads
 - Assay Aggregator
 - View Aggregators
- ▶ Plates
 - Plate Search
 - Plate Map List
- ▶ Post-HTS Analysis
 - New Analysis
 - View Analyses
- ▶ Admin
 - Account Settings

Fetch By ID:

Compound Report

Structure



Powered by ChemAxon

Links

Substance: [197998](#)PubChem Compound: [5284655](#)[Similar Compounds](#)

Bioactivity

Bioactivity: [31 Results](#)

Data

ID: 20098

Name:

13-methyl-6,7,8,9,11,12,14,15,16,17-decahydrocyclopenta[a]phenanthrene-3,6,17-triol

Molecular Formula: C₁₈H₂₄O₃

Molecular Weight: 288.42

Canonical SMILES: CC12CCC3c4ccc(cc4C(CC3C1CCC2O)O)O

Isomeric SMILES:

C[C@]12CC[C@H]3[C@H]([C@@H]1CC[C@@H]2O)C[C@H](C4=C3C=CC(=C4)O)O

Property	Value
Aliphatic Atoms	1
ALogP	2.54
ALogS	-3.51
Andrews Binding Energy	6.7
Barone and Channon Complexity	408
Carbons	18
Charge	0
Charge Count	0
Chiral Centers	0
CLogP	2.147
Functional Groups	3

Property	Value
Halide Fraction	0.00
H-Bond Acceptors	3
H-Bond Donors	3
Heavy Atoms	21
Het:C Ratio	0.17
Heteroatoms	3
Lipinski Acceptors	3
Lipinski Donors	3
Max. Ring Size	6
Min. Ring Size	5

Property	Value
Number of Rings	4
Polar Surface Area	60.69
PSA w/ S and P	60.69
Rigid bonds	6
Ring Degrees of Freedom	7
Ro5 Violations	0
Rotatable Bonds	0
Simple Molecular Complexity Metric	60.141
Solubility	moderate
Unbranched Atoms	0

Figure 7

- Repeat the steps for the other two compounds: dihydrotestosterone and 4-hydroxytamoxifen
- Compare the results

How many compounds were found and how many of them were found in each of the query (common)?