Chemoinformatics: chemical data retrieval, mining and storage

Samuel Egieyeh

Outline

- Chemoinformatics
- Data types
- Chemical structures representations:
- The database: Structure Data Format (SDF)
- The Tools:
  - KNIME and DataWarrior
- Case study:
  - Creating a database of molecular scaffolds from compounds with 70% similarity to current antimalarial drugs.
- Conclusions: Advantages and disadvantages

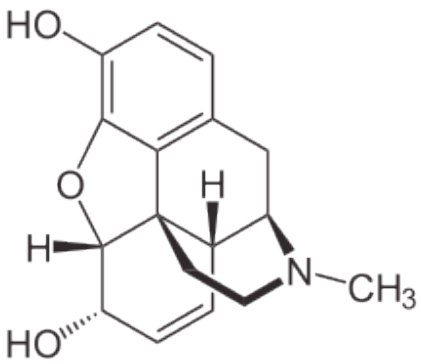
Chemoinformatics

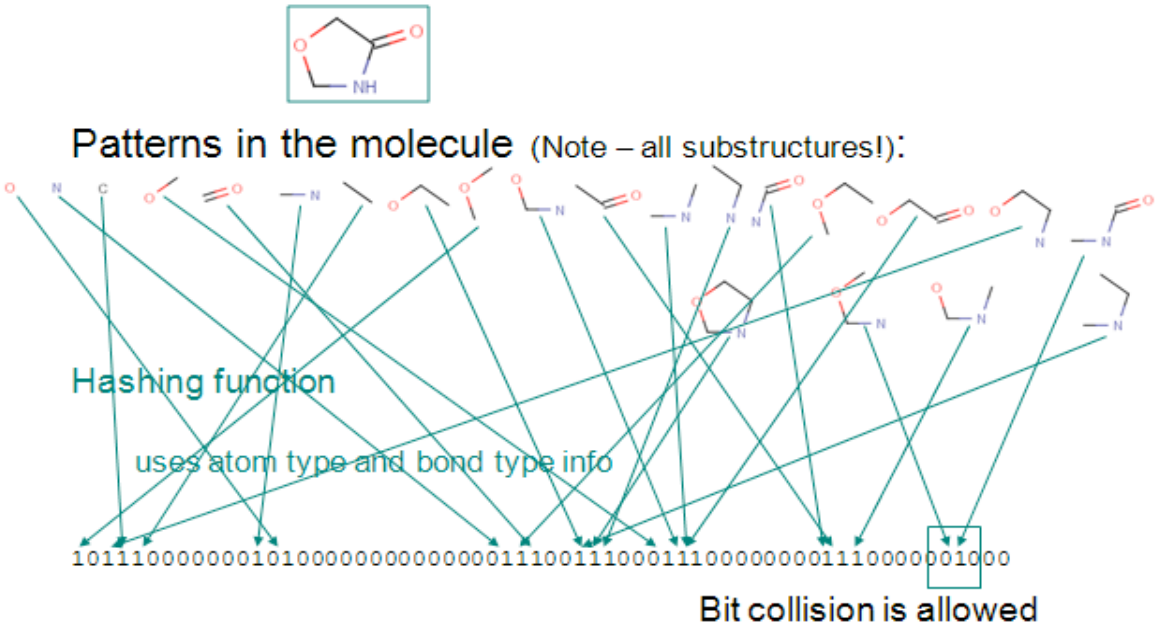
- “Chemoinformatics is the mixing of information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.” F.K. Brown 1998

- Primary application
  - storage
  - indexing
  - search of information relating to compounds.

Data types

- Chemical structures
- Bit-type fingerprints
- Numeric molecular descriptors
- Bioactivity data





Chemical structures representations

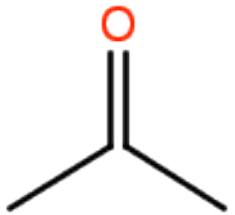
- Canonicalized **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification (SMILES).
  - Unique string that can be used as a universal identifier for a specific chemical structure.
  - Example: SMILES for Casamembrol A
  - [H][C@@]12C[C@H](OC(=O)C(C)CC)C=C3C(OC(C)=O)OC(OC(C)=O)[C@@]13[C@@H](O)C[C@@H](C)[C@@]2(C)C\C=C(/C)C=C

- The IUPAC International Chemical Identifier (InChI™)
  - A textual unique identifier for chemical substances
  - Facilitate the search for such information in databases and on the web.
  - Example: CH_3CH_2OH
  - InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3 (standard InChI)

Chemical structures representations

- Mol file format
  - File format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.
  - Example for acetone:

```
Acetone
creator program

  4  3  0  0  0  0  0  0  0  0999 V2000
   -6.6000    2.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -6.6000    4.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -7.8990    2.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -5.3010    2.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
M  END
```



The chemical database

- Structure Data Format (*MDL SDF* or *SDfile*)
- Developed by Molecular Design Limited
- Handle a list of molecular structures with associated properties[1].
- SDF toolkit is used to read and parse SDFs, filter, and add/ remove properties.
- The SDF can be visualized with many other programs.

1. Dalby et al., *J. Chem. Inf. Comput. Sci.* 1992, **32**: 244-255.

The chemical database

Structure Data Format

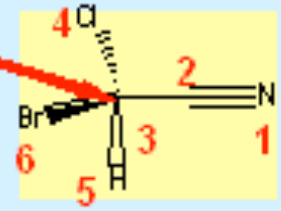
Header note

x,y,z atomic coordinates; Atomic symbols;

Connection table; Bond type

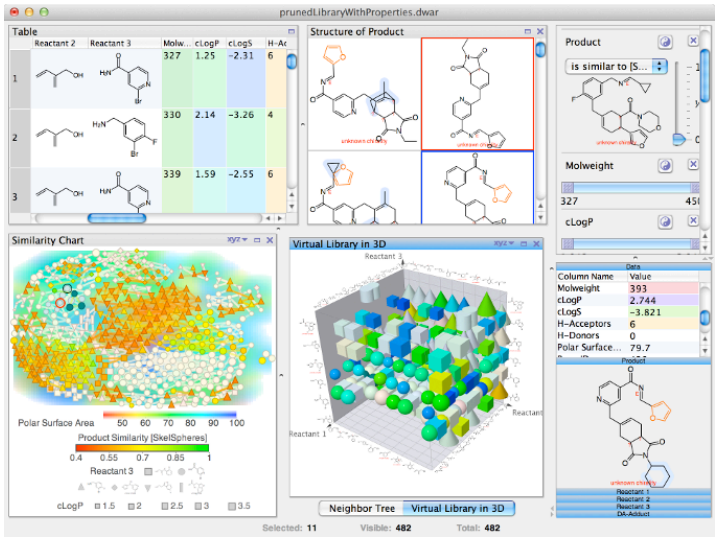
Chiral center

```
csChFmd70/11150310042D

  6  5  0  0  0  0  0  0  0  0999 V2000
    3.7195    1.2124    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.3195    1.2124    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.3523    1.2124    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
    0.9900    2.5647    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    1.5748    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.6523    0.0000    0.0000 Br  0  0  0  0  0  0  0  0  0  0  0  0
  2  1  3  0  0  0  0
  3  2  1  0  0  0  0
  3  4  1  1  0  0  0
  3  5  1  6  0  0  0
  3  6  1  0  0  0  0
M  END
>  <Formula> (2)
C2HBrClN

>  <MolWeight> (2)
154.3936

>  <SMILES> (2)
N#C[C@@](Cl)([H])Br

>  <ChemName> (2)
Bromochloroacetonitrile

$$$$
```



Tools

- KNME (Konstanz Information Miner)
  - Modern data analytics platform that allows one to perform:
  - Sophisticated statistics
  - Data mining
  - Analyze trends
  - Predict potential results.
- DataWarrior
  - Combines dynamic graphical views and interactive row filtering with chemical intelligence.



Case study

- Creating a database of molecular scaffolds from compounds with substructural match to current antimalarial drugs.