MALTAOMICS SUMMER SCHOOL

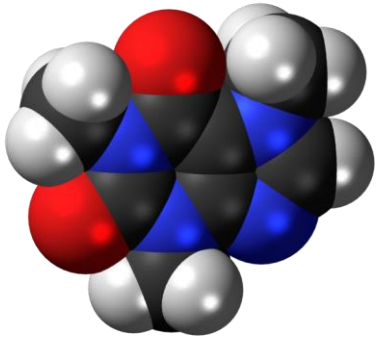
INTRODUCTION TO COMPUTER-AIDED DRUG DESIGN

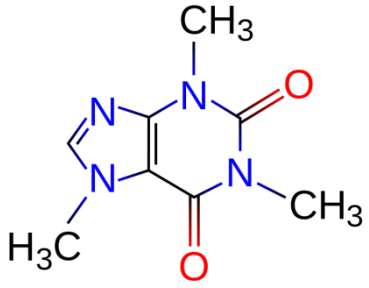
DAY 4 – 10:40-11:30

Dr Jean-Paul Ebejer
jean.p.ebejer@um.edu.mt

WHAT DOES A DRUG LOOK LIKE?

- Various ways how to represent the same compound (molecule)
- An Example: **Caffeine** (or 1,3,7-trimethyl-3,7-dihydro-1H-purine-2,6-dione if you prefer)
- Chemical formula (1D): $C_8H_{10}N_4O_2$
- SMILES String (1D): Cn1cnc2c1c(=O)n(c(=O)n2C)C
- 2D Representation in 3D (Graph)





HOW DO WE REPRESENT MOLECULES IN COMPUTER SYSTEMS?

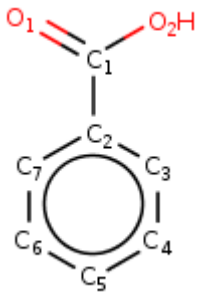
- Any representation of the molecules (which is not the molecule itself) is called a “Descriptor”
  - As the name implies, describes a molecule
- We can represent molecules as a vector
  - E.g. <Molecular Mass, Number of +ve charges, Volume, Log P>
- Countless different representations exist; computed on 1D (properties), 2D (topology; adjacency information), and 3D (geometry) properties of the molecule
- How do we represent connectivity of atoms?

MOLECULAR GRAPHS

- 2D information of a molecule represented by a graph (mathematical notation)
- A graph is made of:
  - A set of vertices which represent atoms ($v \in V$)
  - A set of edges connecting two nodes which represent edges ($e \in E$)
  - Therefore a graph is a tuple G = (V, E)
- Edges have no direction; so graph is said to be undirected
- Nodes and Edges contain information such as atom types, bond order, etc.
- Allows us to use graph theory on molecules; substructure searching etc.
- Only about connectivity

REPRESENTING A MOLECULAR GRAPH

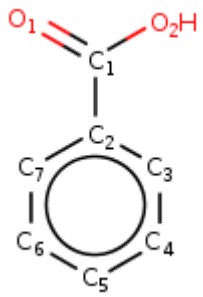
- Using Matrices!
- A molecule with n heavy atoms, may be represented with an n x n matrix
  - Hydrogens often considered “implicit” and omitted
- Adjacency Matrix (indicates which atoms are bonded)



	O1	O2	C1	C2	C3	C4	C5	C6	C7
O1	0	0	1	0	0	0	0	0	0
O2	0	0	1	0	0	0	0	0	0
C1	1	1	0	1	0	0	0	0	0
C2	0	0	1	0	1	0	0	0	1
C3	0	0	0	1	0	1	0	0	0
C4	0	0	0	0	1	0	1	0	0
C5	0	0	0	0	0	1	0	1	0
C6	0	0	0	0	0	0	1	0	1
C7	0	0	0	1	0	0	0	1	0

Can you think of some optimizations?

OPTIMIZATIONS



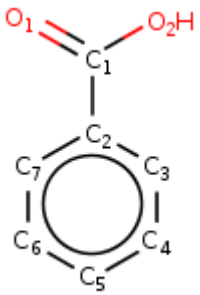
	O1	O2	C1	C2	C3	C4	C5	C6	C7
O1	0	0	1	0	0	0	0	0	0
O2	0	0	1	0	0	0	0	0	0
C1	1	1	0	1	0	0	0	0	0
C2	0	0	1	0	1	0	0	0	1
C3	0	0	0	1	0	1	0	0	0
C4	0	0	0	0	1	0	1	0	0
C5	0	0	0	0	0	1	0	1	0
C6	0	0	0	0	0	0	1	0	1
C7	0	0	0	1	0	0	0	1	0

	O1	O2	C1	C2	C3	C4	C5	C6	C7
O1			1						
O2			1						
C1				1					
C2					1				1
C3						1			
C4							1		
C5								1	
C6									1
C7									

- Matrix is symmetrical
- No need to store 0s
- Each row (or column) must have at least a 1 (otherwise it is not connected to anything!)

DISTANCE MATRIX

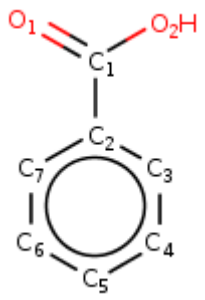
- Encodes distance between atoms
  - Number of bonds between atoms using shortest path
  - Could be 3D distance (in Å)



	O1	O2	C1	C2	C3	C4	C5	C6	C7
O1	0	2	1	2	3	4	5	4	3
O2	2	0	1	2	3	4	5	4	3
C1	1	1	0	1	2	3	4	3	2
C2	2	2	1	0	1	2	3	2	1
C3	3	3	2	1	0	1	2	3	2
C4	4	4	3	2	1	0	1	2	3
C5	5	5	4	3	2	1	0	1	2
C6	4	4	3	2	3	2	1	0	1
C7	3	3	2	1	2	3	2	1	0

BOND MATRIX

- Indicates which atoms are bonded and the corresponding bond orders



	O1	O2	C1	C2	C3	C4	C5	C6	C7
O1	0	0	2	0	0	0	0	0	0
O2	0	0	1	0	0	0	0	0	0
C1	2	1	0	1	0	0	0	0	0
C2	0	0	1	0	1	0	0	0	1
C3	0	0	0	1	0	1	0	0	0
C4	0	0	0	0	1	0	1	0	0
C5	0	0	0	0	0	1	0	1	0
C6	0	0	0	0	0	0	1	0	1
C7	0	0	0	1	0	0	0	1	0

WHY USE MATRICES?

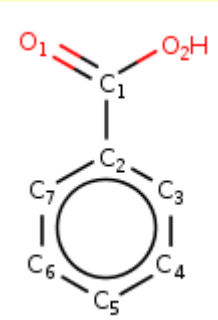
- Advantages
  - Use of Matrix Algebra for comparison etc.
  - Complete representation of graph
- Disadvantages
  - Quadratic size (n^2)
  - Sparsely populated

SOLUTION

- Use of adjacency list (a.k.a. the connection table)
- Linear size
- E.g. MDL SDF (Structure Data File)

```
Benzoic Acid
  Mrv1903 04071910542D

  9  9  0  0  0  0            999 V2000
    0.3348    5.3339    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3796    4.9214    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3796    4.0964    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.3348    3.6839    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0493    4.0964    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0493    4.9214    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.3348    2.8589    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0493    2.4464    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3796    2.4464    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  4  0  0  0  0
  2  3  4  0  0  0  0
  3  4  4  0  0  0  0
  4  5  4  0  0  0  0
  5  6  4  0  0  0  0
  1  6  4  0  0  0  0
  4  7  1  0  0  0  0
  7  8  2  0  0  0  0
  7  9  1  0  0  0  0
M  END
$$$$
```



A LINEAR NOTATION TO REPRESENT MOLECULES

- Simplified Molecular Input Line Specification (SMILES)
- Developed by David Weininger, 1988

SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

DAVID WEININGER

Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987

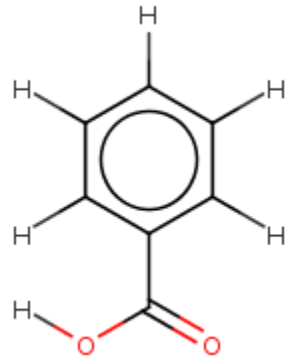
SMILES (Simplified Molecular Input Line Entry System) is a chemical notation system designed for modern chemical information processing. Based on principles of molecular graph theory, SMILES allows rigorous structure specification by use of a very small and natural grammar. The SMILES notation system is also well suited for high-speed machine processing. The resulting ease of usage by the chemist and machine compatability allow many highly efficient chemical computer applications to be designed including generation of a unique notation, constant-speed (zeroeth order) database retrieval, flexible substructure searching, and property prediction models.

SMILES – MAIN POINTS

- Atoms represented by their chemical symbols
  - Uppercase for aliphatic
  - Lowercase for aromatic
- Implicit hydrogen atoms (but explicit definition possible)
- Implicit single bonds
- Greatly reduces complexity and redundancy

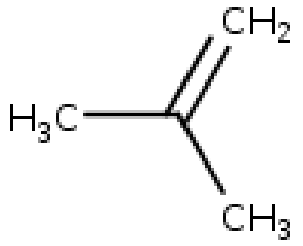

```
OC(=O)c1ccccc1
```

```
[H]OC(=O)c1c([H])c([H])c([H])c([H])c1[H]
```



SMILES (II)

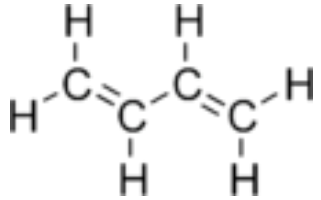
- Double and triple bonds represented with = and #
- Branches represented with parentheses



2-methylprop-1-ene
C=C(C)C

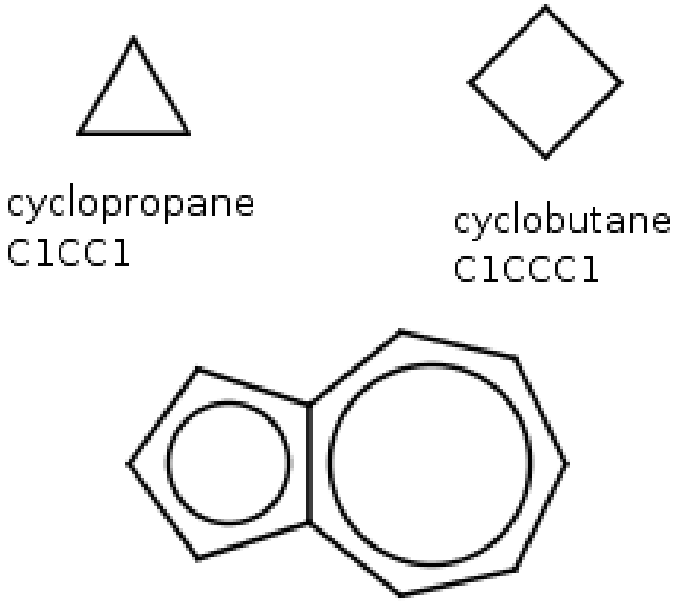
But-1-ene
C=CCC

What about 1,3-Butadiene?



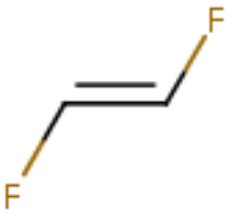
SMILES (III)

- Ring closure represented by numbers



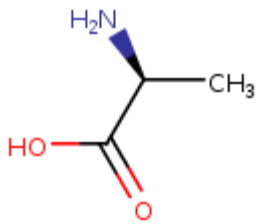
SMILES (IV)

- Stereochemistry at double bonds
  - Z, cis, zusammen, \/ or /\
  - E, trans, Entgegen, // or \\
- Chirality
  - @ counter-clockwise
  - @@ clockwise

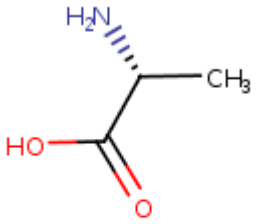


trans-difluoroethene
F/C=C/F

cis-difluoroethene
F/C=C\F



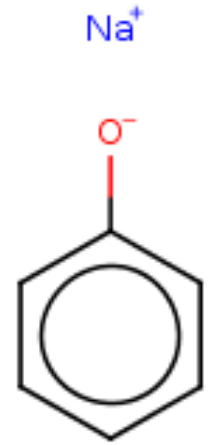
L-alanine
N[C@@H](C)C(=O)O



D-alanine
N[C@H](C)C(=O)O

SMILES (V)

- Atoms of rare elements (not B, C, N, S, P, F, Cl, Br, I) in square brackets; e.g. [Au]
- Charges – use + or -
- Disconnected structures separated by .



sodiumphenolate
[Na+].[O-]c1ccccc1

RECAP

- Limitless ways how to represent molecules in computer systems using notion of “Descriptors”
- Proteins usually stored in PDB file format
- Connectivity information (2D) may be represented using SMILES
- Small molecules may contain 3D information (e.g. SDF).
  - Large sizes


PRACTICAL