Binary Substructure Descriptors for Organic Compounds

K. VARMUZA*, W. DEMUTH, H. SCSIBRANY

Vienna University of Technology
Institute of Chemical Engineering

Laboratory for **C**hemo**M**etrics

* Corresponding and presenting author kvarmuza@email.tuwien.ac.at
www.lcm.tuwien.ac.at
Getreidemarkt 9/166-2
A-1060 Vienna, Austria

Acknowledgment Austrian Science Fund, project P14792-CHE
M. Karlovits, A. Kerber, R. Laue, S. Stein, R. Neudert

Poster Presentation:
Mathematics, Chemistry & Computer Sciences - MATH/CHEM/COMP 2004
21 - 26 June 2004, Dubrovnik, Croatia

Introduction / Overview

A set of 1365 substructures has been defined for the representation of organic compounds by binary vectors.

- Substructure encoding is evident to chemists and easily interpretable.
- Substructure encoding is capable to cover the great diversity of chemical structures.

Software SubMat has been developed for an easy and flexible generation of binary substructure descriptors [1,2].

Only 2-dimensional (connectivity) data of chemical structures are considered.

Applications are reported for

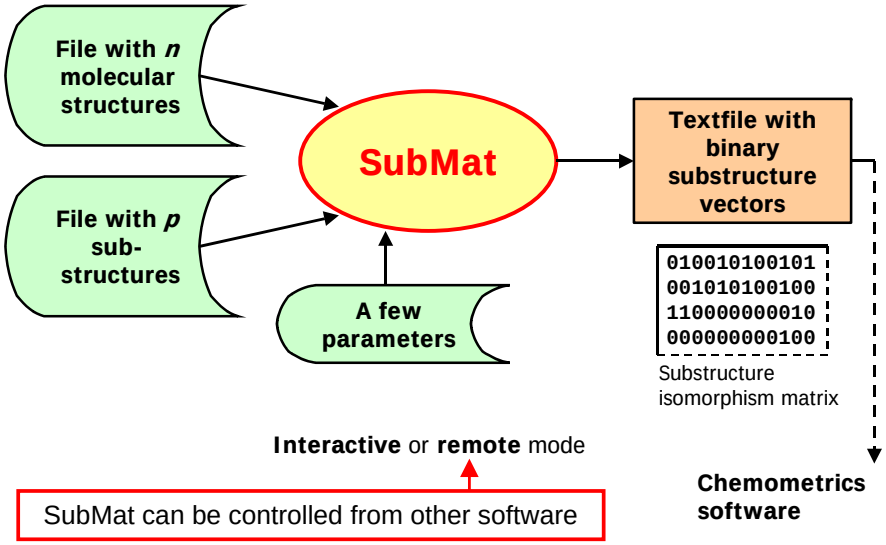
- **characterization of structural diversity** [2],
- **search for similar structures** [2],
- **test of spectra similarity measures** [3,4],
- **cluster analysis of structures** [1,2].

[1] K. Varmuza, H. Scsibrany: J. Chem. Inf. Comput. Sci. **40** (2000) 308-313.
[2] H. Scsibrany, M. Karlovits, W. Demuth, F. Müller, K. Varmuza: Chemom. Intell. Lab. Syst. **67** (2003) 95-108.
[3] K. Varmuza, M. Karlovits, W. Demuth: Anal. Chim. Acta **490** (2003) 313-324.
[4] W. Demuth, M. Karlovits, K. Varmuza: Anal. Chim. Acta (2004) in print.

Software SubMat

SubMat calculates binary substructure descriptors for an input file with molecular structures, and an input file with substructures (all in Molfile format).

SubMat runs under MS Windows operating systems.



Example

n = 1000 molecular structures, and *p* = 200 substructures
1 s computation time (Pentium IV, 2.6 GHz)
5 µs per descriptor value

Demo version and User Guide free at
www.lcm.tuwien.ac.at (Software)

Authors: H. Scsibrany and K. Varmuza

Substructures

Groups

1	Elements (single atoms)	46
2	Two-atom substructures	78
3	Single rings (not aromatic)	404
4	Condensed rings (not aromatic)	130
5	Aromatic rings	97
6	Other rings	39
7	Trees (chains and branches)	418
8	Functional groups	153
	sum	**1365**

Bonds: single, double, triple, aromatic, any type.
Pseudo elements: A (hetero atom), Q (any atom, except H).

Examples

Group 3: single rings, not aromatic

IR: 0.69%	IR: 1.61%	IR: 0.59%	IR: 0.02%
MS: 1.15%	MS: 0.20%	MS: 2.32%	MS: 1.40%

Group 5: aromatic rings

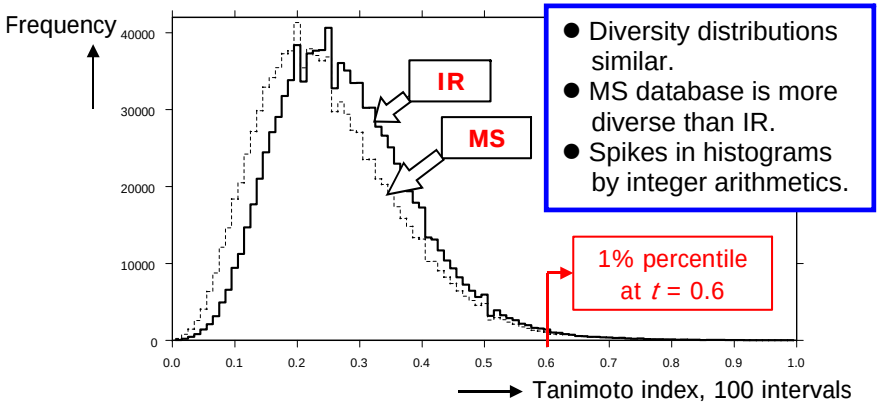
IR: 11.21%	IR: 3.36%	IR: 6.42%	IR: 1.73%
MS: 14.20%	MS: 2.11%	MS: 3.81%	MS: 1.08%

Frequency of compounds containing the substructure are given for two spectroscopic databases. IR, 13,484 compounds; MS, 106,955 compounds.

Applications

Characterization of structural diversity

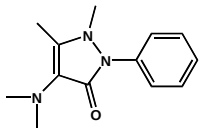
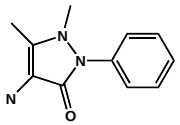
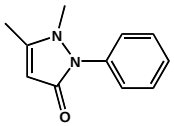
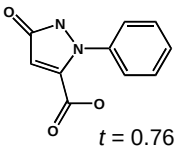
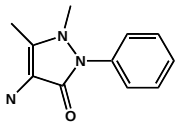
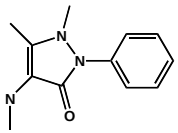
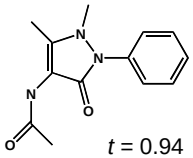
Frequency distributions of **Tanimoto indices** (*t*, a measure for structural similarity) from 10^6 randomly selected structure pairs from two spectroscopic databases. IR, 13,484 compounds; MS, 106,955 compounds.



- Diversity distributions similar.
- MS database is more diverse than IR.
- Spikes in histograms by integer arithmetics.

Search for similar structures

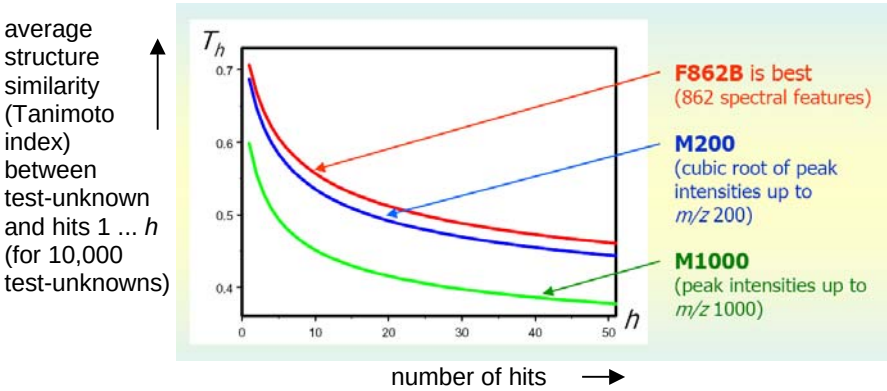
The query structure has been searched in two spectroscopic databases. IR, 13,484 compounds; MS, 106,955 compounds. Hits 1 - 3 are shown, including Tanimoto indices, *t*.

query structure	data base	hit 1	hit 2	hit 3
	IR	 *t* = 0.98	 *t* = 0.90	 *t* = 0.76
	MS	 *t* = 0.98	 *t* = 0.98	 *t* = 0.94

Applications

Spectral versus structural similarity

Three methods (F862B, M200, M1000) for mass spectra similarity searches have been compared with 10,000 "test-unknowns". Method F862 yields hitlists with best chemical structure information. Database: 106,955 comps.



Cluster analysis of structures

A spectral similarity search (IR) for the "test-unknown" **3-amino-benzyl-alcohol** gave 25 compounds. PCA with 18 binary substructure descriptors (selected by maximum variance) shows potential substance classes for the test-unknown. PC1,2: 36%, 28% of total variance, resp.