Diversity of chemical structure libraries characterized by the distribution of Tanimoto indices

K. VARMUZA* and 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

Vienna University of Technology
Institute of Chemical Engineering

Getreidemarkt 9/166
A-1060 Vienna, Austria

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

Poster Presentation:
7th International Conference on Chemical Structures
5 - 9 June 2005, Noordwijkerhout, The Netherlands

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 [1-4].

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

Applications are reported here for

- **characterization of structural diversity** [2]**,**
- **search for similar structures** [2]**,**
- **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.

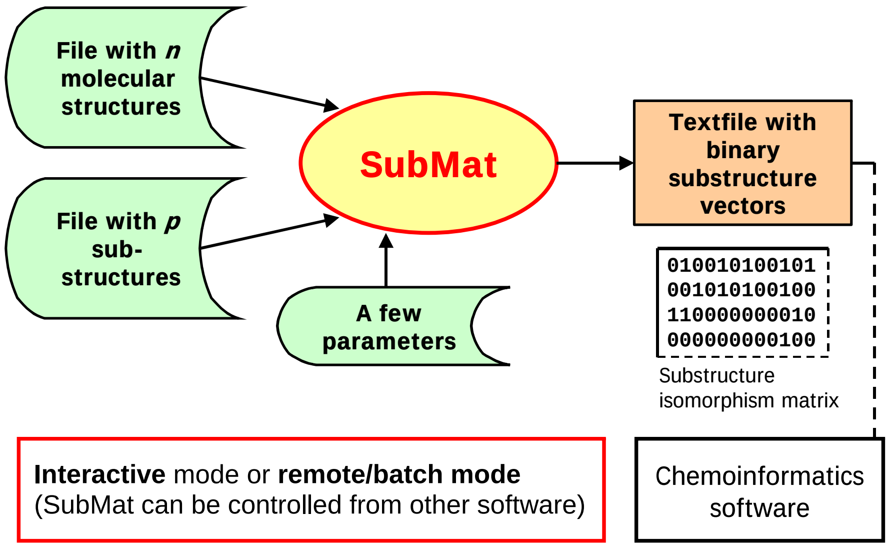
[3] K. Varmuza, M. Karlovits, W. Demuth: Anal. Chim. Acta **490** (2003) 313-324.

[4] W. Demuth, M. Karlovits, K. Varmuza: Anal. Chim. Acta **516** (2004) 75-85.

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.



File with *n* molecular structures

File with *p* sub-structures

A few parameters

SubMat

Textfile with binary substructure vectors

```
010010100101
001010100100
110000000010
000000000100
```

Substructure isomorphism matrix

Interactive mode or **remote/batch mode** (SubMat can be controlled from other software)

Chemoinformatics software

Example

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

Demo version and User Guide free

www.lcm.tuwien.ac.at (Software)

Substructures

Groups

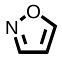
1	**Elements (single atoms)**	**46**
2	**Two-atom substructures**	**78**
3	**Single rings (not aromatic)**	**404**
4	**Condensed rings (not aromatic)**	**130**
5	**Aromatic rings**	**97**
6	**Other rings**	**39**
7	**Trees (chains and branches)**	**418**
8	**Functional groups**	**153**
	Total number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

Bonds: single, double, triple, aromatic, any type.
Pseudo elements: A (hetero atom), Q (any atom, except H) [5].

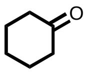
Examples

Group 3: single rings, not aromatic

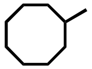
IR: 0.69%
MS: 1.15%



IR: 1.61%
MS: 0.20%

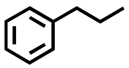


IR: 0.59%
MS: 2.32%

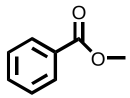


IR: 0.02%
MS: 1.40%

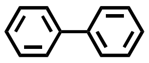
Group 5: aromatic rings



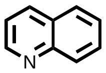
IR: 11.21%
MS: 14.20%



IR: 3.36%
MS: 2.11%



IR: 6.42%
MS: 3.81%



IR: 1.73%
MS: 1.08%

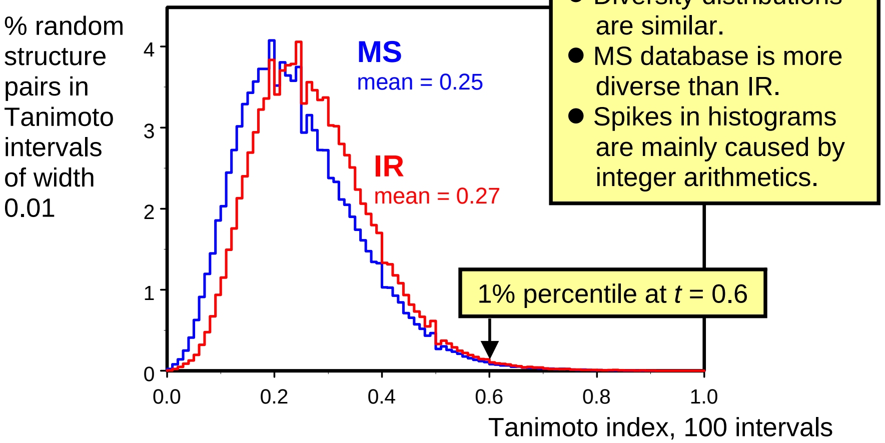
Frequencies of compounds containing the substructure are given for two spectroscopic databases. IR, 13,484 compounds; MS, 106,955 compounds.

[5] K. Varmuza, W. Demuth, M. Karlovits, H. Scsibrany: Croat. Chem. Acta, in print.

Applications

Characterization of structural diversity

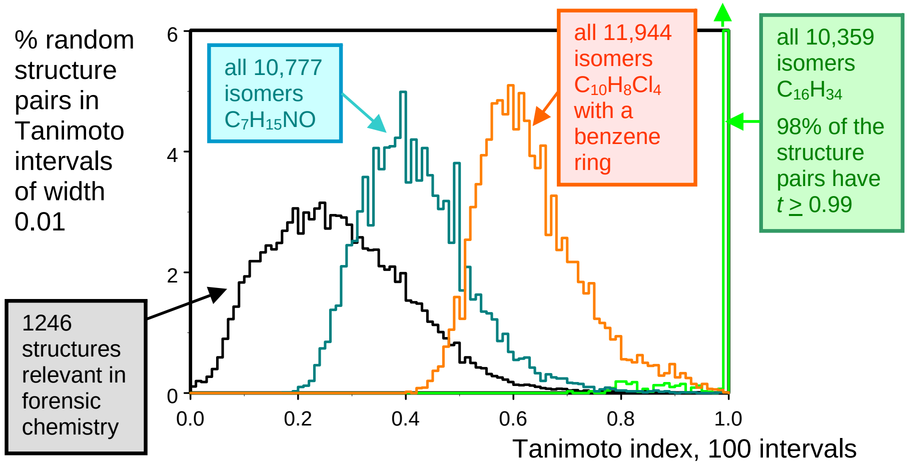
- Frequency distributions of **Tanimoto indices** (*t*) for **500,000 randomly selected structure pairs** from two spectroscopic databases. IR, 13,484 compounds; MS, 106,955 compounds.



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

1% percentile at *t* = 0.6

- Frequency distributions of **Tanimoto indices** (*t*) for 10,000 - 100,000 **randomly selected structure pairs** from four structure libraries.



all 10,777 isomers $C_7H_{15}NO$

all 11,944 isomers $C_{10}H_8Cl_4$ with a benzene ring

all 10,359 isomers $C_{16}H_{34}$

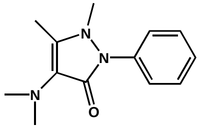
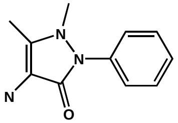
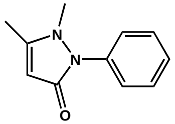
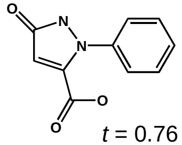
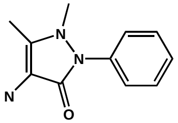
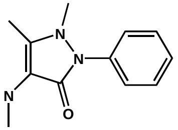
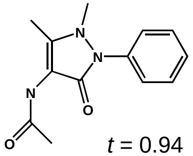
98% of the structure pairs have *t* ≥ 0.99

1246 structures relevant in forensic chemistry

Applications

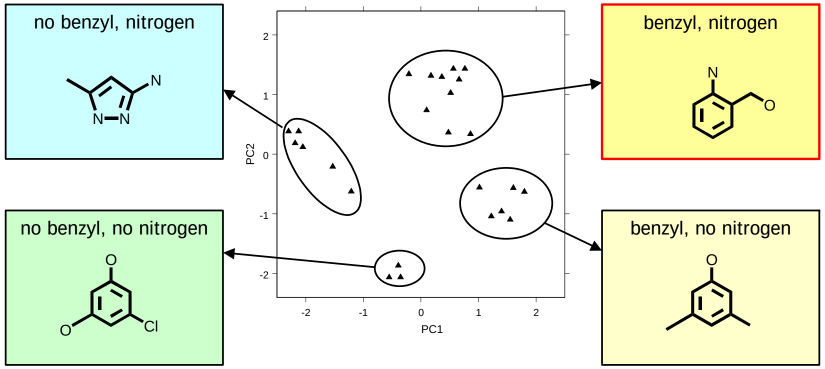
Search for similar structures

A 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

Cluster analysis of structures

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


no benzyl, nitrogen
benzyl, nitrogen
no benzyl, no nitrogen
benzyl, no nitrogen
PC1
PC2