New Topological Information Indices Based on the Full Neighborhood of All Atoms

Kurt Varmuza[1]*, Matthias Dehmer[2], Stephan Borgert[3]

[1] Vienna University of Technology, Austria
Institute of Chemical Engineering

[2] University of Coimbra, Portugal, Center for Mathematics

[3] Darmstadt University of Technology, Germany
Department of Computer Science

* Presenting author
Laboratory for Chemometrics,
Institute of Chemical Engineering, Vienna University of Technology
Getreidemarkt 9/166, A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at, www.lcm.tuwien.ac.at

Poster Presentation: MATH/CHEM/COMP 2008 Conference (MCC 08)
June 10–13, 2008, Verbania, Italy

Introduction

We present a new family of topological information indices based on the full neighborhood of all atoms.

In previous definitions, mostly certain partitions of atoms have been used for this purpose [1-3].
We consider each atom of a molecular structure as a sub-system.

For each atom the complete neighborhood is characterized by an information functional [4], considering the number of atoms in all possible spheres around the atom.

An appropriate weighting scheme combines the number of atoms in the different spheres resulting in a characteristic property of the atom.

The properties of all atoms are normalized to give "probabilities for the sub-systems" necessary for the computation of an entropy measure.

In the current version only skeletons of the chemical structures are considered, with all atoms equal and all bonds equal.

Method

A chemical structure is represented by a graph.

The graph consists of n subsystems corresponding to the n atoms of the structure.

For each subsystem the value f_i of an invariant is calculated based on the complete neighborhood.
We call f_i a special information functional.

The values of the invariants are normalized to give "probabilities" p_i that are combined to an entropy measure E, defining a molecular descriptor.

$$f_i = c_1 s_{i1} + c_2 s_{i2} + \ldots + c_d s_{id}$$

s_{ik} number of atoms in sphere k of atom i

c_k weight for sphere k (e.g. linearly decreasing with increasing k ; $c_1 = d$; $c_2 = d - 1$; ... , $c_d = 1$)

d topological diameter of the structure/graph

$$p_i = f_i \Big/ \sum_{j=1}^{n} f_j$$

$$E = a \left(\mathrm{ld}\, n + \sum_{i=1}^{n} p_i \,\mathrm{ld}\, p_i \right)$$

a is a scaling constant, e.g. 1000.

- If all atoms are topological equivalent (vertex transitive), $E = 0$.
  Examples: rings, prisman, tetrahedron, cube
- E increaes with increasing "neighborhood-diversity" of the atoms
  Examples: chain structures have high values for E.

Results

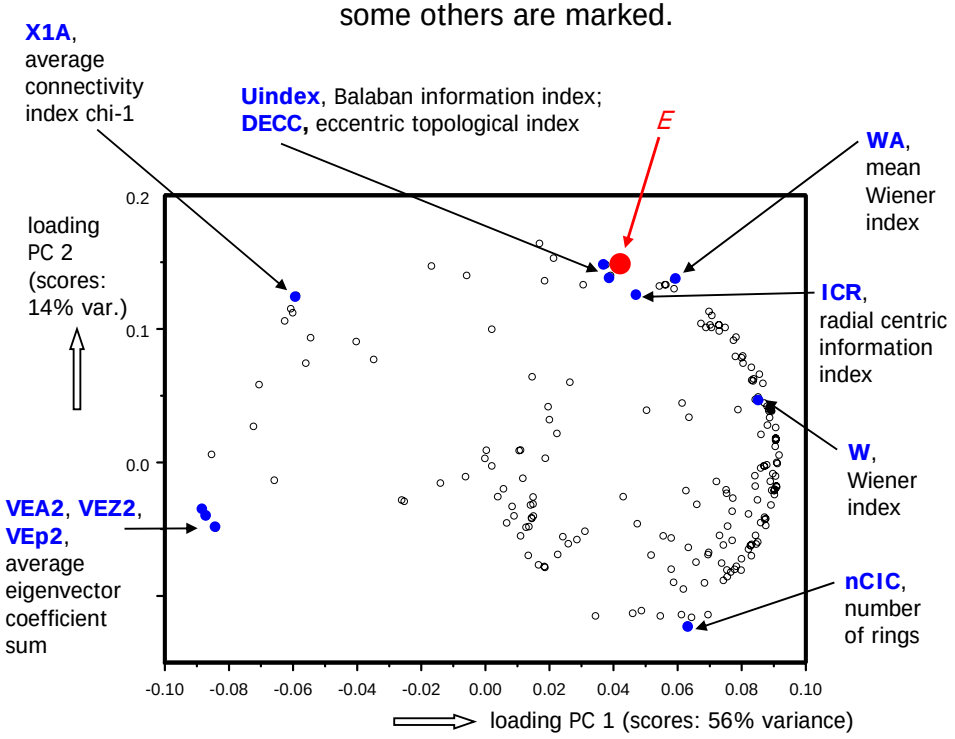
Relationship to other molecular descriptors

A PCA loading plot is used to characterize the multivariate similarity of molecular descriptors, including the new information index E.

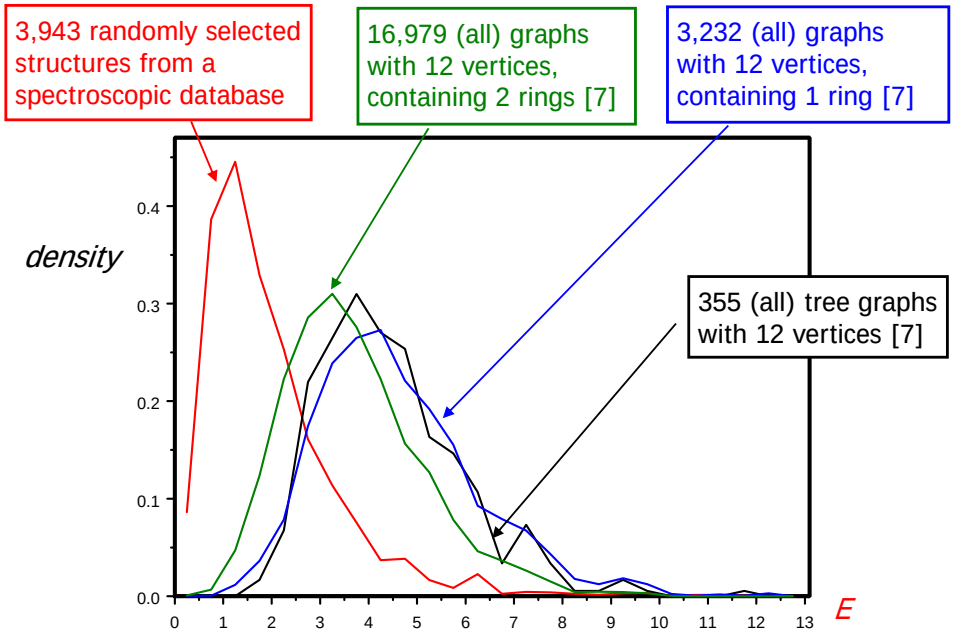
Data n = 3,943 chemical structures, randomly selected from a spectroscopic database [5].

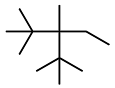
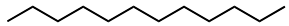
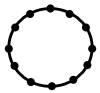
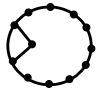
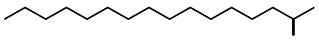
m = 211 molecular descriptors calculated by software *Dragon* [6] from 2D H-depleted structures.

PCA loading plot Calculated from autoscaled descriptors. Descriptors most similar to E, as well as some others are marked.



Distribution of *E* for various structures



	minimum *E*	maximum *E*
trees	**1.8413** 	**11.6165** 
one ring	**0** 12-ring 	**12.4312** 
two rings	**0.1678** 11-ring and 3-ring 	**12.7809** 
spec database	**0** 	**11.6431** 

- *E* characterizes the diversity of the atoms in terms of neighborhood, and thereby a special type of structural complexity and inner symmetry.
- In contrary to previously defined information indices, *E* uses each atom separately (and not in groups), and the neighborhood of atoms considers the whole molecule.
- Extensions of the basic measure *E* for colored graphs (different atoms and different bonds) is under development.
- We generalized the classical information indices because our measure is parameterized and allows the incorporation of various information functionals. Thus, these molecular descriptors can be optimized by machine learning techniques using appropriate data sets.

References

[1] R. Todeschini, V. Consonni, Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany, 2000.
[2] D. Bonchev, N. Trinajstic, J. Chem. Phys. 67 (1977) 4517-4533. Information theory, distance matrix, and molecular branching.
[3] A. Mowshowitz, Bull. Math. Biophys. 30 (1968) 175-204. Entropy and the complexity of the graphs I: An index of the relative complexity of a graph.
[4] M. Dehmer, F. Emmert-Streib, Computational Biology and Chemistry 32 (2008) 131-138. Structural information content of networks: Graph entropy based on local vertex functionals.
[5] NIST, Mass Spectral Database 98, National Institute of Standards and Technology, www.nist.gov/srd/nist1a.htm, Gaithersburg, MD, USA (1998).
[6] Software Dragon, 5.0, Talete srl, www.talete.mi.it, Milan, Italy (2004).
[7] Molgen isomer generator software, Institute for Mathematics II, University of Bayreuth, www.molgen.de, Bayreuth, Germany (2000).