Molecular Descriptors Based on Entropy and the Full Topological 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:
4th German Conference on Chemoinformatics / 22nd CIC-Workshop
Goslar, November 9-11, 2008

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].
Here, we consider each atom of a molecular structure as a sub-system. Topological properties of all atoms together give the value of a descriptor.

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

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

The topological properties of all atoms are normalized to give "probabilities for the sub-systems" necessary for the computation of an entropy measure (the value of a descriptor).

In the current version only skeletons of the chemical structures are considered, with all atoms being equal and all bonds being 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 (topological property) is calculated based on the complete neighborhood.

Here, f_i represents a special information functional.

The values of the invariants are normalized to give "probabilities" p_i that are combined to an entropy measure I, 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 or quadratically decreasing with increasing *k*)
d	topological diameter of the structure/graph

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

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

$I_{LIN\,01}$	linear decrease of c_k
$I_{QUAD\,01}$	quadratic decrease of c_k

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

❍ If all atoms are topologically equivalent (vertex transitive), *I* = 0.

Examples: rings, prisman, tetrahedron, cube

❍ *I* increaes with increasing "neighborhood-diversity" of the atoms

Examples: chain structures have high values for *I* .

Results (1)

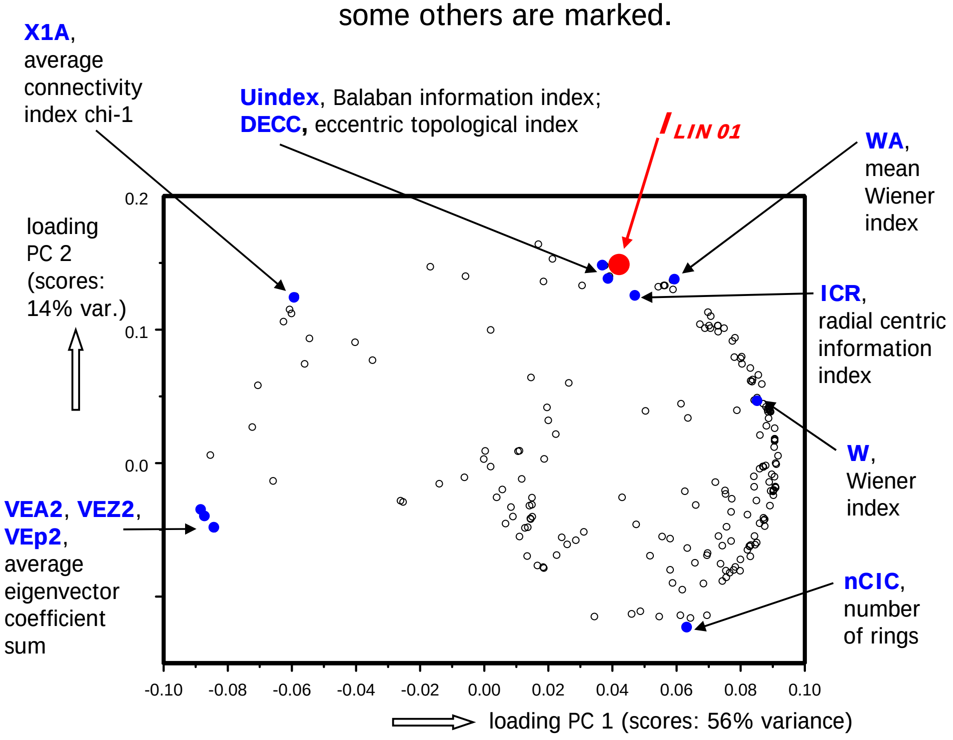
Relationship to other molecular descriptors

A PCA loading plot is used to characterize the multivariate similarity of molecular descriptors, including the new information index $I_{LIN\,01}$

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 $I_{LIN\,01}$, as well as some others are marked.



X1A, average connectivity index chi-1

Uindex, Balaban information index; **DECC**, eccentric topological index

$I_{LIN\,01}$

WA, mean Wiener index

ICR, radial centric information index

W, Wiener index

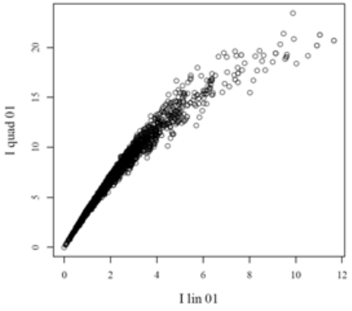
VEA2, **VEZ2**, **VEp2**, average eigenvector coefficient sum

nCIC, number of rings

Results (2)

Relationship to other molecular descriptors

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



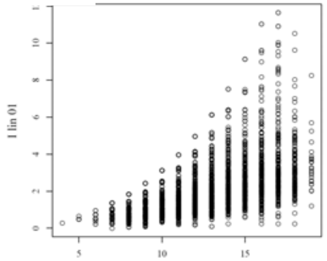
$I_{LIN\,01}$ (linear decrease of neighborhood weights, c_k)

and

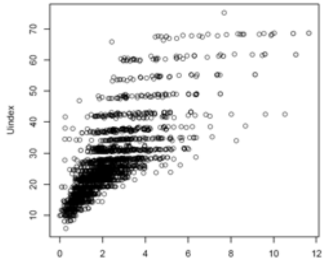
$I_{QUAD\,01}$ (quadratic decrease of neighborhood weights, c_k)

are highly correlated.

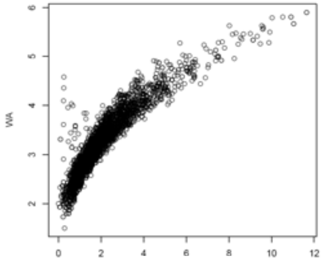
$I_{LIN\,01}$



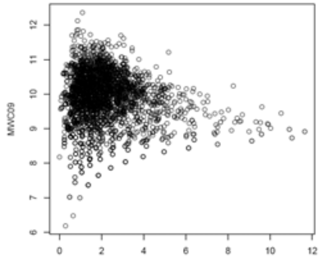
Balaban information index



Mean Wiener index

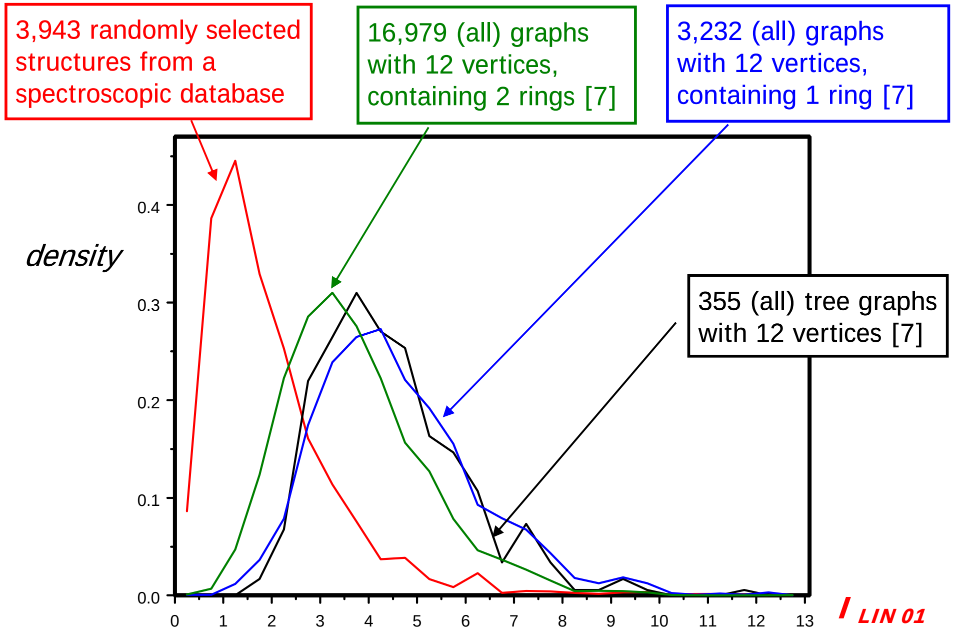


A walk-and-path count index



Results (3)

Distribution of $I_{LIN\,01}$ for various structure sets

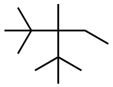
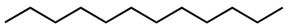
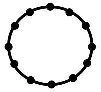
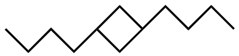
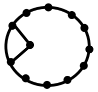
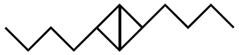
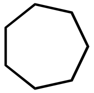
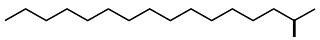


3,943 randomly selected structures from a spectroscopic database

16,979 (all) graphs with 12 vertices, containing 2 rings [7]

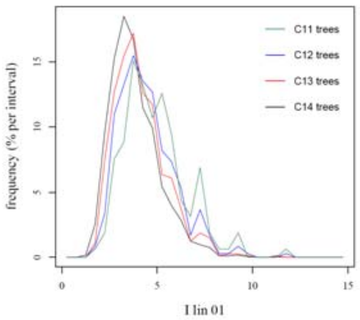
3,232 (all) graphs with 12 vertices, containing 1 ring [7]

355 (all) tree graphs with 12 vertices [7]

	minimum $I_{LIN\,01}$		maximum $I_{LIN\,01}$
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** 

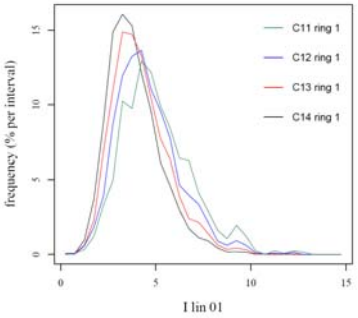
Results (4)

Distribution of $I_{LIN\ 01}$ for various structure sets



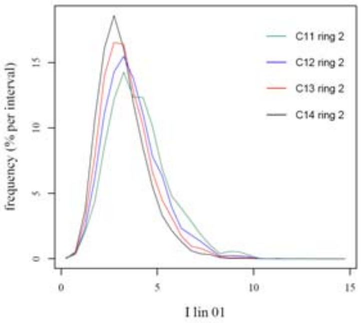
All **trees**
containing 11 to 14 vertices [7]:

C11:	n = 159
C12:	n = 355
C13:	n = 802
C14:	n = 1858



All **graphs with 1 ring**
containing 11 to 14 vertices [7]:

C11:	n = 1231
C12:	n = 3232
C13:	n = 8506
C14:	n = 22,565



All **graphs with 2 rings**
containing 11 to 14 vertices [7]:

C11:	n = 5533
C12:	n = 16,977
C13:	n = 51,652
C14:	n = 156,291

Summary

- The new descriptor $I_{LIN\ 01}$ characterizes the diversity of the atoms in terms of neighborhood, that is a special aspect of structural complexity and inner symmetry.
- In contrary to previously defined information indices, each atom is treated separately (and not in groups), and the neighborhood of atoms considers the whole molecule.
- Extension 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).