Uniqueness of molecular descriptors for organic compounds

Kurt Varmuza[1*] and Matthias Dehmer[2]

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

[2] **UMIT, The Health and Life Sciences University,**
Institute for Bioinformatics and Translational Research
Eduard-Wallnöfer Zentrum 1, A-6060 Hall in Tyrol, Austria
matthias.dehmer@umit.at, www.dehmer.org, www.umit.at

* Presenting author

Poster Presentation

Conferentia Chemometrica 2011 (CC 2011)

September 18–21, 2011, Sümeg, Hungary

Acknowledgments. We gratefully acknowledge support by the Austrian Science Fund (FWF), project P22029-N13. We thank Martin Grabner for collaboration.

Introduction

Uniqueness (or discrimination power or degeneracy) of a molecular descriptor [1] measures the ability to distinguish among different chemical structures. A high uniqueness means that many (ideally all) chemical structures in a given set of structures have different values of the descriptor.

Chemical structures can be represented by graphs [2]. Then uniqueness refers to the ability to distinguish nonisomorphic graphs.

Descriptors with a very high uniqueness are powerful in searches for identical chemical structures (isomorphism tests) in databases. For QSAR/QSPR models a moderate uniqueness is optimal.

Uniqueness and measures used (see page 2) **depend on**

- considered chemical structures (graphs) - the sample,
- number of considered chemical structures (graphs),
- definition of isomorphism (skeleton only, colored graphs),
- accuracy of numerical values (no. of decimal digits).

Usefulness of descriptors with high uniqueness requires

- clear, unambiguous definition of descriptor,
- fast and insensitive algorithm,
- unambiguous structural data (e. g., no 3D),
- adequate for the used chemical structures.

In this preliminary study, we summarize measures for uniqueness, and we compare a few relevant descriptors using exhaustive sets of graphs related to chemical structures.

[1] Todeschini R., Consonni V., Mannhold R.: Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany (2002)
[2] Gasteiger J. (ed.): Handbook of Chemoinformatics - From Data to Knowledge (4 vol.), Wiley-VCH, Weinheim, Germany (2003)

Method

Definitions

x [1:*n*] vector with descriptor values for *n* nonisomorphic structures
Example: *n* = 7; ***x*** = **(1, 2, 2, 3, 4, 4, 4)**

g no. of equivalence groups (different values); *g* = 4

d no. of degeneracies (no. of structures with at least another nonisomorphic structure exhibiting the same descriptor value) *d* = 5

u no. of unique descriptor values (only 1 structure has this value (*d* + *u* = *n*) *u* = 2

$\boldsymbol{x}_G$ [1: *g*] descriptor values of the *g* equivalence groups (1, 2, 3, 4)

$\boldsymbol{n}_G$ [1: *g*] number of structure in the *g* equivalence groups (1, 2, 1, 3)

Measures for uniqueness

$U_U = u / n$ ratio unique values / *n*

$U_G = g / n$ ratio no. of equivalence groups / *n*

$U_K = (n - d) / n$ Konstantinova's index sensitivity [3,4]

$U_H = 1 - (\mathrm{ld}\ n - H) / \mathrm{ld}\ n$ Todeschini's entropy measure [5]

$H = \Sigma(n_G[j] / n)\ \mathrm{ld}\ (n_G[j] / n) \quad j = 1 \ldots g$

- All these uniqueness measures (*U* *) are in the range [0, 1].
- *U* * = 1 if *u* = *g* = *n* (all different); *d* = 0; n_G [all] = 1/*n*
- $U_K = U_H = 0$ if all equal; *u*=*g*=1; *d*=*n*; n_G [1] = *n*; $U_U = U_G = 1/n$
- $U_K = 0$ also if no unique values; *u*=0; *d*=*n*; *e.g.,* (1, 1, 2, 2)
- Instead of *U* * : negative logarithm (base 10) of (1 - *U* *), equivalent to the number of consecutive "9" after decimal point.

[3] Konstantinova E.V., Vidyuk M.V.: J. Chem. Inf. Comput. Sci., **43**, 1860 (2003)
[4] Dehmer M., Barbarini N., Varmuza K., Graber A.: BMC Structural Biology, **10**, 18 (2010)
[5] Todeschini, R., Consonni V., Maiochhi A.: Chemom. Intell. Lab. Syst., **46**, 13 (1999)
[6] Molgen isomer generator software, Inst. Math., Univ. Bayreuth, www.molgen.de, Germany (2000)

Results for alkane skeletons

Exhaustive sets of alkane isomers up to 22 carbon atoms have been generated by software Molgen [6]. These chemical structures (H-depleted) comprise all connected, noncolored tree graphs with a maximum of 4 edges per vertex.

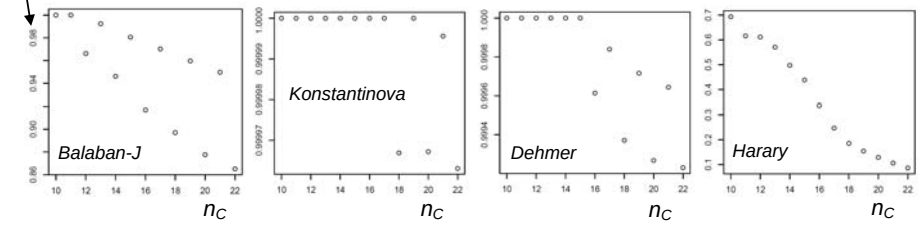
Results for 3 descriptors with high uniqueness are presented, together with a more general descriptor.

Balaban-J	high uniqueness [1]
Konstantinova, information index	high uniqueness [3]
Dehmer, parametrized entropy	high uniqueness, quadrat. weighting [7]
Harary	molecular compactness [1]

All computations were performed with programs written in ***R*** [8] using the free R-library **QuACN** (Müller L.A.J., Dehmer M.).

		Balaban-J		Konstantinova		Dehmer		Harary	
n_C	*n*	*u*	U_U	*u*	U_U	*u*	U_U	*u*	U_U
10	75	75	1.000000	75	1.000000	75	1.000000	52	0.693
11	159	159	1.000000	159	1.000000	159	1.000000	98	0.616
12	355	343	0.966197	355	1.000000	355	1.000000	217	0.611
13	802	796	0.992519	802	1.000000	802	1.000000	458	0.571
14	1858	1758	0.946179	1858	1.000000	1858	1.000000	925	0.498
15	4347	4263	0.980676	4347	1.000000	4347	1.000000	1904	0.438
16	10359	9500	0.917077	10359	1.000000	10355	0.999613	3495	0.337
17	24894	24153	0.970234	24894	1.000000	24890	0.998393	6159	0.247
18	60523	54292	0.897047	60521	0.999967	60485	0.993721	11228	0.186
19	148284	142317	0.959760	148284	1.000000	148242	0.997168	22994	0.155
20	366319	321519	0.877702	366307	0.999967	366051	0.992684	47198	0.129
21	910726	865023	0.949817	910722	0.999996	910402	0.996442	97112	0.107
22	2278658	1971747	0.865311	2278574	0.999963	2276906	0.992311	197505	0.087

n_C no. of C-atoms (vertices) *n* no. of alkane isomers
u no. of unique descriptor values among the *n* isomers
U_U measure for uniqueness ($U_U = u / n$); $U_U = 1$ if all values are different ($u = n$)



[7] Dehmer M., Varmuza K., Borgert S., Emmert-Streib F.: J. Chem. Inf. Model., **49**, 1655 (2009)
[8] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, 2011. www.r-project.org

Results for general isomeric skeletons

Exhaustive sets of isomeric skeletons with 8 to 12 carbon atoms have been generated by software Molgen [6]. These chemical structures (H-depleted) comprise all connected, noncolored graphs (trees, any cycles with ≥3 vertices) with a maximum of 4 edges per vertex.

Results for 3 descriptors with high uniqueness are presented, together with a more general descriptor.

Balaban-J	high uniqueness [1]
Konstantinova, information index	high uniqueness [3]
Dehmer, parametrized entropy	high uniqueness, quadrat. weighting [7]
Harary	molecular compactness [1]

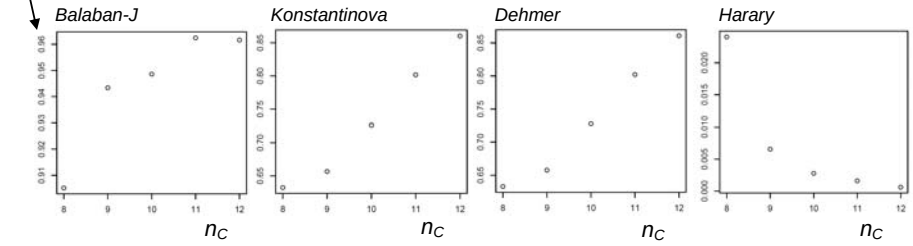
All computations were performed with programs written in ***R*** [8] using the free R-library **QuACN** (Müller L.A.J., Dehmer M.).

		Balaban-J		Konstantinova		Dehmer		Harary	
n_C	n	u	U_U	u	U_U	u	U_U	u	U_U
8	1456	1318	0.90522	921	0.63255	922	0.63324	35	0.02404
9	7000	6604	0.94343	4596	0.65657	4605	0.65786	46	0.00657
10	33840	32101	0.94861	24575	0.72621	24635	0.72798	94	0.00278
11	160294	154271	0.96243	128490	0.80159	128599	0.80227	258	0.00161
12	738928	710554	0.9616	635674	0.86027	636050	0.86077	474	0.00064

n_C no. of C-atoms (vertices) n no. of isomers

u no. of unique descriptor values among the n isomers

U_U measure for uniqueness ($U_U = u / n$); $U_U = 1$ if all values are different ($u = n$)

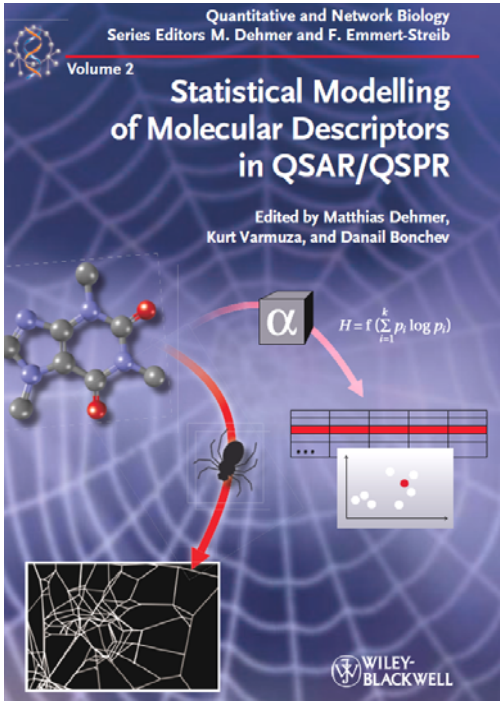


Conclusions

Alkan skeletons: Konstantinova- and Dehmer-indices better than Balaban-J.

General skeletons (chemical structures): Balaban-J index better than Konstantinova- and Dehmer-indices.

Related



Dehmer, Matthias,
Varmuza, Kurt,
Bonchev, Danail (eds.):

Statistical Modelling of Molecular Descriptors in QSAR/QSPR

Wiley-Blackwell

March 2012

ISBN-10: 3-527-32434-8
ISBN-13: 978-3-527-32434-7

ca. 130 Euro

It has been reported that spiders produce disordered nets under the influence of caffeine. Rather positive effects of caffeine are recognized for most humans. However, can simple approaches in QSAR be successful - just using a set of numbers to characterize molecules and applying rather simple (mostly linear) chemometric methods for creating empirical models ?!

TRY but EVALUATE CAREFULLY !

Selected contributions

- Current Modeling Methods used in QSAR/QSPR (L.C. Yee, Y.C. Wei).
- Multivariate Analysis of Molecular Descriptors (V. Consonni, R. Todeschini).
- Structural Similarity based Approaches for the Development of Clustering and QSPR/QSAR Models in Chemical Databases (I.L. Ruiz, G. Cerruela García, Miguel.Á. Gómez-Nieto).
- Consensus Models of Activity Landscapes (J.L. Medina-Franco, A.B. Yongye, F. López-Vallejo).
- Reduction of Dimensionality, Order and Classification in Spaces of Theoretical Descriptions of Molecules. An Approach based on Metrics, Pattern Recognition Techniques and Graph Theoretic Considerations (G. Maroulis).
- Molecular Descriptors and the Electronic Structure (H. Bögel).
- New types of Descriptors and Models in QSAR/QSPR (C. Kramer, T. Clark).