CROATICA CHEMICA ACTA
CCACAA **78** (4) 489–492 (2005)
ISSN-0011-1643
CCA-3039
Original Scientific Paper

On Reducing Correlations between Topological Indices*

Boris Hollas,[a,] Ivan Gutman,[b] and Nenad Trinajstić[c]**

[a] *Theoretische Informatik, Universität Ulm, D-89081 Ulm, Germany*

[b] *Faculty of Science, University of Kragujevac, 34000 Kragujevac, Serbia and Montenegro*

[c] *The Rugjer Bošković Institute, P. O. B. 180, HR-10002 Zagreb, Croatia*

RECEIVED APRIL 11, 2005; REVISED APRIL 29, 2005; ACCEPTED MAY 3, 2005

Keywords
topological index
QSPR
QSAR
Platt number
connectivity index
Zagreb index

A very large number of molecular-graph-based structure descriptors, the so-called *topological indices* (TIs), have been proposed in the recent and current chemical literature. Many of these are highly intercorrelated, which makes their application in QSPR and QSAR studies difficult and purposeless. A class of such TIs (including the Platt number, the connectivity index, and the Zagreb indices) has been examined by methods of mathematical statistics and probability theory, and the reasons for their mutual correlation are revealed. The analysis has shown that by a slight modification of these TIs, their mutual correlation can be reduced or completely eliminated. These theoretical inferences have been corroborated by a computer experiment done on a database consisting of over 126000 distinct molecular structures.

INTRODUCTION

The idea of representing relevant structural features of an organic molecule by means of a number that can be deduced from its structural formula or (in more recent interpretations) from its molecular graph is more than a century old.[1,2] Nowadays, such molecular-graph-based structure-descriptors are usually called *topological indices* (TIs), the term that was proposed by Hosoya[3] and was eventually generally accepted.[2] Until the end of the 1970s, only a limited number (about a dozen) of TIs had appeared in the chemical literature.[4] In more recent times, the number of proposed TIs has enormously increased (and is still increasing) and exceeds one thousand.[5]

The fact that many of the proposed TIs are mutually correlated (almost always linearly) was reported on many occasions.[6–10] This causes major problems in their applications in designing QSPR and QSAR models. (Recall that almost all QSPR and QSAR approaches are based on constructing linear combinations of molecular-structure descriptors, not all of which need to be TIs. If two such descriptors are highly linearly correlated, then the outcomes of the respective model become arbitrary and meaningless.)

In view of the above, it is of great practical importance to know when (and why) two TIs are mutually correlated, and if yes, how this correlation could be reduced or eliminated.

In a series of recently published papers,[11–14] one of the present authors arrived at a solution to this problem, applicable to a wide class of TIs. Because these papers employ sophisticated methods of mathematical statistics and probability theory,[15] they may evade the attention of chemists interested in practical aspects and applications

* Dedicated to Dr. Edward C. Kirby in happy celebration of his 70th birthday

** Author to whom correspondence should be addressed. (E-mail: hollas@informatik.uni-ulm.de)

of QSPR and QSAR. In order to bridge this gap, in this paper we briefly and in a somewhat simplified manner re-state the main result of Refs. 11–14, and then illustrate them by means of a pertinently designed computer experiment.

At this point it is worth mentioning that another approach to reducing correlation between TIs was proposed by Randić:[16,17] the use of orthogonalized linear combinations of (several) TIs. This method was eventually much applied and further elaborated.[18–21] It, however, has nothing in common with the approach outlined in the present paper.

THEORY

In Refs. 11–13 random graph models were used to analyze TIs of the form

$$TI_X(G) = \sum_{\{uv\} \in E} X_u X_v \qquad (1)$$

where $G = (\boldsymbol{V}, \boldsymbol{E})$ is a molecular graph, and $\boldsymbol{V}$ and $\boldsymbol{E}$ are, respectively, its vertex- and edge-sets. Thus, the summation on the right-hand side of (1) goes over all pairs u,v of adjacent vertices of the graph G, *i.e.*, over all edges $\{u,v\}$ of G.

Under a random graph model[11–13] we understand a set of graphs with a probability distribution defined on it. Thus, our *random graph* is a graph chosen at random from the respective (large) set of graphs, according to the respective probability distribution. The random graph models considered in Refs. 11–13 differ in the assumptions made on the underlying probability distribution.

In what follows, the number of edges of the graph G, *i.e.*, the number of elements of the set $\boldsymbol{E}$, will be denoted by m.

The quantity X_v in formula (1) is some property associated with the vertex v. This quantity is viewed as a random variable and it is assumed that its expectation value[15] E(X) is independent of the vertex v and also independent of the graph G.

Under these assumptions, the following holds:[11]

(i) Topological indices TI_X and TI_Y are linearly correlated if the expectation values E(X) and E(Y) of the vertex properties X and Y are large. As E(X) and E(Y) tend to infinity, the correlation coefficient[15] $Corr(TI_X,TI_Y)$ tends to 1. Then, in addition, these TIs are linearly correlated with the parameter m.

(ii) Topological indices TI_X, TI_Y are uncorrelated (and their correlation coefficient is equal to zero) if E(X) = 0 or E(Y) = 0.

(iii) Topological indices TI_X and m are uncorrelated if E(X) = 0.

Thus, even if completely different properties X, Y are encoded, the resultant indices TI_X, TI_Y are strongly correlated if X_u, X_v have large expectation values. All information, except the number of edges, is lost. In particular, it was shown[11] that if E(X), E(Y) > 4.1 then $Corr(TI_X,TI_Y) > 0.7$.

On the other hand, according to (ii), these correlations are eliminated if on the right-hand side of (1) $X_u X_v$ is replaced by $X_u X_v - E(X) E(X)$.

The above results hold also in the case when the summation in (1) goes over pairs of vertices u,v at a fixed distance d, $d \neq 1$.

* * *

Although the above results may look interesting, their practical applicability is limited. Namely, contrary to the assumptions on which the results (i)–(iii) are based, the vertex properties of interest in chemical applications are not independent of the molecular graph. In view of this, a different model was considered,[14] in which the vertex property depends on the degree of a vertex. (Recall that the degree deg(v) of the vertex v is the number of its first neighbors.)

Let

$$TI_X(G) = \sum_{\{uv\} \in E} X_{uv} \qquad (2)$$

be a topological index with

$$X_{uv} = f[\deg(u)\, \deg(v)] \qquad (3)$$

where f is some function. The Platt number,[22] the connectivity index,[23,24] the 2nd Zagreb index,[25–28] and its modified version[27,29] are topological indices of this kind; for details see the book.[5]

As in the case of independent vertex properties, it was shown that TI_X can be transformed to an index $TI_{\tilde{X}}$, such that

$$\tilde{X}_{uv} = X_{uv} - \frac{Cov(TI_X, m)}{Var(m)} \qquad (4)$$

where *Var* and *Cov* stand for variance and covariance,[15] respectively. Then $TI_{\tilde{X}}$ and m are uncorrelated, with a zero correlation coefficient. (This means that there is no linear correlation between $TI_{\tilde{X}}$ and m. It could be shown[14] that also any curvilinear dependence between $TI_{\tilde{X}}$ and m is absent.)

For properties X_u, X_v that both depend on the vertex-degree as in (3), we cannot expect $TI_{\tilde{X}}$ and $TI_{\tilde{Y}}$ to become uncorrelated. It is reasonable to assume that $|Corr(TI_{\tilde{X}}, TI_{\tilde{Y}})| < |Corr(TI_X, TI_Y)|$ as the correlation with m is eliminated. The validity of relations of this kind needs, however, to be tested on concrete examples, which we actually do in what follows.

In the subsequent section we verify these theoretical results for the connectivity index χ, the 2nd Zagreb index M_2, the modified 2nd Zagreb index M'_2 and the Platt number F. These indices are defined as follows:

$$\chi = \sum_{\{uv\}\in E} \frac{1}{\sqrt{\deg(u)\deg(v)}} \quad (5)$$

$$M_2 = \sum_{\{uv\}\in E} \deg(u)\deg(v) \quad (6)$$

$$M'_2 = \sum_{\{uv\}\in E} \frac{1}{\deg(u)\deg(v)} \quad (7)$$

$$F = \sum_{\{uv\}\in E} [\deg(u) + \deg(v) - 2] \quad (8)$$

COMPUTATIONAL RESULTS

The molecular structures used in our calculations were taken from the free NCI 127k database[30] of the National Cancer Institute that contains connection tables for about 127,000 structures. These structures are the compounds in the NCI database from July 27, 1993 that met three conditions:

(I) A complete connection table existed.

(II) The compound was not covered by a proprietary agreement.

(III) A CAS registry number was given to the compound.

Some of these structures were not connected and were therefore discarded, leaving 126,674 structures for the computer experiment. Table I shows the relative means (mean/variance) for the vertex properties X_v in formula (1) for the respective TI. For the Platt number F, which is not of the form required by (1), the relative mean of $\deg(u) + \deg(v) - 2$ was used.

TABLE I. Relative means of vertex properties

TI	χ	M_2	M'_2	F
E(TI)/Var(TI)	31.02	3.61	9.33	3.27

As can be expected from the theoretical results, the TIs with highest relative means are those most strongly correlated (see Table II).

TABLE II. Correlation matrix for untransformed topological indices; quantities χ, M_2, M'_2, F are defined *via* Eqs. (5)–(8); m is the number of edges.

	m	χ	M_2	M'_2	F
M	1.0000	0.9939	0.9545	0.9748	0.9703
χ		1.0000	0.9214	0.9930	0.9415
M_2			1.0000	0.8830	0.9950
M'_2				1.0000	0.9052
F					1.0000

Table III shows the correlations for these indices after X_{uv} in (2) has been replaced by $\tilde{X}_{uv}$, see Eq. (4).

TABLE III. Correlation matrix for topological indices transformed according to Eq. (4). Other data same as in Table II.

	m	$\tilde{\chi}$	$\tilde{M}_2$	$\tilde{M}'_2$	$\tilde{F}$
m	1.0000	0.0000	0.0000	0.0000	0.0000
$\tilde{\chi}$		1.0000	–0.8256	0.9788	–0.8518
$\tilde{M}_2$			1.0000	–0.7119	0.9547
$\tilde{M}'_2$				1.0000	–0.7519
$\tilde{F}$					1.0000

The correlations with m are eliminated by construction, whereas all other correlations are significantly reduced. As already mentioned, we cannot expect all correlations to vanish since all four TIs are sums of vertex properties that depend on the vertex-degree. A greater reduction of correlations is to be expected for TIs with substantially different vertex properties.

REFERENCES

1. D. H. Rouvray, in: D. Bonchev and D. H. Rouvray (Eds.), *Chemical Graph Theory – Introduction and Fundamentals*, Gordon & Breach, New York, 1990, pp. 22–39.
2. A. T. Balaban and O. Ivanciuc, in: J. Devillers and A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon & Breach, Amsterdam, 1999, pp. 21–57.
3. H. Hosoya, *Bull. Chem. Soc. Jpn.* **44** (1971) 2332–2339.
4. A. T. Balaban, I. Motoc, D. Bonchev, and O. Mekenyan, *Topics Curr. Chem.* **114** (1983) 21–55.
5. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
6. I. Motoc, A. T. Balaban, O. Mekenyan, and D. Bonchev, *MATCH–Commun. Math. Comput. Chem.* **13** (1982) 369–404.
7. K. Kovačević, D. Plavšić, N. Trinajstić, and D. Horvat, *Stud. Phys. Theor. Chem.* **63** (1989) 213–224.
8. D. Horvat, A. Graovac, D. Plavšić, N. Trinajstić, and M. Strunje, *Int. J. Quantum Chem.: Quantum Chem. Symp.* **26** (1992) 401–408.
9. S. C. Basak, B. D. Gute, and A. T. Balaban, *Croat. Chem. Acta* **77** (2004) 331–344.
10. A. Kerber, R. Laue, M. Meringer, and C. Rücker, *MATCH–Commun. Math. Comput. Chem.* **51** (2004) 187–204.
11. B. Hollas, *MATCH–Commun. Math. Comput. Chem.* **45** (2002) 27–33.
12. B. Hollas, *J. Math. Chem.* **33** (2003) 91–101.
13. B. Hollas, *MATCH–Commun. Math. Comput. Chem.* **47** (2003) 79–86.
14. B. Hollas, *MATCH–Commun. Math. Comput. Chem.* **54** (2005) 341–350.
15. For details on statistical concepts encountered in this paper (expectation value, variance, covariance, correlation coefficient) see appropriate textbooks, *e.g.*, B. L. Van der Werden, *Mathematical Statistics*, Springer-Verlag, Berlin, 1969; J. Czerminski, A. Iwasiewicz, and Z. Paszek, *Statistical Methods in Applied Chemistry,* Elsevier, Amsterdam, 1990.
16. M. Randić, *New J. Chem.* **15** (1991) 517–525.

17. M. Randić, *J. Chem. Inf. Comput. Sci.* **31** (1991) 311–320.
18. B. Lučić, S. Nikolić, N. Trinajstić, and D. Juretić, *J. Chem. Inf. Comput. Sci.* **35** (1995) 532–538.
19. M. Šoškić, D. Plavšić, and N. Trinajstić, *J. Chem. Inf. Comput. Sci*. **36** (1996) 829–832.
20. O. Araujo and D. A. Morales, *Chem. Phys. Lett.* **257** (1996) 393–396.
21. O. Araujo and D. A. Morales, *J. Chem. Inf. Comput. Sci*. **38** (1998) 1031–1037.
22. J. A. Platt, *J. Chem. Phys.* **15** (1947) 419–420.
23. M. Randić, *J. Am. Chem. Soc*. **97** (1975) 6609–6615.
24. M. Randić, *J. Mol. Graphics Modell.* **20** (2001) 19–35.
25. I. Gutman and N. Trinajstić, *Chem. Phys. Lett.* **17** (1972) 535–537.
26. I. Gutman, B. Ruščić, N. Trinajstić, and C. F. Wilcox, *J. Chem. Phys.* **62** (1975) 535–538.
27. S. Nikolić, G. Kovačević, A. Miličević, and N. Trinajstić, *Croat. Chem. Acta* **76** (2003) 113–124.
28. K. C. Das and I. Gutman, *MATCH–Commun. Math. Comput. Chem*. **52** (2004) 103–112.
29. D. Vukičević and N. Trinajstić, *Croat. Chem. Acta* **76** (2003) 183–187.
30. National Cancer Institute, Connection Tables for 127000 Structures: ftp://helix.nih.gov/ncidata/2D/nciopen.mol.Z.

SAŽETAK

O umanjivanju korelacije između topoloških indeksa

Boris Hollas, Ivan Gutman i Nenad Trinajstić

U novijoj je kemijskoj literaturi predložen veliki broj strukturnih deskriptora zasnovanih na molekularnome grafu, takozvanih *topoloških indeksa*. Mnogi od njih su u velikoj mjeri međusobno korelirani što otežava ili onemogućava njihovu primjenu u QSPR i QSAR studijama. Jedna skupina ovakvih topoloških indeksa (koja obuhvaća Plattov broj, indeks povezanosti kao i Zagrebačke indekse) proučavana je s pomoću metoda matematičke statistike i teorije vjerojatnosti. Otkriveni su razlozi za njihovu uzajamnu koreliranost. Analiza je pokazala, da se malom modifikacijom ovih topoloških indeksa njihova koreliranost može umanjiti ili potpuno otkloniti. Dobiveni teorijski zaključci potvrđeni su kompjutorskim eksperimentom u kojem je upotrebljena jedna baza podataka s više od 126000 molekularnih struktura.