Using Real Numbers as Vertex Invariants for Third-Generation Topological Indexes

ALEXANDRU T. BALABAN

Department of Organic Chemistry, Technical University (Polytechnic Institute), Splaiul Independent 313, 77206 Bucharest, Romania

Received August 26, 1991

First-generation topological indexes (TI's) were integer numbers obtained by simple ("bookkeeping") operations from local vertex invariants (LOVI's), which were integer numbers. Second-generation TI's were real numbers obtained via sophisticated ("structural") operations from integer LOVI's. Third-generation TI's are real numbers based on real-number LOVI's. In successive generations, there is an increasing correlational ability and a decreasing degeneracy of TI's. Four types of newly developed real-number LOVI's are reviewed: (i) Information-based LOVI's obtained from topological distances to all other graph vertexes; (ii) Solutions of linear equation systems obtained from triplets consisting of a matrix (adjacency or distance matrix) and two column vectors; (iii) LOVI's based on eigenvalues and eigenvectors of the two above matrices; (iv) Regressive distance sums and regressive vertex degrees, which are the corresponding LOVI's (distance sums or vertex degrees) augmented slightly by all other vertexes, whose contributions decrease with increasing distance. When the LOVI's are based on topological distances, it is easy to include information on the presence and location of multiple bonds and/or heteroatoms. All LOVI's are validated by intramolecular comparison within various alkanes, and all TI's are validated both by intermolecular comparison within series of isomeric alkanes and by correlations with physicochemical properties.

INTRODUCTION: LOCAL VERTEX INVARIANTS, FIRST- AND SECOND-GENERATION TOPOLOGICAL INDEXES

Topological indexes (TI's) are numbers associated with chemical structures for the purpose of quantitative structure–activity (or structure–property) relationships. Their use in drug design, along with such parameters as Hammett or Taft constants or hydrophobic parameters, has gained acceptance during the last few years. They have been reviewed by various authors.[1-11]

Two characteristics of TI's have to be emphasized. First, until now no single TI could be used universally in optimal correlations; therefore, more than 100 TI's are in existence. Second, whereas any constitution of a normal covalent compound may be converted into various TI's according to definite procedures, the reverse is not true: retrieval of the chemical constitution from the numerical value of the TI is not possible. One of the reasons for the latter fact is the degeneracy of TI's, i.e., more than one chemical constitution may give rise to the same value for the TI. It is desirable to have TI's with good correlational ability and with low degeneracy.

Topological indexes are obtained from the hydrogen-suppressed (constitutional) graph of the molecule on the basis of local graph invariants. These are usually local vertex invariants (LOVI's), but one may also conceive bond (edge) invariants. Such invariants are numbers associated with vertexes symbolizing atoms (or correspondingly with edges symbolizing covalent bonds) in such a way as to be independent of arbitrary vertex labelings.

The number of vertexes in the hydrogen-suppressed graph will be denoted by n. Acyclic graphs are called trees; when all vertexes have degrees lower than or equal to 4, the corresponding trees are called 4-trees. The degree of a vertex is the number of edges meeting at that vertex.

Examples of simple LOVI's are the vertex degrees D_i; these are familiar to organic chemists because they translate arithmetically the nature of carbon atoms in hydrocarbons: primary, secondary, tertiary, or quaternary carbon atoms correspond to vertex degrees 1, 2, 3, and 4, respectively. A different derivation of the vertex degree is by summing over rows or columns the entries a_{ij} in the adjacency matrix A corresponding to the constitutional graph ($a_{ij} = 1$ for adjacent vertices, and 0 otherwise).

Other examples for LOVI's are the distance sums S_i, which represent the sums over rows or columns in the distance matrix D. Entries d_{ij} in this matrix are the topological distances between vertexes, i.e., the number of edges along the shortest path between vertexes i and j.

Both these examples are integer number LOVI's; normally, D_i values are 1, 2, 3, and 4. By contrast, values for S_i may be indefinitely large for indefinitely large molecules. Other integer LOVI's are obtained from 1-vertex fragments.[11]

First-generation TI's were integer numbers obtained by operations involving integer-number LOVI's; operations were simple ("bookkeeping operations"), involving only one vertex at a time. Thus, the very first TI was the Wiener Index, which represents the half-sum of all S_i values in matrix D.[12] Other such first-generation TI's are the following

Zagreb group TI[13] denoted by M_1:

$$M_1 = \sum_i D_i^2$$

Gordon–Scantlebury index denoted by N_2:[14]

$$N_2 = \sum_i D_i(D_i - 1)/2$$

quadratic index denoted by Q:[15]

$$Q = [\sum_i (i^2 - 2i)D_i + 2]/2$$

the last three indexes are interrelated:

$$Q = N_2 - n + 2 = 3 - 2n + (M_1/2)$$

$$M_1 = 2(N_2 + n - 1)$$

centric indexes denoted by B and C[15]

Hosoya's index denoted by Z[16]

a number of independent vertex sets (Merrifield–Simmon's σ index)[17]

All these first-generation TI's have a high degeneracy.

Second-generation TI's are real numbers obtained from LOVI's, which are integer numbers, by applying sophisticated ("structural") operations involving more than one vertex at a time. Examples are Randić's molecular connectivity[18]

$$\chi = \sum_{\text{edges ij}} (D_i D_j)^{-1/2}$$

0095-2338/92/1632-0023$03.00/0 © 1992 American Chemical Society

Table I. Topological Indexes (TI's) Based on Local Vertex Invariants (LOVI's)

TI generation	LOVI	operation	TI	degeneracy of TI
first	integer no.	bookkeeping	integer no.	high
second	integer or rational no.	structural	real no.	medium or low
third[a]	real no.	both types	real no.	low

[a] Includes TI's based on matrices whose entries are real numbers.

or Kier and Hall's extended connectivities for paths of certain lengths greater than one (edges are paths of length 1; when m nonrepeating edges form a continuous path between two vertexes, the length of this path is m)[5]

$$^{m}\chi = \sum_{\text{path}} (D_i D_j \ldots D_k)^{-1/2}$$

Other examples of 2nd generation TI's are the mean-square distance[4] between all vertexes in the graph

$$D^{(2)} = \frac{1}{n(n-1)} (\sum_{i,j} d_{ij}^2)^{-1/2}$$

and the information-theoretic indexes obtained by applying Shannon's formula, e.g., to all distances d_{ij} in the distance matrix.[6,19]

A last example of a second-generation TI is the J index (average distance sum connectivity), which reflects the "topological shape" of the constitutional graph (q denotes the number of edges, and μ is the cyclomatic number of the graph, equal to $q - n + 1$)[20-24]

$$J = \frac{q}{\mu + 1} \sum_{\text{edges } ij} (S_i S_j)^{-1/2}$$

SPECIAL PROBLEMS: THE PRESENCE OF MULTIPLE BONDS AND/OR HETEROATOMS

There exist simple algorithms for obtaining the distance matrix of a graph from its adjacency matrix.[25] Unlike the adjacency matrix, whose information is limited to account for neighboring or nonneighboring atoms, the distance matrix can store more than its initial information content, which is the topological distance between vertexes (an integer number). Thus, the presence of multiple bonds can be indicated by including for such bonds a rational number, namely, the inverse of bond order: 1/2, 1/3, or 2/3 for double, triple, or aromatic bonds, respectively. In this case, the graph invariants S_i (distance sums) will no longer be integer numbers but rational numbers.

Also, if atom i is a heteroatom (different from carbon or hydrogen), then some characteristic value for this atom can be used for multiplying the corresponding vertex invariant or the entries in the ith row and column. For instance, if one wishes to express the chemical identity of atom i by its atomic number Z_i, then one multiplies the corresponding entries by some function of Z_i, such as $6/Z_i$ as advocated by Barysz et al.[26] For correlating with physicochemical properties, however, other characteristics are better suited: for instance, covalent radii or electronegativities are characteristics that present a periodicity with respect to increasing Z_i values. Coefficients expressing such periodical variations have been devised and used in connection with the topological index J; the coefficient for carbon atoms was taken conventionally to be equal to 1.[24]

STRATEGIES FOR DEVISING REAL-NUMBER LOVI'S

Table I summarizes the characteristics of LOVI's and TI's as well as those of the operations converting LOVI's into first- and second-generation TI's. It may be seen that the obvious strategy for devising TI's with lower degeneracy is to base them on real-number LOVI's. Indeed, in the last few years several such LOVI's and TI's have been proposed, and some of them will be considered in detail here.

Owing to its superior ability to code for the presence of multiple bonding and/or heteroatoms, the distance matrix will be preferred over the adjacency matrix. Both these matrices specify a graph uniquely (up to isomorphism). No reduced form of these matrices has yet been devised which would be able to characterize a graph up to isomorphism.

A feature that must be emphasized is the following one: we wish to obtain LOVI's characterizing vertexes individually, starting from the typical form for inputting information about the graph structure, namely via its A or D matrix. It should be recalled that entries in such matrices represent binary relationships between two vertexes, yet we wish to obtain a LOVI end-product referring to one vertex only. Then several solutions emerge: all other vertexes must be either summed (as in the distance sum acting as LOVI) or averaged according to the number of atoms, edges, or cycles. It will be shown below that other types of matrices may also be imagined and used for this purpose.

After a strategy for devising LOVI's has been implemented, it must be tested in order to see if the resulting LOVI's have reasonable ranges of values, and if they behave reasonably in comparing related sets of molecules. We consider that the best validation tests for LOVI's are the intramolecular comparison for linear and branched alkane chains and the intermolecular comparison within sets of isomeric alkanes. We have consistently applied these tests to the LOVI's we devised using all alkanes with 2–9 carbon atoms.

When discussing degenerate TI's, one must distinguish between assignment degeneracy and operational degeneracy;[27] the former degeneracy arises when LOVI's of nonequivalent vertexes are equal (for instance, all vertex degrees for secondary carbon atoms are $D_i = 2$; therefore, n-pentane has assignment degeneracy for such carbon atoms when the LOVI's are vertex degrees); the latter degeneracy arises when the same TI value is obtained for different graphs due to the use of an operation that was too indiscriminate.

REAL-NUMBER LOVI'S

1. Information-Based LOVI's. The first authors who used Shannon's formula for devising LOVI's were Basak, Raychaudhury, Klopman, and co-workers for molecular graphs containing all atoms, including hydrogens.[28-32]

We proposed[33] four new LOVI's based on the distance vectors for each vertex. Let the distance vector for vertex i consist of g_j times each distance $j = 1, 2, \ldots, d_{i\,max}$. By definition, $S_i \sum_j j g_j$. Then the local information and the mean local information on the magnitude of distances for vertex i are, respectively

$$v_i = S_i \log_2 S_i - u_i$$

$$u_i = -\sum_j \frac{j g_j}{S_i} \log_2 \frac{j}{S_i}$$

The extended local information and the mean extended local information on the magnitude of distances are, respectively

$$x_i = S_i \log_2 S_i - y_i$$

$$y_i = \sum_j j g_j \log_2 j$$

Since in all these distance vectors one may include data on the presence of multiple bonds and/or heteroatoms, these information-theoretic LOVI's may easily code for such chemically relevant systems.

The smallest 4-tree which presents assignment degeneracy (originating in the fact that it has nonequivalent vertexes possessing identical distance vectors) is the molecular graph

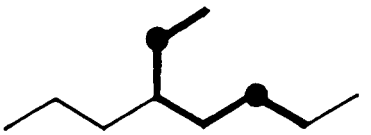


Figure 1. Hydrogen-depleted graph of the decane isomer for which the marked vertexes have the same distance vector, i.e., the same distribution of distances to all other graph vertexes.

Table II. Two Smallest Identity Trees with Information-Based LOVI's

	3-methylhexane				3-methylheptane			
vertex	U_i	v_i	x_i	y_i	u_i	v_i	x_i	y_i
1	2.441	72.62	43.94	31.12	2.642	107.40	63.41	43.63
2	2.412	45.69	31.35	16.75	2.594	72.46	46.69	28.36
3	2.446	30.77	24.46	8.75	2.611	50.69	36.55	16.76
4	2.482	35.57	27.30	10.75	2.699	50.60	37.79	15.51
5	2.414	50.89	33.79	19.51	2.656	61.34	42.49	21.51
6	2.439	78.27	46.35	34.36	2.604	83.83	52.07	34.36
7	2.473	56.13	37.09	21.51	2.643	119.57	68.73	53.48
8					2.666	83.77	53.32	33.12

of 4-ethyloctane (Figure 1): the two marked vertexes have the same distances to all other vertexes, namely, 1,1,2,2,3,3,4,4,5. By adding pairwise new vertexes in symmetrical positions along the marked path of length 5 (Figure 1), one may obtain new graphs with degenerate distance vectors, having 12, 14, etc. vertexes.

Table II illustrates values for the new LOVI's in the smallest identity tree corresponding to the molecular graph of 3-methylhexane. It may be seen that v_i, x_i, and y_i have wider ranges than u_i. All three former LOVI's have larger values for the endpoints (vertexes of degree 1) numbered 1, 6, and 7, and smaller values for vertexes of higher degree, especially the ternary vertex numbered 3. On comparing vertexes of equal degree in the same molecule, one sees that the three former LOVI's have higher values for vertexes farther from the center or from a vertex of high degree. All these three LOVI's intramolecularly order the carbon atoms of 3-methylhexane exactly like the HOC (hierarchically extended ordered connectivities) ordering.[34]

However, these three LOVI's differ in their intramolecular ordering of vertexes for the next higher homologous identity tree, corresponding to 3-methylheptane. Indeed, as seen from Table II, x_i values are influenced more by the vertex degree than by the centricity of the vertex, while for v_i and y_i values the opposite holds.

By applying a formula analogous to that employed for J to the new LOVI's, one obtains new TI's denoted by U, V, X, and Y. It may be seen from Table III that the intermolecular orderings according to V and X are practically identical with one another (only one inversion is observed for 4-methylheptane and 2,5-dimethylhexane), but Y and especially U lead to different orderings of alkanes, which do not agree with the ideal ordering advocated by Bertz.[35] Again, the range in U values (as for u_i values) is not convenient because it increases to infinity in the series of n-alkanes and it has narrow ranges for each isomer series.

2. Solutions of Linear Equation Systems. A method for inputting topological or chemical information and resulting in useful real-number LOVI's is to combine the $n \times n$ matrix (preferably the distance matrix for reasons explained earlier) with two column $l \times n$ vectors (which may be identical or distinct): one for the diagonal elements of the matrix, and the other for the free term. Thus a system of n linear equations is obtained whose solutions are the LOVI's.[27]

Table III. Information-Based TI's for Octane Isomers Ordered According to Increasing X Values[a]

isomer	U	V	X	Y
C8	18.80	0.6170	0.9707	1.5743
2M-C7	18.75	0.6803	1.0380	1.8121
3M-C7	18.60	0.7336	1.0855	2.0490
4M-C7	18.52	0.7557	1.1026	2.1650
25MM-C6	18.63	0.7551	1.1116	2.1249
3E-C6	18.34	0.8144	1.1502	2.4808
24MM-C6	18.45	0.8202	1.1659	2.4512
22MM-C6	18.57	0.8255	1.1791	2.4356
23MM-C6	18.40	0.8492	1.1898	2.6100
34MM-C6	18.28	0.8972	1.2276	2.8870
3E-2M-C5	18.18	0.9227	1.2445	3.0824
33MM-C6	18.29	0.9310	1.2591	3.0706
224MMM-C5	18.36	0.9319	1.2694	3.0046
234MMM-C5	18.21	0.9638	1.2868	3.2648
3E-3M-C5	18.08	1.0191	1.3217	3.7505
223MMM-C5	18.16	1.0306	1.3428	3.6700
233MMM-C5	18.08	1.0684	1.3678	4.0126
2233MMMM-C4	17.97	1.2012	1.4745	4.9756

[a] Nomenclature according to IUPAC rules; M = methyl, E = ethyl.

Some of the column vectors that were tested included chemical data (atomic number Z_i, electronegativity, or covalent radius), topological data (vertex degree denoted in this section by V, distance sum S, total number of non-hydrogen atoms or vertexes, denoted in this section by N), or constants (e.g., 1; in some cases the constant was the number N of vertexes or for hydrocarbons it was the atomic number 6 for carbon atoms).

By testing various triplets (matrix–vector–vector, specified by their symbols) it was found[27,36] that the resulting LOVI's may increase or decrease from the periphery toward the center of the graph and that in linear chains this variation may be either monotonous or alternating. Since the last type of variation has very limited applications (it occurs occasionally only for triplets originating with the adjacency matrix), it will not be illustrated here. We present in Table IV examples of LOVI's resulting from various triplets for the molecular graphs of n-pentane and isopentane (2-methylbutane) having Z_i = 6 and N = 5, with the IUPAC vertex numbering.

It may be seen from Table IV that LOVI's can be positive or negative (the latter case occurs for the DSV triplet). Another remark concerns the fact that when the free term is a constant, the resulting LOVI's are proportional to the value of this constant. This is why in Table IV the only constant value of the free term was taken to be N.

Simple summation of LOVI's for the whole graph may lead to interesting TI's. In the case of the AZV triplet, a good quadratic equation was found to correlate the boiling points (°C) of 50 alkanes with 2–9 carbon atoms: $BP = -30.23-(TI_{AZV})^2 + 203.25(TI_{AZV}) - 142.89$.

The correlation coefficient is r = 0.9983, and the standard deviation is s = 2.9°.

However, operational degeneracy may appear, as was observed for the topological index TI_{AZV} obtained for two dodecane isomers (Figure 2).

3. LOVI's Based on Graph Eigenvalues or Eigenvectors. If the eigenvalues of matrices A or D are calculated, it was found earlier[37] that the lowest eigenvalue $E(A)$ of trees may serve as a TI for alkanes, and algebraic formulas were found for calculating this eigenvalue for special types of trees: paths and stars. Randić proposed to use the first eigenvector for the purpose of vertex labeling,[39] and in the Schultz paper, eigenvalues were discussed as possible TI's for various graphs.[38]

We examined the eigenvectors corresponding to the lowest (largest negative) eigenvalue $E(A)$ of all alkanes with up to 8 vertexes.[40] It was observed that assignment degeneracy is obtained for vertexes of 2-methylhexane, 2,4-dimethylpentane, 4-methylheptane, 2,5-dimethylhexane, and 2,3-dimethyl-

Table IV. LOVI's for Isopentane and *n*-Pentane as Solutions of Linear Equation Systems Derived from Various Triplets[a]

	2-methylbutane				*n*-pentane		
triplet	1 and 5	2	3	4	1 and 5	2 and 4	3
DN^2N	0.1508	0.1699	0.1633	0.1437	0.1392	0.1585	0.1651
DSN	0.2825	0.7272	0.4988	0.1502	0.1520	0.4223	0.5912
DSV	0.0096	0.5673	0.2404	−0.0481	−0.0270	0.2027	0.2838
increase							
ASV	0.0587	0.5304	0.2307	0.0855	0.0761	0.2386	0.2538
ASN	0.5433	0.6540	0.6437	0.4840	0.4442	0.5584	0.6472
DN^2S	0.2459	0.1505	0.1834	0.2816	0.3041	0.2077	0.1747
AZS	1.2860	0.2842	0.7227	1.3796	1.5353	0.7879	0.7374
decrease							
ANS	1.5579	0.2106	0.8311	1.6338	1.8273	0.8636	0.8546

[a] LOVI's may increase or decrease toward center.

Table V. LOVI's Obtained as Eigenvectors for the Lowest Eigenvalue $E(A)$ or $E(D)$ of the Two Smallest Identity Trees from Table II

	3-methylhexane		3-methylheptane	
vertex	eigenvector for $E(A)$	eigenvector for $E(D)$	eigenvector for $E(A)$	eigenvector for $E(D)$
1	0.2073	0.4561	0.1927	0.4296
2	0.4083	0.3425	0.3834	0.3324
3	0.5968	0.2750	0.5698	0.2688
4	0.4642	0.2978	0.4635	0.2700
5	0.3176	0.3607	0.3592	0.2986
6	0.1612	0.4722	0.2369	0.3574
7	0.3030	0.3966	0.1191	0.4523
8			0.2865	0.3719

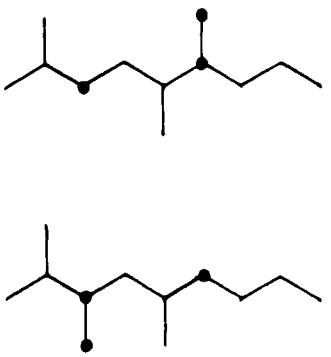


Figure 2. Two isomers of dodecane for which the LOVI values based on the AZV triplet are pairwise equal for the marked vertexes of degrees 1, 2, and 3.

hexane. Lower values of the eigenvectors correspond to vertexes of lower degree, farther from the center or from a vertex of high degree, as seen in Table V for the two smallest identity trees.

More interesting are eigenvectors corresponding to the unique negative eigenvalue of the distance matrix, $E(D)$.[40] No assignment degeneracy was observed for alkanes with up to 8 vertexes. Higher values of eigenvectors correspond to vertexes of lower degree, farther from the center or from a vertex of high degree, as seen in Table V. Chemical information (multiple bonding, heteroatoms) can be input into the distance matrix, hence into the corresponding eigenvectors functioning as LOVI's.

4. Regressive Distance Sums and Regressive Vertex Degrees. The vertex degrees and the distance sums have been the LOVI's on which most of the first- and second-generation TI's were based. Whereas the vertex degree D_j gives indications about the immediate neighborhood of a vertex j; the distance sum S_j places the major emphasis on the more remote vertexes which make the major contribution to S_j.

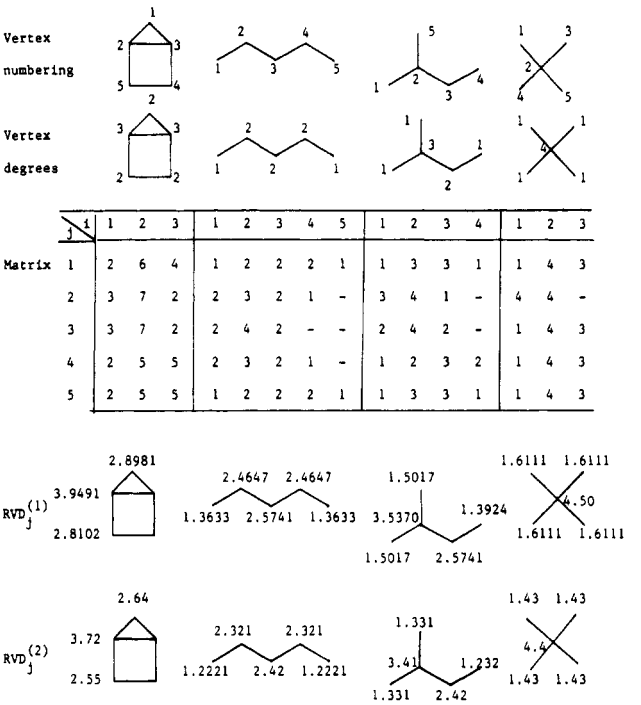
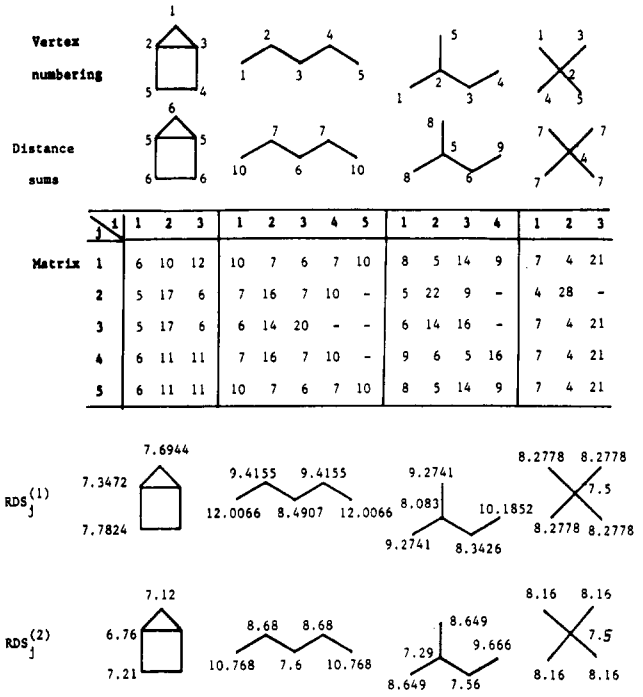


Figure 3. Four graphs with five vertexes with an arbitrary vertex numbering, followed by the vertex degrees, the matrices for the derivation of regressive vertex degrees, and the corresponding values of these LOVI's according to formula 1 or 2.

Regressive vertex degrees are slightly increased vertex degrees, so as to take into account the contribution of more remote vertexes, attenuated by increasing distances. Correspondingly, regressive distance sums are similarly increased distance sums. In both cases, the resulting LOVI's are real numbers, which have far less assignment degeneracy than D_j or S_j.

For obtaining these new LOVI's, one starts by constructing from the integer LOVI's D_j and S_j, respectively, new matrices reflecting the various shells $i = 1, 2, \ldots$ at increasing distance around each vertex. For four graphs with 5 vertexes, we shall exemplify in Figure 3 the regressive vertex degrees and in Figure 4 the regressive distance sums.

In the i = 1st column, the entries y_{ij} are the LOVI's for the vertex j itself. In the i = 2nd column, one sums LOVI's for all vertexes in the shell surrounding the vertex j and directly connected to it (i.e., at distance 1); in the ith column, the entries y_{ij} are sums of LOVI's in the shell at distance $i - 1$ from vertex j. The maximum distance between two vertexes of a graph is called the diameter of that graph, and this is also the number of columns in the new matrices.



Matrix	i \ j	1	2	3	1	2	3	4	5	1	2	3	4	1	2	3
	1	6	10	12	10	7	6	7	10	8	5	14	9	7	4	21
	2	5	17	6	7	16	7	10	-	5	22	9	-	4	28	-
	3	5	17	6	6	14	20	-	-	6	14	16	-	7	4	21
	4	6	11	11	7	16	7	10	-	9	6	5	16	7	4	21
	5	6	11	11	10	7	6	7	10	8	5	14	9	7	4	21

Figure 4. Same four graphs as in Figure 3 with the arbitrary vertex numbering, the distance sums, the matrices for the derivation of regressive distance sums, and the corresponding values for these LOVI's according to formula 1 or 2.

It will be observed that in these matrices, all sums over rows are equal to $2w$ for the RDS (regressive distance sums) matrix, and to $2q$ for the RVD (regressive vertex degrees) matrix (w is Wiener's TI and q is the number of edges in the graph).

For the next step, two possibilities exist according to formulas 1 and 2:

$$\mathrm{LOVI}_j^{(1)} = \sum_i i^{-3} y_{ij} \quad (1)$$

$$\mathrm{LOVI}_j^{(2)} = \sum_i 10^{1-i} y_{ij} \quad (2)$$

The reciprocal cubic regressive formula 1 leads to slightly higher values than the decimal regressive formula 2 because in the former case the ith layer entries (i = 2, 3, 4, etc.) are divided by 8, 27, 64, etc., while in the latter case the divisors are 10, 100, 1000, etc.

With the new LOVI's denoted here by RVD_j and RDS_j, one can proceed, via various operations, to obtain new TI's. In a recently published paper,[41] we indicated how to input information on bond multiplicity or on the presence of heteroatoms into RVD_j-type indexes. In that paper, intra- and intermolecular validation of the new LOVI's was presented. New TI's obtained by simple summation of RVD_j possess low degeneracy and good correlational ability, e.g., for boiling points of 35 alkanes (n = 5–8) we obtained a correlation coefficient of r = 0.985. Intermolecular ordering of alkanes, alkenes, or haloalkanes is reasonable. Algebraic recursive formulas for the new TI's were displayed.[41]

It is easier, via topological distances as shown in the preceding sections, to introduce into RDS_j-type LOVI's information on the presence of multiple bonds and/or heteroatoms.[42]

CONCLUSIONS

In the preceding sections we have presented four new LOVI's which are real numbers and which were validated by intra- and intermolecular comparisons using the series of alkane molecules with 2–9 carbon atoms as test series. All these new LOVI's can easily, by means of adjusted topological distances, account for the presence of multiple bonds and/or heteroatoms in the molecules under study. On the basis of the new LOVI's, by various operations which may be more or less sophisticated, one can obtain new TI's that also were tested in a few correlations. In addition to the four types of LOVI's discussed above, we have also developed a fifth type of real-number LOVI's, called distance-enhanced exponential connectivities; they vary in the range from 0 to 1, and their summation over the whole graph yields new TI's that correlate satisfactorily with boiling points and critical pressures of alkanes with 4–9 carbon atoms.[43] However, for these LOVI's that do not depend explicitly on distances, it is less easy to include information on multiple bonds and heteroatoms.

Third-generation TI's will undoubtedly emerge from real-number LOVI's; such TI's can be designed to have extremely low or, possibly, no degeneracy, and at the same time to possess good correlational ability.

ACKNOWLEDGMENT

This summary of work carried out during the last years has benefitted from cooperation with several co-workers from Bucharest, Cluj, and Timisoara (Romania), listed alphabetically below: T.-S. Balaban, C. Catana, D. Ciubotariu, M. Diudea, P. Filip, O. Ivanciuc, M. Medeleanu, and O. Minailiuc. Thanks are expressed to the Computer Division of the ACS and to Professor J. R. Dias for the support which enabled me to present this paper at the 202nd American Chemical Society Meeting in New York in August 1991.

REFERENCES AND NOTES

(1) Balaban, A. T.; Chiriac, A.; Motoc, I.; Simon, Z. *Steric Fit in QSAR*; Lecture Notes in Chemistry No. 15; Springer: Berlin, 1980.
(2) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Top. Curr. Chem.* **1983**, *114*, 21.
(3) Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334.
(4) Balaban, A. T. *Pure Appl. Chem.* **1983**, *55*, 199.
(5) (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976. (b) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Wiley: New York, 1986.
(6) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, 1983.
(7) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. 2, Chapter 4, p 105.
(8) (a) Rouvray, D. H. *Math. Chem.* **1975**, *1*, 125. (b) Rouvray, D. H. *J. Comput. Chem.* **1987**, *8*, 470. (c) Rouvray, D. H. *Discrete Appl. Math.* **1988**, *19*, 317. (d) Rouvray, D. H. *Sci. Am.* **1986**, *255* (Sept) 36.
(9) Stankevich, M. I.; Stankevich, I. V.; Zefirov, N. S. *Usp. Khim.* **1988**, *57*, 337.
(10) Sabljic, A.; Trinajstić, N. *Acta Pharm. Jugosl.* **1982**, *31*, 189.
(11) Mekenyan, O.; Bonchev, D.; Balaban, A. T. *J. Math. Chem.* **1988**, *2*, 347.
(12) (a) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17, 2636. (b) Wiener, H. *J. Chem. Phys.* **1947**, *15*, 766. (c) Wiener, H. *J. Phys. Chem.* **1948**, *52*, 425, 1082.
(13) Gutman, I.; Ruscić, B.; Trinajstić, N.; Wilcox, C. F., Jr. *J. Chem. Phys.* **1975**, *62*, 3399.
(14) Gordon, M.; Scantlebury, G. R. *Trans. Faraday Soc.* **1964**, *60*, 605.
(15) (a) Balaban, A. T. *Theor. Chim. Acta (Berlin)* **1979**, *53*, 355. (b) Bonchev, D.; Balaban, A. T.; Randić, M. *Int. J. Quantum Chem.* **1981**, *19*, 61.
(16) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(17) Merrifield, R. E.; Simmons, H. E. *Topological Methods in Chemistry*; Wiley: New York, 1989.
(18) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(19) Bonchev, D.; Trinajstić, N. *J. Chem. Phys.* **1977**, *67*, 4517.
(20) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
(21) Balaban, A. T.; Filip, P. *Math. Chem.* **1984**, *16*, 163.
(22) Balaban, A. T.; Ivanciuc, O. MATH/CHEM/COMP 1988. *Stud. Phys. Theor. Chem.* **1989**, *63*, 193.
(23) Balaban, A. T.; Ionescu-Pallas, N.; Balaban, T. S. *Math. Chem.* **1984**, *16*, 163.
(24) Balaban, A. T. *Math. Chem.* **1986**, *21*, 115.
(25) Barysz, M.; Plavsić, D.; Trinajstić, N. *Math. Chem.* **1986**, *19*, 89.
(26) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstić, N. Chemical Applications of Topology and Graph Theory. *Stud. Phys. Theor. Chem.* **1983**, *28*, 222.
(27) Filip, P. A.; Balaban, T. S.; Balaban, A. T. *J. Math. Chem.* **1987**, *1*, 61.
(28) (a) Roy, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, A. A. *Arzneimitt.-Forsch.* **1982**, *32*, 322. (b) *Ibid.* **1983**, *33*, 352.

(29) Roy, A. B.; Raychaudhury, C.; Ghosh, J. J.; Roy, S. K.; Basak, S. C. In Quantitative Approaches to Drug Design; Dearden, J. C., Ed.; Elsevier: Amsterdam, 1983; p 75.
(30) Raychaudhury, C.; Roy, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1989**, *5*, 581.
(31) Klopman, G.; Raychaudhury, C.; Henderson, R. V. *Math. Comput. Modelling* **1988**, *11*, 635.
(32) Klopman, G.; Raychaudhury, C. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 12.
(33) Balaban, A. T.; Balaban, T. S. *J. Math. Chem.* **1991**, *8*, 383.
(34) Balaban, A. T.; Mekenyan, O.; Bonchev, D. *J. Comput. Chem.* **1985**, *6*, 538.
(35) Bertz, S. H. *Discrete Appl. Math.* **1988**, *19*, 65.
(36) Balaban, A. T. Graph Theory and Topology in Chemistry. *Study Phys. Theor. Chem.* **1987**, *51*, 159.
(37) Lovasz, L.; Pelikan, J. *Period. Math. Hung.* **1973**, *3*, 175.
(38) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 27.
(39) Randić, M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105.
(40) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517.
(41) Diudea, M.; Minailiuc, O.; Balaban, A. T. *J. Comput. Chem.* **1991**, *12*, 527.
(42) Balaban, A. T.; Diudea, M., unpublished results.
(43) Balaban, A. T.; Catana, C. *J. Comput. Chem.*, in press.

Comparative Study of Molecular Descriptors Derived from the Distance Matrix

ZLATKO MIHALIĆ

Faculty of Science and Mathematics, The University of Zagreb, 41000 Zagreb, The Republic of Croatia, Yugoslavia

SONJA NIKOLIĆ and NENAD TRINAJSTIĆ*

The Rugjer Bošković Institute, 41001 Zagreb, The Republic of Croatia, Yugoslavia

Received August 16, 1991

A comparative study of 10 distance indices derived from the distance matrix either in the graph-theoretical (topological) form or in the geometric (topographic) form is carried out. They are partitioned into five topological indices and five topographic indices. The adjective topological or topographic indicates which matrix each family of distance indices has originated from. All five topological indices have been known in the literature, while four out of five topographic indices are introduced in this work. Only the 3-D Wiener number has been proposed earlier. Most of distance indices are found to be intercorrelated, i.e., they contain similar structural information. Only 2-J and 3-J indices did not intercorrelate with any other distance index but themselves. The three most accurate structure–property models for predicting boiling points of alkanes are based on the connectivity index χ, its variant χ' ($= N\chi$), and on the topological distance index 2-TI. It is unclear at present why the 3-D distance indices have produced inferior structure–boiling point models in comparison with models based on the 2-D distance indices and connectivity indices.

INTRODUCTION

The distance matrix appears to be a convenient source for deriving molecular descriptors.[1-4] This matrix can be given in two forms,[5] i.e., as the graph-theoretical (topological) distance matrix[6-8a] and the geometric (topographic) distance matrix.[9-11] Molecular descriptors that can be derived from the topological distance matrix belong to the class of topological indices,[12] while those that can be obtained from the topographic distance matrix are topographic indices.[13] Topological indices and topographic indices represent a subgroup of molecular descriptors,[14] i.e., they are used to characterize the constitution and the configuration of a molecule by a single number. In order to simplify the presentation, topological and topographic distance-matrix-related indices will be called by a common term, distance indices.

There are a number of distance indices available in the literature.[15a] Most of them will be discussed here. We will not, however, consider the information-theoretic distance indices.[16-18] The current interest in distance indices as well as in other molecular descriptors is stimulated by their use in the nonempirical[19] quantitative structure–property relationships (QSPR)[2] and quantitative structure–activity relationships (QSAR).[20]

The present work is motivated by an aim to compare distance indices in two ways. First, we will try to answer the question as to the extent the distance indices are intercorrelated. In other words, we will investigate to what degree they contain the same type of constitutional and geometric information. Second, we will examine how the distance indices perform in a given structure–property correlation. For the latter purpose, the boiling points of the first 150 alkanes are selected. Each distance index considered will be used to build a QSPR model for boiling points, and the quality of each model will be judged on the basis of its statistical characteristics. In this study will also be included the connectivity index[21] and the most-used topological index in QSPR and QSAR to date.[22] Hence, the QSPR model based on the connectivity index (or on one of its variants) will be used as a standard against which the QSPR models based on distance indices will be measured.

Throughout the article we will use the graph-theoretical language.[6] Chemical structures will be represented by hydrogen-depleted graphs in the standard manner.[8b,15b]

The structure of this article will be as follows. The second section will contain the definitions of the distance indices based on the graph-theoretical distance matrix. In the third section, the definitions of the distance indices based on the geometric distance matrix will be presented. The fourth section will include the definition of the connectivity index and one of its variants. The intercorrelation of distance indices will be discussed in the fifth section. In the sixth section will be given all QSPR models that were considered, their statistical

0095-2338/92/1632-0028$03.00/0 © 1992 American Chemical Society