Theor Chim Acta (1986) 70: 365-378

Theoretica
Chimica Acta

© Springer-Verlag 1986

Discrimination and ordering of chemical structures by the number of walks*

Marko Razinger

Boris Kidrič Institute of Chemistry, POB 30, 61115 Ljubljana, Yugoslavia

(Received February 14, revised August 28/Accepted September 3, 1986)

The use of the number of walks for the discrimination of graphs representing chemical structures is discussed. A highly selective graph-theoretical index based on the number of walks is defined. The index values were calculated for 661 acyclic and 376 cyclic structures. The selectivity of the new index is compared to that of some of the most selective previously defined indexes. The ordering of structures induced by the value of the index is also considered.

Key words: Graph theory — Number of walks — Topological index

Introduction

Graph theory has an important role in the definition and development of various numerical and alphanumerical descriptors of chemical structure used in documentation, structure - property correlation, and for other purposes [1]. A host of different graph invariants can be used for the discrimination and/or characterization of graphs representing chemical structures [2].

The number of walks is a graph invariant used in the past in many different ways for the characterization of molecular graphs. In this paper we consider as molecular graphs only graphs representing saturated hydrocarbon molecules depleted of hydrogen atoms, i.e. connected undirected graphs with vertex degree between 1 and 4. The term "walk" is used in its broadest sense [3] as representing

---

* Presented in part at the 7th International Conference on Computers in Chemical Research and Education, held in Garmisch-Partenkirchen, Germany, June 10-14, 1985

an alternating sequence of vertexes and edges beginning and ending with vertexes. Several attempts to characterize molecular graphs using smaller subgroups of all possible walks have been made. Let us cite as examples "self returning walks" (or closed walks) [4, 5], "unusual random walks" [6], and "self avoiding walks" (or paths) [7-9]. Still another example of the use of the number of walks can be found in the Chemical Abstracts Registry File, the largest existing structural data file (over 7 million structures in 1985 [1]). The canonical connection tables in this file are derived via Morgan's algorithm [10]. This algorithm is used for the canonical labelling of vertexes by means of their extended connectivities. The chemical community uses extended connectivity for the discrimination of chemical structures through CAS Online and related systems, taking advantage of the fact that the extended connectivity of order $\underline{i}$ of a given vertex is identical to the number of walks of length $\underline{i}$, starting from this vertex, as we demonstrated in a previous paper [11].

In this paper we propose a new numerical descriptor of chemical structure based only on the number of walks as the basic graph invariant. Its structural selectivity is equal to or better than that of the most discriminating descriptors defined until now.

Definition of the index

In the formulation of any numerical descriptor of chemical structure two steps can usually be distinguished. In the first step a graph invariant, characterizing selected structural features by its numerical value, is chosen. The second step consists in the choice of an algorithm for the transformation (i.e. compression) of information, obtained in the first step, into a single numerical value, called graph theoretical or sometimes topological index. This index can be used for characterization of a given molecular graph or at least for its discrimination from other nonisomorphic graphs.

As our choice of graph invariant used for numerical characterization of molecular graphs was the number of walks, we considered and compared the discriminating power of walks of different lengths: for each vertex in test structures the number of all walks up to the length N (N being the number of graph vertexes in all further text), and of the walks of length N only, were enumerated.

After having obtained the numerical value characterizing each and every vertex we proceeded to the choice of the transformation algorithm. One of the possibilities used for the definition of several indexes [12, 13] is a straightforward summation of numerical values, attributed to individual vertexes, over all vertexes. Such a simple method of numerical transformation causes considerable loss of information and hence results in indexes of low structural selectivity. For indexes of high discriminating power other algorithms must be employed.

Both most selective graph-theoretical indexes defined until now [13], namely Balaban's distance based index J [14] and Randić's molecular ID (15), implement the identical transformation algorithm. The molecular ID has the unsurpassed structural selectivity having no duplicate values in all alkane isomers up to structures with 16 carbon atoms and only two nonisomorphic structures with the same index value in C-15 and C-16 isomers [18].

Actually, this transformation algorithm was first used by Randić in the definition of the connectivity index $^1\chi$ [16]. From the aspect of structural selectivity this index was the best numerical structural descriptor ten years ago and even today it is not yet surpassed from the aspect of structure-property correlation performance [1, 12, 17].

The general form of Randić's algorithm, used in all above mentioned indexes, is the following:

$$\text{Index value} = \sum_{\text{edge}} (I_i \cdot I_j)^{-1/2} \tag{1}$$

where I_i and I_j stand for the selected invariant values of adjacent vertexes $\underset{\sim}{i}$ and $\underset{\sim}{j}$ respectively. The summation is done over all pairs of adjacent vertexes, i.e. over all edges in the graph.

In the graph-theoretical indexes mentioned above (Eq. (1)) the selected invariants were vertex degree for $^1\chi$ (16), averaged distance sum for J [14], and again the vertex degree (used in the calculation of weighted path numbers) for molecular ID [15].

The graph invariant, used in the present work, is the number of walks. Individual graph vertexes are characterized by its value and Randić's formula is employed subsequently to obtain a numerical value characterizing the graph as a whole.

Generation of index values

For the testing of the new index, the calculation of index values for a large number of structures was carried out using a program developed for automatic generation of numerical descriptors and described in detail elsewhere [11, 19]. The calculation of various descriptors is based on the adjacency matrix $\underset{\sim}{A}$ of the graph, representing a given chemical structure, and higher powers of $\underset{\sim}{A}$, i.e. $\underset{\sim}{A}^2, \underset{\sim}{A}^3, \ldots, \underset{\sim}{A}^N$. There are many graph-theoretical characteristics known to be deducible from these powers $\underset{\sim}{A}^k$ (3, 20), the number of walks among them. The value of element $(\underset{\sim}{A}^k)_{ij}$ of the $\underset{\sim}{k}$th power of adjacency matrix $\underset{\sim}{A}$ is equal to the number of all possible walks of length $\underset{\sim}{k}$, starting from vertex $\underset{\sim}{i}$ and ending in vertex $\underset{\sim}{j}$. Thus the sum of these elements $(\underset{\sim}{A}^k)_{ij}$ over all values of $\underset{\sim}{j}$ (i.e. the sum of the $\underset{\sim}{i}$th row) gives the number of all possible walks of length $\underset{\sim}{k}$ from vertex $\underset{\sim}{i}$ to all other vertexes in the graph - and this is just the invariant needed for the calculation of the index.

In this manner the number of all walks of length 1 to N (NW for total Number of Walks in further text) and the number of walks of length N only (LW for Longest considered Walk in further text) were calculated for each vertex in a considerable number of structures. Employing the following two formulas

$$I(NW) = \sum_{\text{adj}\cdot i,j} (NW_i \cdot NW_j)^{-1/2} \tag{2}$$

where

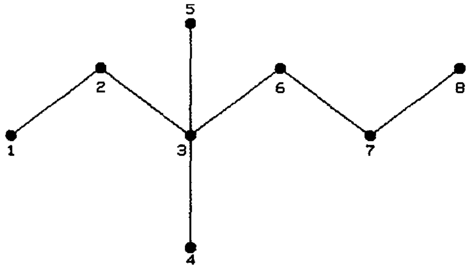
$$NW_i = \sum_{k=1}^{N} \sum_{j=1}^{N} (\underline{A}^k)_{ij}$$

$$I(LW) = \sum_{\mathrm{adj}\cdot i,j} (LW_i \cdot LW_j)^{-1/2} \tag{3}$$

where

$$LW_i = \sum_{j=1}^{N} (\underline{A}^N)_{ij}$$

the index values were calculated for all test structures. In example, the calculation of index values for 3,3 dimethyl hexane, using Eq. 3, is given in extenso in Fig. 1.



Power 1

	1	2	3	4	5	6	7	8	Sum $a(i,j)$ ($j=1,\ldots 8$)
1	0	1	0	0	0	0	0	0	1
2	1	0	1	0	0	0	0	0	2
3	0	1	0	1	1	1	0	0	4
4	0	0	1	0	0	0	0	0	1
5	0	0	1	0	0	0	0	0	1
6	0	0	1	0	0	0	1	0	2
7	0	0	0	0	0	1	0	1	2
8	0	0	0	0	0	0	1	0	1

Power 2

	1	2	3	4	5	6	7	8	Sum $a(i,j)$ ($j=1,\ldots 8$)
1	1	0	1	0	0	0	0	0	2
2	0	2	0	1	1	1	0	0	5
3	1	0	4	0	0	0	1	0	6
4	0	1	0	1	1	1	0	0	4
5	0	1	0	1	1	1	0	0	4
6	0	1	0	1	1	2	0	1	6
7	0	0	1	0	0	0	2	0	3
8	0	0	0	0	0	1	0	1	2

Fig. 1. Calculation of index $I(LW)$ for 3, 3 dimethyl hexane

⋮

Power 7

	1	2	3	4	5	6	7	8	Sum $a(i,j)$ ($j=1,\dots 8$)
1	0	30	0	23	23	30	0	7	113
2	30	0	106	0	0	0	37	0	173
3	0	106	0	83	83	113	0	30	415
4	23	0	83	0	0	0	30	0	136
5	23	0	83	0	0	0	30	0	136
6	30	0	113	0	0	0	46	0	189
7	0	37	0	30	30	46	0	16	159
8	7	0	30	0	0	0	16	0	53

Power 8

	1	2	3	4	5	6	7	8	Sum $a(i,j)$ ($j=1,\dots 8$)
1	30	0	106	0	0	0	37	0	173
2	0	136	0	106	106	143	0	37	528
3	106	0	385	0	0	0	143	0	634
4	0	106	0	83	83	113	0	30	415
5	0	106	0	83	83	113	0	30	415
6	0	143	0	113	113	159	0	46	574
7	37	0	143	0	0	0	62	0	242
8	0	37	0	30	30	46	0	16	159

$$\begin{aligned} I(LW) = {} & (LW_1 \times LW_2)^{-1/2} + (LW_2 \times LW_3)^{-1/2} + (LW_3 \times LW_4)^{-1/2} \\ & + (LW_3 \times LW_5)^{-1/2} + (LW_3 \times LW_6)^{-1/2} + (LW_6 \times LW_7)^{-1/2} \\ & + (LW_7 \times LW_8)^{-1/2} \end{aligned}$$

$$\begin{aligned} \underline{I(LW)} = {} & (173 \times 528)^{-1/2} + (528 \times 634)^{-1/2} + (634 \times 415)^{-1/2} \\ & + (634 \times 415)^{-1/2} + (634 \times 574)^{-1/2} + (574 \times 242)^{-1/2} \\ & + (242 \times 159)^{-1/2} = \underline{0.018374869} \end{aligned}$$

Fig. 1. (cont.)

The test data set of structures consisted of 661 trees (all alkane isomers from butane to dodecane) and 376 graphs containing cycles (all mono- and bicyclic structures with 4–8 vertexes [21]). The structural discriminating power of the new index was examined and compared to that of some other highly selective indexes. The ordering of structures according to their index value was scrutinized also.

Discussion

At the initial stage of the formulation of the new index, two different graph invariants were considered, namely *NW* and *LW* (see above). It is well known that the discrimination of nonequivalent vertexes by the extended connectivity (or the number of walks, since these two invariants are closely related) may not be successful, and also that the number of classes of nonequivalent vertexes in a graph may oscillate during the iterative calculation of extended connectivity

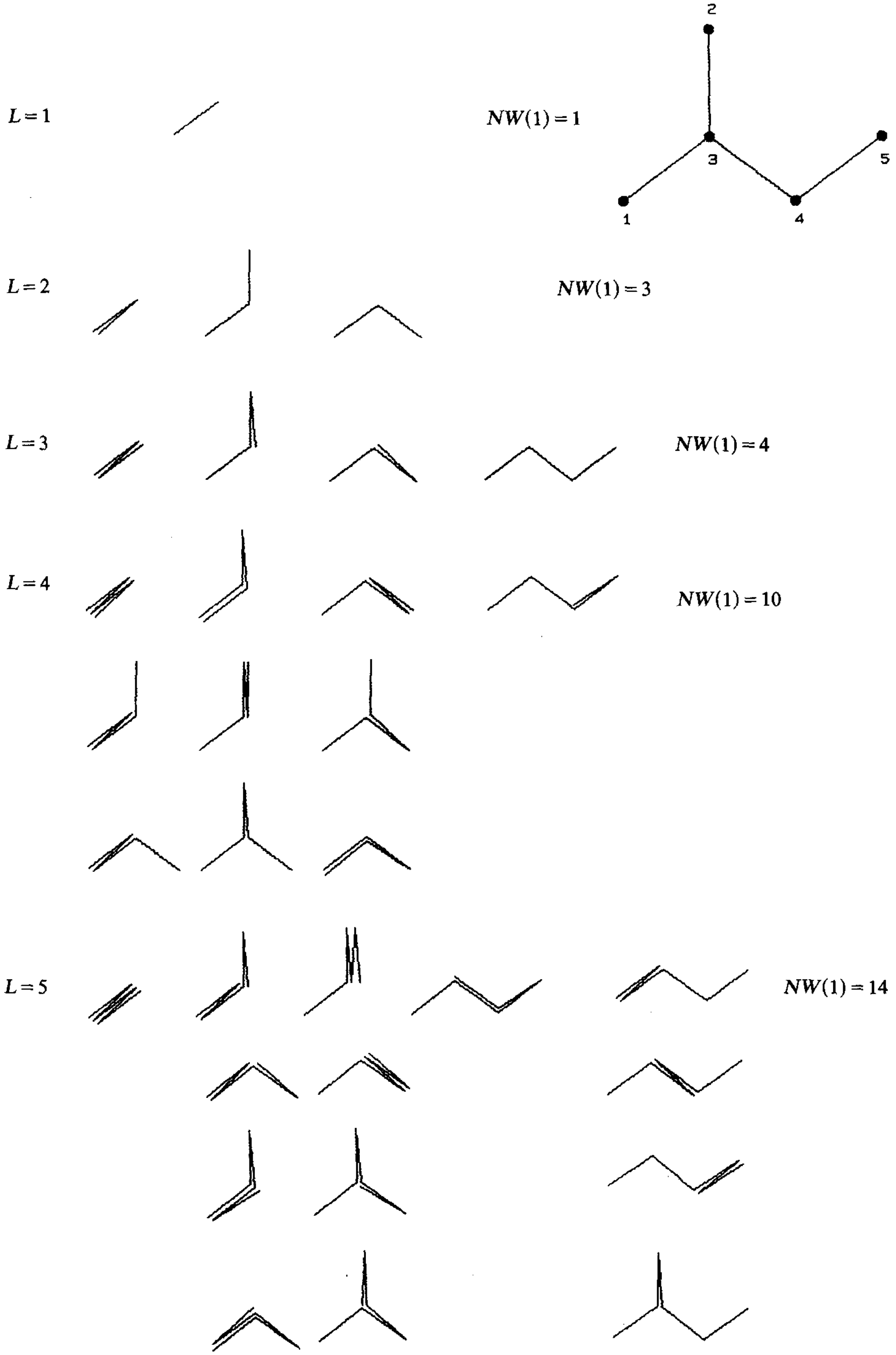


Fig. 2. All possible walks of length 1 to 5, starting in vertex No. 1 of the graph representing 2-methyl butane

[10, 22, 23]. Because of this, the number of walks of length $N-1$, $N-2$, etc. might in some cases discriminate nonequivalent vertexes better than the number of walks of length N does. Applying the above argument it was expected that, in general, the index based on NW would bear more information about the graph than the one based on LW and, consequently, would be more selective. This conjecture proved to be false since both indexes discriminated all 1037 test structures equally well.

The reason why the mentioned conclusion about the comparative selectivity of the two indexes was found to be erroneous becomes evident from the close inspection of Fig. 2. In this figure are represented all possible walks of length from 1 to 5, starting in vertex 1. They are shown schematically as multigraphs superimposed on the molecular graph of 2-methyl butane. It is evident that each of the multigraphs in family $L(i)$ is a subgraph of at least one multigraph in family $L(i+1)$ and hence all information on graphs in family $L(i)$ is inherent in family $L(i+1)$.

Since both indexes $I(NW)$ and $I(LW)$ were found to be equally selective, the enumeration of all walks up to the length N was discarded and it was decided to retain only the number of walks of length N as the basic invariant.

For trees the index $I(LW)$ was found to be totally selective since the test on 661 nonisomorphic graphs resulted in 661 different index values. A comparison of structural selectivity of 4 indexes (the three indexes mentioned above and the new index $I(LW)$) is given in Table 1. Because in lower alkane structures (with number of carbon atoms equal or less than 8) all four indexes are totally selective, the data on their selectivity for these structures are not included in the table.

On the other hand, none of the three indexes, shown in Table 2, was totally selective in the case of cyclic graphs. While index $I(LW)$ had no degenerated values for the tested acyclic and monocyclic structures, 7 pairs of isocodal bicyclic graphs (i.e. two nonisomorphic graphs having the same index value) were found among 376 tested cyclic structures. The 7 pairs of graphs with degenerated $I(LW)$ values are shown in Fig. 3. Even if the number of isocodal pairs of graphs is small, it is sufficient to generalize the rule to predict the type of structures that will have equal index values. Some of the predicted isocodal graph pairs with

Table 1. Structural selectivity[a] of some graph theoretical indexes in alkane series

Alkane	Nonane	Decane	Undecane	Dodecane
N(str.)	35	75	159	355
Index	Number of different index values			
$^1\chi$	31	55	92	179
J	35	75	159	349
ID	35	75	159	355
$I(LW)$	35	75	159	355

[a] All index values were calculated to 16 decimal places since the usually reported precision of 4 decimal places may lead to wrong conclusions on selectivity [13, 15]

Table 2. Structural selectivity of some[a] graph theoretical indexes for all mono- and bicyclic graphs with 4 to 8 vertexes

N(vert.)	4	5	6	7	8
N(str.)	3	10	29	84	250
Index	Number of different index values				
$^1\chi$	3	10	24	60	144
$I(LW)$	3	10	28	83	245
J	3	10	29	84	248

[a] We have not calculated the selectivity of ID on the above structural set but the literature data [15] indicate its total selectivity for all monocyclic graphs with 6 and 7 vertexes, selected monocyclic graphs with 8 vertexes, and some other cyclic structures

the number of vertexes greater than 8 (hence these structures were not present in the test structural data bank) are shown in Fig. 3.

The general form of bicyclic structures having the same $I(LW)$ value is shown in Fig. 4. The structures of type A correspond to nonseparable graphs (blocks) if they are not substituted (i.e. if all vertexes are members of at least one cycle). The matching isocodal structures of type B have the same number of bridges as the A type structures have common edges in the two cycles. Of course, the term "bridge" is understood in the graph-theoretical [3] and not in chemical meaning of the word, i.e. a bridge being an edge whose removal causes the disconnection of the graph.

In addition to the unsubstituted structures in Fig. 4, the same structures substituted symmetrically with two isomorphic chains or trees are also isocodal. The symmetry of the isocodal graphs representing these structures may be described by automorphic mappings expressed by the following permutations of the vertexes in Fig. 4:

$$(1 \quad k)(2 \quad k+1)(3 \quad k+2)\cdots$$

$$(1 \quad k)(2 \quad k-1)(3 \quad k-2)\cdots.$$

In a language more familiar to the chemists, the first permutation corresponds to the inversion of all atoms through the inversion center or to the rotation about a two-fold axis perpendicular to the common edge (type A) or bridge (type B), while the second permutation corresponds to the reflection in the plane perpendicular to the common edge or bridge (Fig. 3).

The substituted structures of type A are no longer blocks but their origin can be traced back to their nonseparable ancestor.

To conclude the discussion of its selectivity, it may be stated that the index $I(LW)$ discriminates the tested acyclic structures equally well as molecular ID and better than all other graph-theoretical indexes defined until now. For the cyclic structures its discriminating power is lesser but still comparable to that of Blaban's J and Randić's ID. The simplicity with which the index predicts isocodal

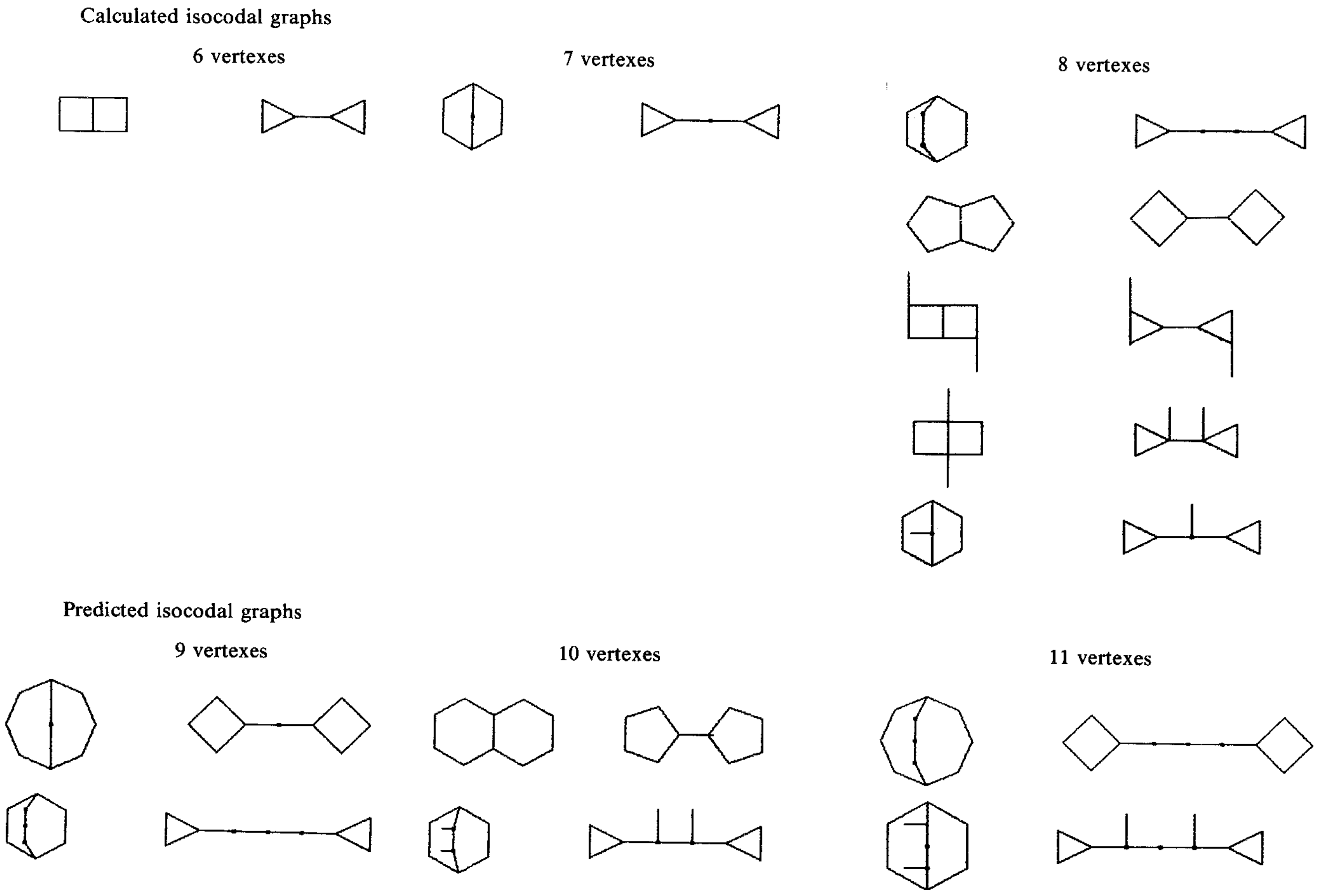


Fig. 3. Calculated and predicted isocodal nonisomorphic bicyclic graphs

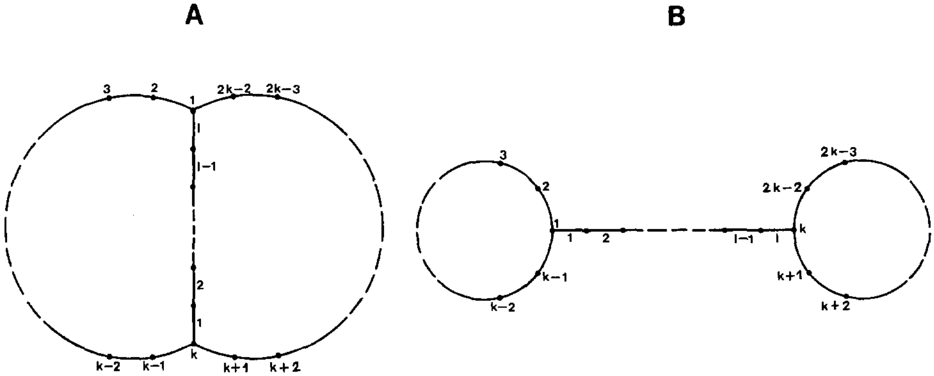


Fig. 4. (A) Nonseparable bicyclic graph consisting of two isomorphic cycles with $k+l-1$ vertexes each. Common edges form a path of length l. (B) Separable bicyclic graph consisting of two isomorphic cycles with $k-1$ vertexes each. A path consisting of l bridges connects the two cycles

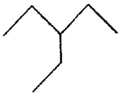
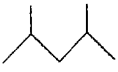
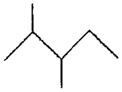
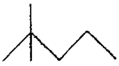
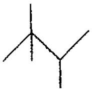
bicyclic structures is also worth mentioning from the graph-theoretical point of view.

The values of $I(LW)$ can be used for a quick if rather pragmatical test of graph isomorphism. Since the index is not absolutely discriminating, the eventual non-isomorphism of isocodal graphs must be ascertained by other means; nevertheless, because of the excellent structural selectivity of $I(LW)$, there is a high probability that isocodal graphs will be isomorphic also. By this method, we found recently five pairs of isomorphic graphs in the plotouts of a large collection of 427 graphs which were supposed to be all nonisomorphic. Correspondence with authors clarified this as a fault in the plotter routines.

Apart from the selectivity, there is another important aspect of graph-theoretical descriptors: they can be used to order the structures according to the attributed numerical values. This ordering may or may not parallel the ordering of structures by their properties. In the following paragraphs we shall touch this vast subject only briefly to compare the order induced on some structural populations by $I(LW)$ and several other indexes.

In Table 3 the values of 6 different indexes for all heptane isomers are given. Two of the indexes not yet cited are also well known: the Wiener number W [24] and Bonchev's information on distances I_D [25]. All values given in Table 3 were calculated with the precision of at least 8 decimal places but only the first four (five for ID) numerals are shown. The $I(LW)$ values in the table are the real values multiplied by 1000. Beside the values of indexes calculated for 9 heptane isomers, the ordering of structures induced by the decreasing value of separate indexes is given in parenthesis. The ordering of structures by $I(LW)$ is identical to that of I_D and W, it parallels that of ID closely and that of ${}^1\chi$ generally, and it is practically the inverse of the ordering by J. The number of walks is closely related to the distribution of distances in the graph – the more compact (branched, sterically hindered, etc. . .) is the structure, the lesser is its

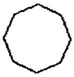
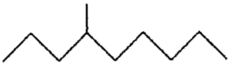
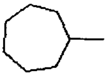
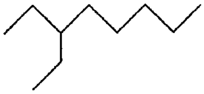
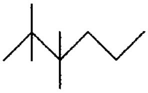
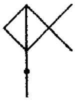
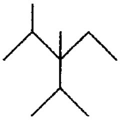
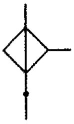
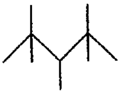
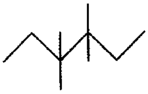
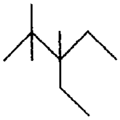
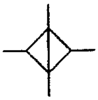
Table 3. Ordering of 9 heptane isomers by index value for 6 different graph-theoretical indexes

Structure	$I(LW)$ $\times 1000$	I_D	J	${}^1\chi$	ID	W
	90.92 (1)	233.5 (1)	2.447 (9)	3.414 (1)	12.834 (1)	56 (1)
	67.06 (2)	218.2 (2)	2.678 (8)	3.270 (4)	12.670 (2)	52 (2)
	59.31 (3)	210.5 (3)	2.832 (7)	3.308 (3)	12.660 (4)	50 (3)
	52.28 (4)	203.5 (4)	2.992 (5)	3.346 (2)	12.669 (3)	48 (4, 5)
	48.88 (5)	203.1 (5)	2.953 (6)	3.125 (6)	12.509 (5)	48 (4, 5)
	41.58 (6)	195.6 (6)	3.144 (4)	3.181 (5)	12.505 (6)	46 (6, 7)
	39.09 (7)	195.1 (7)	3.154 (3)	3.061 (8)	12.449 (7)	46 (6, 7)
	32.46 (8)	187.7 (8)	3.360 (2)	3.121 (7)	12.443 (8)	44 (8)
	26.23 (9)	179.9 (9)	3.541 (1)	2.943 (9)	12.293 (9)	42 (9)

$I(LW)$ value. All other indexes shown in the table are more or less directly related to distance distribution with the exception of ${}^1\chi$ which is based on vertex degree so that the general good agreement in ordering is not surprising. The index $I(LW)$ follows the rules, established by Bonchev and Trinajstić [25] for the quantification of molecular branching very well. A recent work on chemical ordering of molecules [26] gives identical ordering of heptane isomers which correlates well with several physico-chemical properties.

An example of ordering of larger structural populations using $I(LW)$ values is given in Table 4. All 75 decane isomers and all 250 mono- and bicyclic structures with 8 vertexes are ordered according to the attributed $I(LW)$ value. Only the first and last 6 structures of each class are shown in the table. The ordering of decanes is "conventional", going from n-decane to 2,2,3,3,4-pentamethyl hexane.

Table 4. Ordering of 75 decane isomers and 250 mono- and bicyclic structures having 8 vertexes by I(LW) value

Decane isomers	$I(LW) \times 1000$	Mono- and bicyclic str. with 8 vertices	$I(LW) \times 1000$
	15.98		31.25
	12.38		25.32
	11.09		22.53
	10.55		22.39
	10.34		20.45
	9.40		20.03
⋮ 63 structures	⋮	⋮ 238 structures	⋮
	2.29		2.63
	2.20		2.52
	1.94		2.26
	1.90		2.09
	1.87		1.97
	1.63		1.82

In the class of mono- and bicyclic structures the main parameter for ordering appears to be the number of cycles as this is the structural feature that increases the number of walks the most. Apart from this, the ordering follows the usual trend from the least to the most sterically hindered structure.

Concluding remarks

The comparison of index $I(LW)$ with other highly discriminating graph-theoretical indexes is favourable. The calculation of index values is straightforward and not demanding in computer time. The algorithms for the generation of values for large number of structures from their adjacency matrices are simple and efficient. Moreover, they can be implemented without too much effort in any chemical information system having access to structural data either in the form of adjacency matrices, connection tables, Wiswesser linear notation, etc. Based on an easily calculable graph invariant, the new index discriminates well similar nonisomorphic structures and is undoubtedly one of the three most structurally selective graph-theoretical indexes defined until now. In this presentation only homoatomic chemical structures were discussed but the possibilities to include heteroatoms in the scheme are under study.

The possibility of prediction of isocodal structures, described in the discsussion, removes in fact the degeneration of index values as the structures in question can be discriminated by other means. Of course extensive further study is needed to substantiate this conjecture for bicyclic systems and to generalize it to systems with higher number of cycles and substituents.

Although the potential of the new index for the characterization of molecular graphs in structure — properties correlation was not yet studied in detail, the preliminary trials of ordering the structures by index values are promising.

Acknowledgement. Discussions with Professor Jure Zupan and Dr. Marjana Novič and the financial support of the Research Community of Slovenia are gratefully acknowledged.

References

1. Balaban A T (1985) J Chem Inf Comput Sci 25:334-343
2. Trinajstić N (1983)) Chemical graph theory, vol II. CRC Press, Boca Raton, Fl
3. Harary F (1972) Graph theory. Addison-Wesley, Reading, MA
4. Randić M (1980) J Comput Chem 4:386-399
5. Knop J V, Muller W R, Szymanski K, Randić M (1983) Croat Chem Acta 56:405-409
6. Randić M, Woodworth W L, Graovac A (1983) Int J Quant Chem 24:435-452
7. Randić M, Wilkins C L (1979) J Phys Chem 83:1525-1540
8. Randić M, Wilkins C L (1980) Int J Quant Chem 18:1005-1027
9. Wilkins C L, Randić M (1980) Theor Chim Acta 58:45-68
10. Morgan H L (1965) J Chem Doc 5:107-113

11 Razinger M (1982) Theor Chim Acta 61:581-586

12. Balaban A T, Chiriac A, Motoc I, Simon Z (1980) Steric fit in quantitative structure - activity relations. Lect Notes Chem 15, chap 3. Springer, Berlin Heidelberg New York

13. Razinger M, Chrétien J R, Dubois J E (1985) J Chem Inf Comput Sci 25:23-27
14. Balaban A T (1982) Chem Phys Letters 89:399-404
15. Randić M (1984) J Chem Inf Comput Sci 24:164-175
16. Randić M (1975) J Am Chem Soc 97: 6609-6615
17. Kier L B, Hall L H (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
18. Szymanski K, Muller W R, Knop J V, Trinajstić N (1985) J Chem Inf Comput Sci 25:413-415
19. Razinger M, Chrétien J R, Dubois J E (1984) Vestn Slov Kem Drus 31:211-227
20. Deo N (1974) Graph theory with applications to engineering and computer science. Prentice-Hall, Englewood Cliffs, NJ
21. Bonchev D, Mekenyan O, Trinajstić N (1981) J Comput Chem 2:127-148
22. Randić M (1975) J Chem Inf Comput Sci 15:105-108
23. Golender V E, Drboglav V V, Rosenblit A B (1981) J Chem Inf Comput Sci 21:196-204
24. Wiener H (1947) J Am Chem Soc 69:17-20
25. Bonchev D, Trinajstić N, (1977) J Chem Phys 67:4517-4533
26. Grossman S C (1985) Int J Quant Chem 28: 1-16