Journal of Mathematical Chemistry 13(1993)95–106

On the search for the best correlation between graph theoretical invariants and physicochemical properties

Daniel A. Morales
Grupo de Química Teórica, Departamento de Química, Facultad de Ciencias, Universidad de los Andes, Mérida 5101, Venezuela

and

Oswaldo Araujo
Departamenta de Matemática, Facultad de Ciencias, Universidad de los Andes, Mérida 5101, Venezuela

Received 12 August 1991; revised 26 March 1992

We have examined the correlation of the Randić connectivity index with the Hosoya topological index, the Wiener number and the molecular identification number in search of the optimal functional relation between those indices and the boiling points of alkanes. We found that some functional relations used empirically in the literature can be understood using the known fact that the Randić connectivity index is the most successful single descriptor of molecular structure.

1. Introduction

It is well known that molecular topology determines a large number of molecular properties, including not only those depending on molecular size and shape such as boiling points, molecular volumes, solubilities, refractive indices, etc., but also the quantum mechanical characteristics of molecules, such as energy levels, electronic populations, etc. which depend, essentially, on the connectivity of the atoms [1].

Thus, it would be of great interest to have some quantitative measure reflecting the essential features of a given topological structure. Such measures are usually called topological indices or, more exactly, graph theoretical invariants, since these numbers are the same for isomorphic graphs [2]. These indices in some way reflect not only the size and shape of a molecule but also their connectivity, i.e. the way their atoms are linked.

Many topological indices have been developed through the years and correlated with many physicochemical properties [2–5]. Recently, Randić et al. [6] made a critical study of the correlation of four indices: the Wiener number (W), the Hosoya

topological index (Z), the Randić connectivity index (χ), and the molecular identification number (ID) with the boiling points of alkanes. They found different functional forms for the correlations, which led to significant improvements in the correlation coefficients compared to those using only the plain topological indices. However, the search for the best functional form was somewhat empirical. On the other hand, their study shows that there exist correlations among the different topological indices, although no mathematical relations were established. It is the purpose of this paper to establish and test approximate functional relations among the four indices previously mentioned, and to explain why the functional forms proposed by Randić et al. work, taking as a basis the correlation between the boiling points of alkanes and the Randić connectivity index.

This paper is organized as follows. In section 2, we review the definitions of the four indices mentioned above and prove their invariance. In section 3, we establish approximate relations between the Randić connectivity index and the Wiener number, the Hosoya topological index, and the molecular identification index to test the expressions obtained with actual correlations. Also, in that section, we correlate the boiling points of the lower alkanes with the functional relations encountered. Section 4 is a summary of our results.

2. Overview of topological indices

We shall assume that the reader has some familiarity with the terminology of graph theory; Harary's book is a good reference [7]. Also, we suppose that all molecular species can be represented by means of an appropriate chemical graph.

An *invariant* of a graph G is any number associated with G which has the same value for any graph isomorphic to G. In a connected graph G, the distance $d(i, j)$ between two vertices i and j is the length of the shortest path joining them.

In what follows, we review the Wiener number, the Hosoya topological index, the Randić connectivity index, and the molecular identification number and prove their invariance.

2.1. WIENER NUMBER

The Wiener number was introduced in 1947 by Wiener [8] and is based on the graph concept of distance. The Wiener number is equal to one-half the sum of the elements of the distance matrix of graph G [9], i.e.

$$W = \frac{1}{2} \sum_{ij} d_{ij}, \tag{1}$$

where the d_{ij}'s are the coefficients of the distance matrix $D = (d_{ij})$.

The Wiener number of any graph G is an invariant. In fact, if G' is a graph isomorphic to G, then $f: G \to G'$ is a bijection which preserves adjacency. Now, if $D = (d_{ij})$ and $D' = (d'_{kl})$ are the distance matrices of G and G', respectively, we have

$d_{ij} = \min d(i, j) = d'_{f(i)f(j)}$, that is, $W \simeq W'$, where W' is the Wiener number of the graph G'.

For normal alkanes, i.e. unbranched alkanes, we have the following formula:

$$W_G = \tfrac{1}{6}(n^3 - n), \tag{2}$$

where n is the number of vertices.

The Wiener number is an index which gives a measure of the branching of a graph [10]. Thus, the Wiener number of a branched molecule is less than that of a linear or less compact molecule. Platt [11] has suggested that the cubic root of the Wiener number is a measure of the mean distance among carbon atoms in a molecule and is, approximately, a measure of the probability that one part of a molecule will be attracted by another by van der Waals forces. Since its introduction, it has been found that this index correlates very well with properties such as boiling point, viscosity, surface tension and refractive index. On the other hand, it has been found that, in general, a system is at its minimum of energy when its Wiener number is a minimum [12]. This idea has been applied to model several processes in which foreign atoms fill interstitial areas in a crystal. By calculating the three-dimensional Wiener number of each possible configuration of crystal-lattice atoms and foreign atoms, and using the above-mentioned idea, it is possible to determine the configurations most likely to be realized. This last finding could be useful in the search for the most probable geometry of clusters.

2.2. THE RANDIĆ CONNECTIVITY INDEX

The Randić connectivity index χ depends on the graph concept of degree [13]. It is defined by

$$\chi = \sum (1/mn)^{1/2}. \tag{3}$$

This summation is over all edges in the graph G and it includes one term for each edge in G. The variables m and n are the degrees of the adjacent points joined by each edge. Since the degree of any graph is an invariant, χ is also an invariant.

Let G be the graph of a normal alkane; then

$$\chi_{4+n'} = \sqrt{2} + (1 + n')/2, \tag{4}$$

where $4 + n'$ is the number of vertices and $n' = -1, 0, 1, 2, \ldots$. It has been found that the Randić connectivity index is the one which presents the best correlation with the physicochemical properties of many substances. The Randić connectivity index is also known as the path-one connectivity index. This comes from the fact that χ is calculated summing over all paths of length one in the structural graph. By a natural extension, it is possible to consider additional indices corresponding to paths of lengths greater than one [14].

2.3. THE HOSOYA TOPOLOGICAL INDEX

The Hosoya topological index Z of a graph G is based on the count of non-adjacent edges of G and is defined as [9]

$$Z = \sum_{k=0}^{m} p(G,k), \tag{5}$$

where $p(G, k)$ represents the number of different ways of selecting k non-adjacent edges in graph G, and the summation extends over all the m edges of the graph. By definition, $p(G, 0) = 1$ and $p(G, 1)$ is equal to the number of edges.

Since every isomorphism of a graph preserves both adjacency and non-adjacency, we conclude that the Hosoya topological index is an invariant of G.

The Fibonacci numbers are integers F_n defined by

$$F_n = F_{n-1} + F_{n-2}, \quad n \geq 2; \quad F_0 = F_1 = 1.$$

They have a simple combinatorial meaning: F_{n+1} is the number of subsets of $\{1, \ldots, n\}$ *such that no two elements are adjacent* [15].

Hence, for normal alkanes, the Hosoya topological index is given by

$$Z = F_{n+1} = \left(1/\sqrt{5}\right)\left[\left(\left(1+\sqrt{5}\right)/2\right)^{n+1} - \left(\left(1-\sqrt{5}\right)/2\right)^{n+1}\right]. \tag{6}$$

2.4. MOLECULAR IDENTIFICATION INDEX

The molecular identification index [16] of a graph G is defined as the sum of weighted path numbers. Each path of length zero is given a weight of unity, and for paths of length greater than zero the weight is equal to the product of $(1/mn)^{1/2}$ terms, one term for each edge included in the path, where m and n are the degrees of the vertices joined by the edge. In mathematical terms

$$\mathrm{ID} = \sum_i \omega_{0i} + 1/2 \sum_{ij} \omega_{ij}, \tag{7}$$

where ω_{0i} is the path-weight for all paths of length zero, and ω_{ij} is the path-weight corresponding to all paths of length greater than zero.

Since the Randić connectivity index and the number of vertices of G are invariants, then the molecular identification index is also an invariant.

It is easy to see that, for normal alkanes,

$$\mathrm{ID} = n + \sqrt{2} \sum_{i=0}^{n-3} (1/2^i) + s, \tag{8}$$

where $n \geq 3$ is the number of vertices and

$$s = 1/2^{n-2} + \sum_{i=0}^{n-3} (i/2^{n-i-2}). \tag{9}$$

Using the expression for the geometric progression, the preceding equation can also be written as

$$\mathrm{ID} = 2(2^{2-n} + 2^{i-n} + n - 2) + (2 - 2^{3-n})\sqrt{2}\,. \tag{10}$$

3. Relations between indices

As has been pointed out previously, the Randić connectivity index is the most successful single descriptor of molecular properties; however, if we plot the correlation between the experimental boiling points of 21 alkanes, tabulated in table 1, and the Randić connectivity index, the linear correlation does not adequately fit the experimental data owing to the curvature of the plot. However, a fitting of the data to a second-degree polynomial in χ improves the correlation, as shown by Randić et al. [6].

From the preceding analysis, it is obvious that a search for the best correlation of physicochemical properties with topological indices is in order.

In this section, we derive and test approximate expressions between the Randić connectivity index and the Wiener number, the Hosoya topological index, and the molecular identification index. The relations obtained are derived from exact formulas for linear alkanes. However, *since the change in a topological index with branching is small for small branching, we suppose that the functional relation obtained for linear alkanes is also valid for branched ones except that we allow the coefficients to vary*. We then make a correlation analysis based on the functional relations derived and the actual values for the topological indices of alkanes C_2 through C_7.

3.1. RELATION BETWEEN THE RANDIĆ CONNECTIVITY INDEX AND THE WIENER NUMBER

From eqs. (2) and (4) above, for normal alkanes we can derive the following approximate relation between χ and W:

$$\chi = A_R + B_R(6W)^{1/3} + C_R(6W)^{-1/3} + \ldots\,, \tag{11}$$

where $A_R = \sqrt{2} - 3/2 \cong -0.09$, $B_R = 1/2$, $C_R = 1/6 \cong 0.17, \ldots$. This equation was derived by approximately solving the resulting cubic equation in terms of W and then substituting the result in eq. (4). In table 1, we reproduce the values of the four topological indices considered here for the alkanes C_2–C_7. A linear regression analysis of the values of χ versus $(6W)^{1/3}$, not including ethane, leads to the following coefficients for the correlation: $A_R = -0.12$, and $B_R = 0.50$ with a correlation coefficient (c.c.) of 0.9906 and a standard deviation (s.d.) of 0.0817. If we include ethane, then the results are $A_R = -0.00394$, $B_R = 0.483$ with a c.c. of 0.9920 and a s.d. of 0.0878.

Table 1

Experimental boiling points of lower alkanes and values of the Randić connectivity index χ, the Wiener number W, the Hosoya topological index Z, and the molecular identification index ID.

Compound	BP (obsd.) °C	Randić index (χ)	Wiener no. (W)	Hosoya index (Z)	ID
ethane	−88.63	1.000	1	2	3.0000
propane	−42.07	1.414	4	3	4.9142
2-methylpropane	−11.73	1.732	9	4	6.7321
n-butane	−0.50	1.914	10	5	6.8713
2,2-dimethylpropane	9.50	2.000	16	5	8.5000
2-methylbutane	27.85	2.270	18	7	8.6968
n-pentane	36.07	2.414	20	8	8.8499
2,2-dimethylbutane	49.74	2.561	28	9	10.4660
2,3-dimethylbutane	57.99	2.643	29	10	10.5236
2-methylpentane	60.27	2.770	32	11	10.6792
3-methylpentane	63.28	2.808	31	12	10.6759
n-hexane	68.74	2.914	35	13	10.8391
2,2-dimethylpentane	79.20	3.061	46	14	12.4490
2,4-dimethylpentane	80.50	3.126	48	15	12.5092
2,2,3-trimethylbutane	80.88	2.943	42	13	12.2931
3,3-dimethylpentane	86.03	3.121	44	16	12.4427
2,3-dimethylpentane	89.78	3.181	46	17	12.5052
2-methylhexane	90.05	3.270	52	18	12.6704
3-methylhexane	91.85	3.308	50	19	12.6600
3-ethylpentane	93.48	3.346	48	20	12.6692
n-heptane	98.42	3.414	56	21	12.8338

Since in some of the situations (like the Randić connectivity index), ethane is a special case in that it is a single molecule that involves a single parameter not appearing in other molecules and hence can be adjusted and influences the statistics without actually reflecting improvement in the description of other molecules. Thus, from this result we can conclude that since the Randić connectivity index presents a good correlation with the boiling points of the 21 alkanes listed in table 1 (c.c. of 0.9914 and a s.d. of 6.746), then a good correlation will be shown between the boiling points and the cubic root of the Wiener number. This is indeed what we find; the correlation obtained is given by

$$\mathrm{BP} = -146.2 + 64.87W^{1/3}$$

with a c.c. of 0.9951 and a s.d. of 5.082 (data plotted in fig. 1). This represents an impressive improvement over the correlation between the boiling points and the first power of the Wiener number, which gives a c.c. of 0.9432 and a s.d. of 17.09.

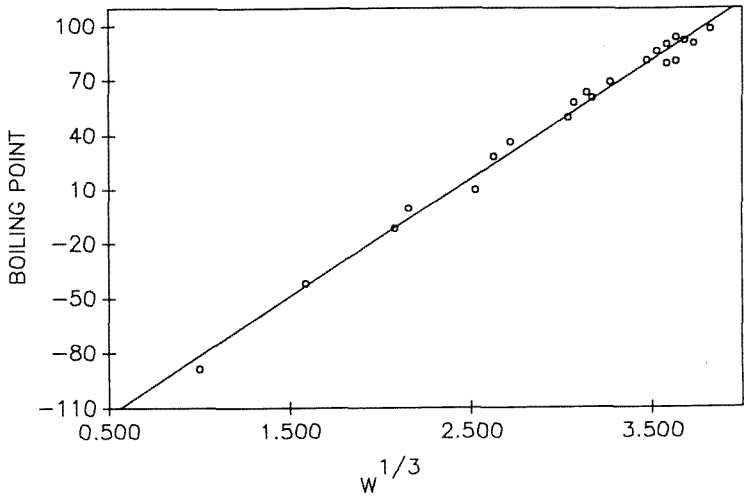

Fig. 1. Correlation between the boiling points of the C_2–C_7 alkanes and the third power of the Wiener number.

Randić et al. [6] tested other powers of W in the correlation with the boiling points of the alkanes listed in table 1, and found that a still better correlation will be obtained using $W^{1/4}$. In a correlation study of boiling points and $W^{1/3}$, we found that a functional dependence of the form

$$\mathrm{BP} = a + bW^{1/3} + c/W^{1/3} \tag{12}$$

suggested by expression (11) with $a = -1.118 \times 10^2$, $b = 5.774 \times 10$, and $c = -3.503 \times 10$ gives a c.c. of 0.9962 and a s.d. of 4.598, which are close to those obtained from the correlation between boiling points and $W^{1/4}$, i.e. 0.9963 and 4.417 [6].

3.2. RELATION BETWEEN THE RANDIĆ CONNECTIVITY AND THE HOSOYA TOPOLOGICAL INDEX

In order to derive a relation between the Randić connectivity index and the Hosoya topological index, we will approximate expression (6) by simply neglecting the term $[(1-\sqrt{5})/2]^{n+1}$ in (6), and combining the resulting equation with eq. (4). Hence, we obtain the following approximate expression which relates χ and Z for linear alkanes:

$$\chi = A_{\mathrm{H}} + B_{\mathrm{H}} \ln Z, \tag{13}$$

where

$$A_{\mathrm{H}} = \sqrt{2} - 2 - \ln 5/(4 \ln \alpha) \cong 0.250,$$

$$B_{\mathrm{H}} = \tfrac{1}{2} \ln \alpha \cong 1.039$$

and α is the golden ratio $(1+\sqrt{5})/2$.

Using the values for χ and Z from table 1, a correlation analysis of χ versus $\ln Z$, not including ethane, produces the following result: $A_H = 0.300$, $B_H = 1.024$ with a c.c. of 0.9988 and a s.d. of 0.0286. If we include the data for ethane, we obtain $A_H = 0.2956$, $B_H = 1.026$ with a c.c. of 0.9992 and a s.d. of 0.0279. This correlation also explains previous correlations of boiling points of alkanes with the logarithm of Z. For the 21 alkanes listed in table 1, we find

$$\mathrm{BP} = -121.95 + 74.99 \ln Z$$

with a c.c. of 0.9902 and a s.d. of 7.185 (fig. 2), a large improvement over the correlation of boiling points with Z (c.c. = 0.9128, s.d. = 21.01).

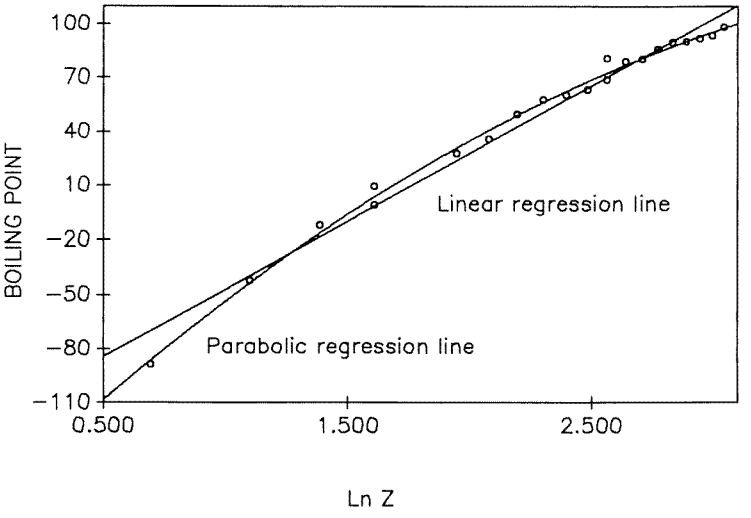

Fig. 2. Correlation between boiling points of the C_2–C_7 alkanes and the logarithm of the Hosoya topological index Z.

Randić et al. [6] tested, empirically, several functional relations of powers Z and found good correlations, especially for negative powers. This finding can be explained using the preceding discussion. Since Z^x can be written in a power series of $\ln Z$ as

$$Z^x = \exp(x \ln Z) = 1 + x \ln Z + \tfrac{1}{2} x^2 \ln^2 Z + \ldots , \tag{14}$$

we see that Z^x is proportional to $\ln Z$ and that a correlation of Z^x with boiling points, for small x, will give good correlation coefficients. Randić et al. find the best fit for $x = -1/3$ (using only fractional powers). The nonlinear terms in expansion (14) above can be thought of as coming from the nonlinear terms neglected in eq. (6) when the approximate formula (13) was obtained. Expansion (14) also suggests that

a still better correlation can be obtained between boiling points and the Hosoya topological index by the functional expression

$$BP = a_H + b_H \ln Z + c_H \ln^2 Z, \tag{15}$$

which, in turn, can be thought of as coming from a nonlinear relationship between the Randić connectivity index and the Hosoya topological index, of the form

$$\chi = A' + B' \ln Z + C' \ln^2 Z. \tag{16}$$

Indeed, a fitting of the Randić connectivity index and the Hosoya topological index to the functional relation (16) produces the following result:

$$\chi = 0.2755 + 1.049 \ln Z - 5.83 \times 10^{-3} \ln^2 Z$$

with a c.c. of 0.9992 and a s.d. of 0.0286.

Fitting the data given in table 1 to the functional relation given by eq. (15) gives the following coefficients: $a_H = -169.92$, $b_H = 130.32$, and $c_H = -13.93$, with a c.c. of 0.9976 and a s.d. of 3.657 (fig. 2). These results are very close to those obtained using $Z^{-1/3}$ (c.c. = 0.9976, s.d. = 3.547).

Also, since $A_H < B_H \ln Z$ in eq. (13), we can rewrite that equation as

$$\chi = B_H \ln Z[1 + A_H/B_H \ln Z] \equiv B_H \ln Z. \tag{17}$$

Randić found that the best correlation is obtained using $\chi^{1/3}$; then we can write eq. (17) as

$$\chi^{1/3} = B_H^{1/3} (\ln Z)^{1/3}. \tag{18}$$

Thus, a correlation of boiling points and $(\ln Z)^{1/3}$ must also give a better fit than the correlation using $\ln Z$. In fact, a regression analysis shows that, for the 21 alkanes listed in table 1,

$$BP = -387.56 + 335.56 (\ln Z)^{1/3}$$

with a c.c. of 0.9976 and a s.d. of 3.548, which are close to those obtained using $Z^{-1/3}$ (fig. 3).

3.3. RELATION BETWEEN THE RANDIĆ CONNECTIVITY INDEX AND THE MOLECULAR IDENTIFICATION INDEX

If we combine eqs. (4) and (10), we find a relation between the Randić connectivity index and the molecular identification index given by

$$ID = a_{ID} \exp(-b_{ID}\chi) + c_{ID}\chi + d_{ID}, \tag{19}$$

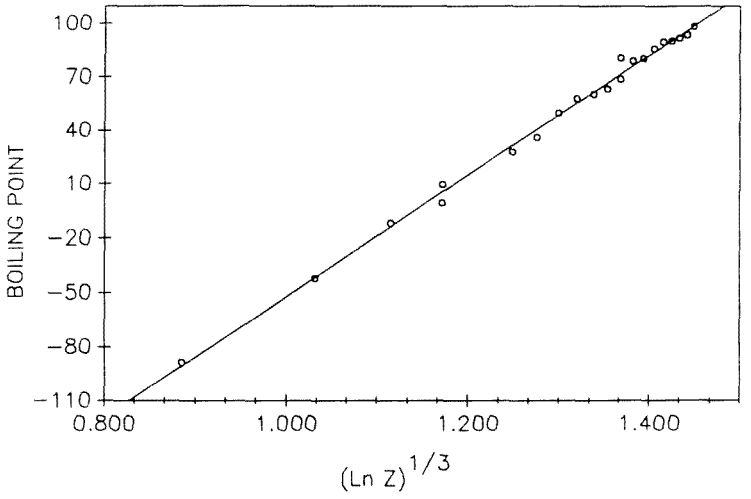

Fig. 3. Correlation between the boiling points of the C_2–C_7 alkanes and the one-third power of the logarithm of the Hosoya topological index Z.

where

$$a_{ID} = 2^{2\sqrt{2}}\left(\tfrac{3}{2}-\sqrt{2}\right),\ b_{ID} = \ln 4,\ c_{ID} = 4,\ d_{ID} = 2\left(1-\sqrt{2}\right) \cong -0.8284.$$

Since the first term in eq. (19) decays very rapidly with χ, we can see that the relation between ID and χ is approximately linear. A linear regression analysis, using the data in table 1, and not including ethane, produces the following values: $c_{ID} = 4.004$ and $d_{ID} = -0.3149$ with a c.c. of 0.9828 and a s.d. of 0.450. If we include ethane, then we obtain $c_{ID} = 4.13$ and $d_{ID} = -0.667$, with a correlation coefficient of 0.9876 and a s.d. of 0.457. Using ID directly in a correlation analysis with boiling points gives

$$BP = -129.685 + 17.514\ ID$$

with a c.c. of 0.9920 and a s.d. of 6.496 (fig. 4). A slight improvement on the correlation is found using the fact that $\chi^{1/3}$ gives the best index. In this way, from eq. (19) we find that $ID^{1/3}$ is proportional to $\chi^{1/3}$. Thus, a regression analysis of boiling points and $ID^{1/3}$ gives

$$BP = -396.471 + 207.641\ ID^{1/3}$$

with a c.c. of 0.9922 and a s.d. of 6.406 (fig. 5). Randić et al. find that the best correlation is obtained using $ID^{1/2}$. The deviation in the powers of 1/3 and 1/2 is, presumably, due to ignoring the exponential term in eq. (19).

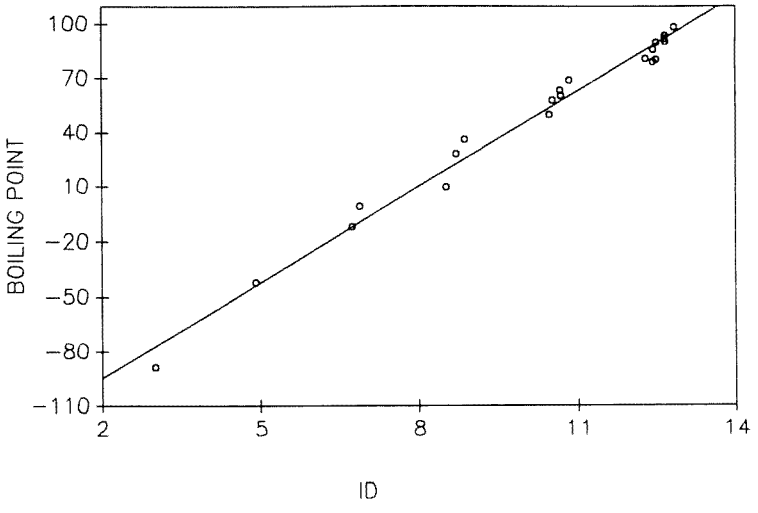

Fig. 4. Correlation between the boiling points of the C_2–C_7 alkanes and the molecular identification index ID.

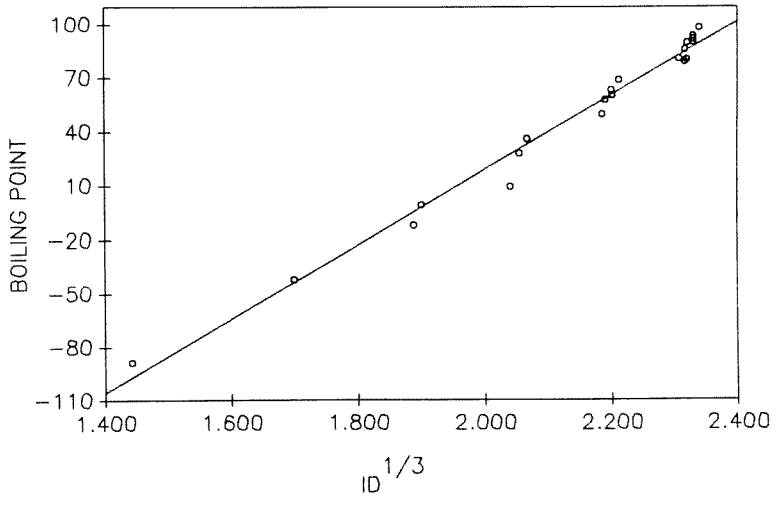

Fig. 5. Correlation between the boiling points of the C_2–C_7 alkanes and the one-third power of the molecular identification index ID.

4. Conclusions

In this paper, we have made an anlysis of the correlations which exist among four important topological indices: the Randić connectivity index, the Wiener number, the Hosoya topological index, and the molecular identification index. We have established approximate relations which have been tested. The functional relations

have then been used in establishing correlations with the boiling points of alkanes. Since it was found that the Randić connectivity index and its one-third power gives the best correlation with boiling points [6], we have used the functional relations among the Randić connectivity index and the three other indices mentioned above to find the best functional form for the correlation between those indices and the boiling points of alkanes, explaining, in this way, some empirical findings of Randić et al. [6].

The analysis of this paper also corroborates a fact pointed out by Randić: owing to the correlation among indices, it is going to be more difficult than anticipated to arrive at new invariants that have novel difference structural bases and cannot be simply (if not trivially) related to those already existing [6].

Finally, we want to point out that recently the interest in topological indices has shifted toward to use of two or more descriptors in structure–property relationships [17]. In this sense, correlations based on a single descriptor may give some insights when one designs multivariate regressions. For example, fig. 4 shows that isomeric variations (represented by points with a slope visibly larger than the slope of the overall linear regression) could be improved with a descriptor which displays the opposite behavior. In fact, a correlation of the boiling points of C_2–C_7 with ID as the first descriptor and 1/ID as the second shows an improvement of the correlation coefficient (0.9944) and a reduction of the standard deviation (5.4319) compared to those values obtained with the use of ID as the sole descriptor.

Acknowledgement

This work was partially supported by the CDCHT of La Universidad de los Andes. We thank the referees for several helpful comments and for pointing out some errors in our manuscript.

References

[1] D.H. Rouvray, Sci. Am. 255(1986)40.
[2] D.H. Rouvray, Am. Sci. 61(1973)729.
[3] N- Trinajstić, *Chemical Graph Theory*, Vol. 2 (CRC Press, Boca Raton, 1983).
[4] R.B. King and D.H. Rouvray (eds.), *Graph Theory and Topology in Chemistry*, sect. B (Elsevier, Amsterdam, 1987).
[5] D.H. Rouvray, J. Comput. Chem. 8(1987)470.
[6] M. Randić, P.J. Hansen and P.C. Jurs, J. Chem. Inf. Comput. Sci. 28(1988)60.
[7] F. Harary, *Graph Theory* (Addison–Wesley, Reading, MA, 1972).
[8] H. Wiener, J. Am. Chem. Soc. 69(1947)17.
[9] H. Hosoya, Bull. Chem. Soc. Japan 44(1971)2332.
[10] D.H. Rouvray, Discr. Appl. Math. 19(1988)317.
[11] J.R. Platt, J. Phys. Chem. 56(1952)328.
[12] D. Bonchev, O. Mekenyan and H.G. Fritsche, Phys. Stat. Sol. (a) 55(1979)181; (a) 56(1979)607; Z. Phys. Chem. (Leipzig) 265(1984)959; J. Cryst. Growth 49(1980)90.
[13] M. Randić, J. Am. Chem. Soc. 97(1975)6609.
[14] L.B. Kier, W.J. Murray, M. Randić and L.H. Hall, J. Pharm. Sci. 65(1976)1226.
[15] L. Comtet, *Advanced Cominatorics* (Reidel, Dordrecht, 1974) p. 45.
[16] M. Randić, J. Chem. Inf. Comput. Sci. 24(1984)164.
[17] M. Randić, J. Chem. Inf. Comput. Sci. 31(1991)311.