Indian Journal of Chemistry
Vol. 37A, October 1998, pp. 849-855

Papers

An atom-bond connectivity index: Modelling the enthalpy of formation of alkanes

Ernesto Estrada[a], Luis Torres[a], Lissette Rodríguez[a] & Ivan Gutman[b]*
[a]Department of Drug Design, Centro de Bioactivos Químicos,
Universidad Central de Las Villas, Santa Clara 54830, Villa Clara, Cuba, and [b]Faculty of Science,
University of Kragujevac, P. O. Box 60, YU - 34000 Kragujevac, Yugoslavia

Received 15 December 1997

The atom-bond connectivity index (ABC), a novel graph theoretical invariant, based on the connectivity between atoms and bonds in a molecule, is proposed. This structure-descriptor is computed from the vertex and edge degrees, but in contrast to the original connectivity index of Randic - ABC does not reflect the extent of branching of the molecule. ABC is used to describe the heats of formation of alkanes, resulting in a good quantitative structure-property relationship (QSPR) model (r = 0.9970). The model is interpreted in such a manner that the intercept and slope of the regression equation have a physical meaning.

The introduction of graph-theoretical structure-descriptors represents an important step forward in the search of predictive models in chemistry and falls within the lines of the increasing use of mathematical and computational methods in contemporary chemistry[1,2]. The basis for these models is the study of the quantitative structure-property and structure-activity relationships (QSPR and QSAR, respectively), in which the structural information of molecules is encoded into numbers obtained from graph-theoretical invariants.

One of the most popular graph invariants is the so-called connectivity index (χ), introduced by Randic[3] in 1975. Originally χ was aimed at the modelling of the branching of the carbon-atom skeleton of alkanes, but eventually proved to be exceptionally successful in correlations with a great variety of physico-chemical and pharmacological properties of many kinds of organic molecules (acyclic and cyclic, hydrocarbons and heteroatom-containing, aliphatic and aromatic)[4,5]. As a consequence, the connectivity index is nowadays the most frequently employed structure-descriptor in QSPR and QSAR studies[6]. The success of χ is based on the fact that numerous molecular properties are prominently branching-determined. On the other hand, for those (few) molecular properties which are weakly dependent on branching (of which enthalpy is a typical representative), the connectivity index is of little practical value. In Fig. 1 the standard heats of formation (in gas phase, at 25°C) of a selected set of alkanes are plotted against the respective χ values. The regression equation and the corresponding statistical parameters (N = sample size, R = correlation coefficient, s - standard deviation) are:

$$-\Delta H^{\circ}_{f}\,(\text{kJ / mol}) = 64.85 + 40.93\,\chi$$
$$N = 48 \qquad R = 0.9534 \qquad s = 12.26$$

It may be observed that with increasing number of carbon atoms both $-\Delta H^{\circ}_{f}$ and χ increase, but within sets of isomers (having constant carbon-atom count) the opposite is the case.

The above mentioned problem has been noticed already by Randic[3]. In order to overcome it, he proposed to use a two-variable equation for ΔH°_{f}, that includes both the number of carbon atoms (n) and χ.

The aim of this work is to design a structure-descriptor that would be as similar as possible to the connectivity index (in particular, defined in terms of the same graph parameters as χ), but would codify structural information unrelated to branching, thus being complementary to χ.

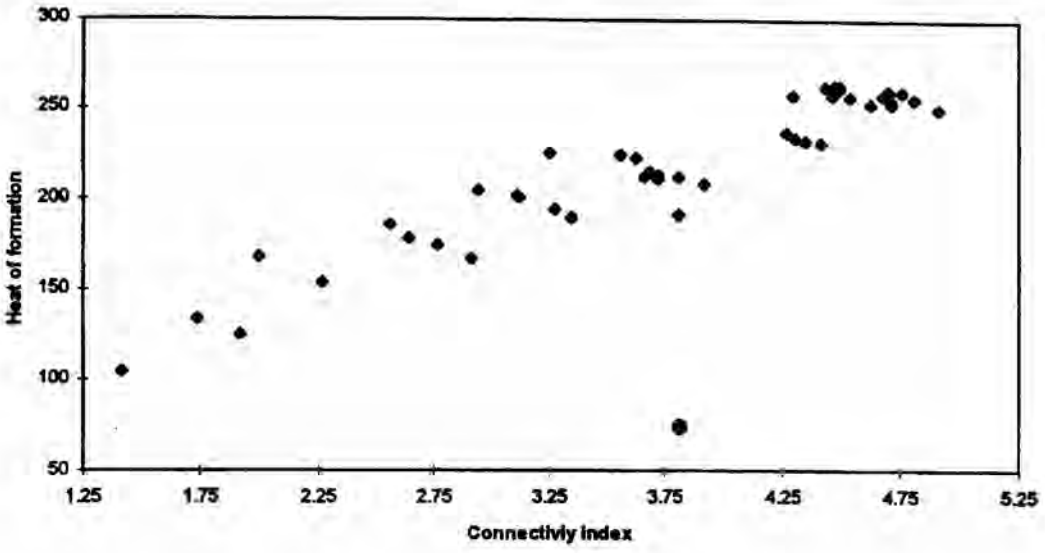

Fig. 1– Heats of formation of the alkanes from training set (see Table 1) vs. the connectivity index, Eq.(1)

The connectivity index is defined as follows. Let G be the molecular graph. Let e be an edge of G, connecting the vertices v_i and v_j. Let the degree (= the number of first neighbours) of the vertex v be denoted by $\delta(v)$. Then the connectivity index of G is given by:

$$\chi = \chi(G) = \sum_{e} [\delta(v_i)\,\delta(v_j).]^{-1/2} \qquad \ldots (1)$$

where the summation goes over all edges of G.

This graph invariant has been the object of several theoretical studies[7-9] and it has served as a basis for the generation of other chemically useful topological indices. For instance, Balban's J index[10] uses distance degrees instead of vertex degrees, arranged in a manner analogous to Eq. (1). The bond connectivity index [11-13], introduced by one of the present authors, uses the edge (bond)degree instead of the vertex (atom) degrees in an analogous way as in the χ index. The edge degree is identical to the vertex degree of the corresponding vertex of the line graph, associated to the molecular graph[14]. As a consequence, the edge connectivity index is equal to the Randic index of the line graph of the molecular graph[15].

The study of graph theoretical indices coming from the line graph of the molecular graph has become one of the newest directions in the search for molecular structure-descriptors[16-21]. The use of line graphs for generation of topological indices and their applications for modelling physico-chemical properties of organic molecules appears to be first reported by Bertz in the early 1980s [22-24]. However, a systematic use of line graphs in designing QSPR models has started relatively recently in a series of papers of Estrada and Gutman[15,16,19-21].

What we report here is the first graph theoretical structure-descriptor that combines concepts coming from the molecular graph and its line graph. The novel descriptor will be referred to as the ABC *(Atom-Bond Connectivity)* index. As we show later on, it has a very good ability to describe and predict the enthalpy of formation of alkanes.

The ABC index

The ABC index is calculated in analogy to the Randic index, Eq. (1), but using both the vertex and edge degrees. ABC is defined by means of the following expression:

$$ABC = \sum_{e}\left[\frac{\delta(v_i)\cdot\delta(v_j)}{2\delta(e)}\right]^{-1/2} \qquad \ldots(2)$$

where $\delta(e)$ stands for the degree of the edge e, namely the number of edges incident to e. As before, the summation goes over all edges of the molecular graph G. Clearly, $\delta(e)$ is the degree of the respective ver-

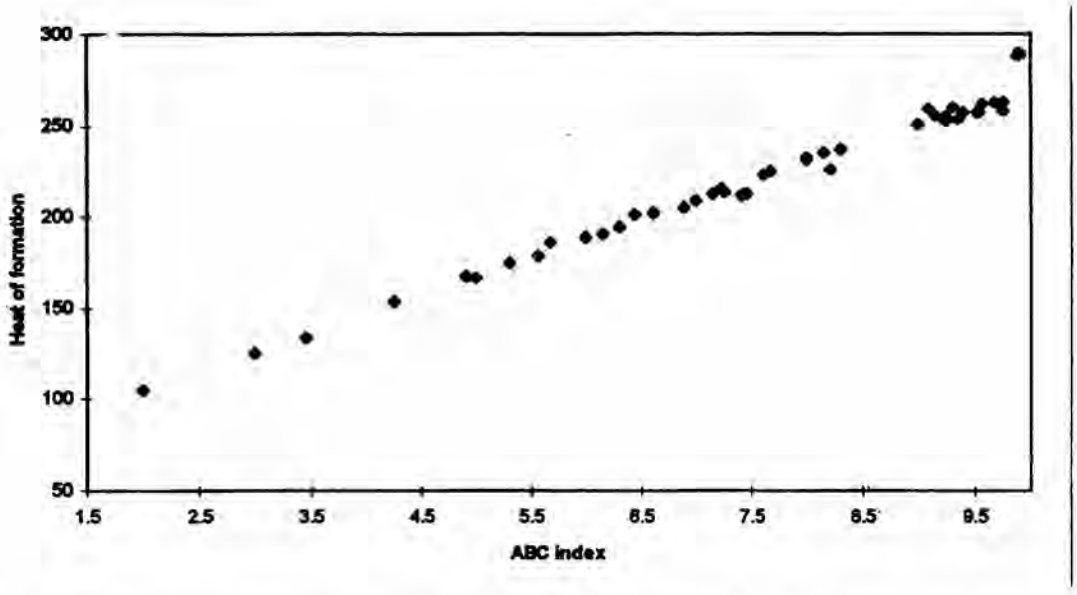

Fig 2 – Heats of formation of the alkanes from training set (see Table 1) vs. the atom bond connectivity index, Eq (2).

tex of the line graph of G. By using the well known relationship between the edge and vertex degrees, Eq. (2) is rewritten as follows:

$$ABC = \sum_{e} \left[\frac{\delta(v_i) \cdot \delta(v_j)}{2(\delta(v_i) + \delta(v_j) - 2)} \right]^{-1/2} \quad ...(3)$$

The right-hand side of Eq. (3) is undefined for methane. For ethane, the molecular graph of which has just one edge, both vertices have unit degrees, and the edge degree is equal to zero. Hence the ABC index of ethane is equal to zero.

For molecular graphs with more than two vertices, the term

$$\left[\frac{\delta(v_i) \cdot \delta(v_j)}{2(\delta(v_i) + \delta(v_j) - 2)} \right]^{-1/2}$$

has the obvious interpretation as being the edge (i.e., bond) contribution to the ABC index. It is easy to see that there are only a few distinct values that this term may assume, depending on the type of the respective edge (i.e. chemical bond). We say that an edge is of type (x,y) if the degree of its two end-vertices are x and y, respectively. In molecular graphs of saturated hydrocarbons (except methane and ethane) the only possible edge types are (1,2), (1,3), (1,4), (2,2), (2,3), (2,4), (3,3), (3,4) & (4,4). If the number of edges of type (x,y) in the molecular graph considered is denoted by $m_{x,y}$ then from Eq. (3) it immediately follows:

$$ABC = m_{1,2} + (2/\sqrt{3})m_{1,3} + (\sqrt{3/2})m_{1,4} + m_{2,2} + m_{2,3} + m_{2,4} + (\sqrt{8}/3)m_{3,3} + \\ + (\sqrt{5/6})m_{3,4} + (\sqrt{3}/2)m_{4,4} \quad ...(4)$$

i.e.,

$$ABC = m_{1,2} + 1.155m_{1,3} + 1.225m_{1,4} + m_{2,2} + m_{2,3} + m_{2,4} + 0.943m_{3,3} + \\ + 0.913m_{3,4} + 0.866m_{4,4}$$

which should be compared with:

$$\chi = (1/\sqrt{2})m_{1,2} + (1/\sqrt{3})m_{1,3} + (1/2)m_{1,4} + (1/2)m_{2,2} + (1/\sqrt{6})m_{2,3} + \\ + (1/\sqrt{8})m_{2,4} + (1/3)m_{3,3} + (1/\sqrt{12})m_{3,4} + (1/4)m_{4,4}$$

Table 1 — The ABC index of compounds in the training set as well as the experimental[26] and calculated standard gas phase heats of formation at 298K.

Alkane	ABC Index	$-\Delta H°_f$ (kJ / mol) obsd	$-\Delta H°_f$ (kJ / mol) calcd.	Residual
3	2.00000	104.67	106.73	-2.06
4	3.00000	125.66	127.10	-1.44
2M3	3.45694	134.19	136.41	-2.22
2M4	4.25286	153.68	152.62	1.06
6	5.00000	167.03	167.84	-0.81
22MM3	4.89902	167.95	165.78	2.16
2M5	5.30944	174.8	174.14	0.65
23MM4	5.56158	178.3	179.28	-0.98
22MM4	5.67430	186.10	181.58	4.52
3E5	6.00000	189.3	188.21	1.09
3M6	6.15472	191.3	191.36	-0.06
2M6	6.30944	194.6	194.52	0.08
33MM5	6.44944	201.2	197.37	3.83
24MM5	6.61886	201.7	200.82	0.88
223MMM4	6.89648	204.5	206.48	-1.98
8	7.00000	208.7	208.58	0.11
3E2M5	7.40688	211.0	216.87	-5.87
4M7	7.15468	212.0	211.74	0.26
24MM6	7.46414	212.6	218.04	-5.4
34MM6	7.25216	212.8	213.72	-0.92
3E3M5	7.22472	214.9	213.16	1.74
25MM6	7.61886	222.5	221.19	1.31
22MM6	7.67430	224.6	222.32	2.28
2233M4	8.21454	225.5	233.33	-7.83
9	8.00000	228.7	228.96	-0.26
3E7	8.00000	231.84	228.96	2.88
4E7	8.00000	231.84	228.96	2.88
4M8	8..15472	234.36	232.11	2.25
2M8	8.30930	236.40	235.26	1.14
10	9.00000	250.31	249.33	0.98
3M2E7	9.25230	252.39	254.47	-2.08
345MMM7	9.34974	253.10	256.45	-3.35
23MM4E6	9.34974	253.19	256.45	-3.26
45MM8	9.25230	254.78	254.47	0.31
2M3E7	9.25230	254.91	254.47	0.44
5M9	9.15472	255.25	252.48	2.77
24MM4E6	9.53416	256.89	260.21	-3.32
23MM8	9.40702	257.26	257.62	-0.36
344MMM7	9.51718	257.52	259.86	-2.34
2345MMMM6	9.75676	257.77	264.74	-6.97
4M3E7	9.09758	259.35	251.32	8.03
36MM8	9.30944	259.53	255.63	3.90
35MM8	9.30944	260.16	255.63	4.53
2M4E7	9.30944	260.16	255.63	4.53
245MMM7	9.56174	261.64	260.77	0.88
236MMM7	9.71644	262.16	263.92	-1.76
22MM4E6	9.67430	262.85	263.06	-0.22
244MMM7	9.75902	262.85	264.79	-1.94

Table 2 — The ABC index of compounds in the external prediction set as well as the experimental[26] and calculated standard gas phase heats at formation at 298 K.

Alkane	ABC Index	$-\Delta H°_f$ (kJ / mol) obsd.	$-\Delta H°_f$ (kJ / mol) calcd.	Residual
5	4.00000	146.77	147.47	0.70
3M5	5.15472	172.10	170.99	-1.11
7	6.00000	187.7	188.21	0.51
23MM5	6.40688	198.0	196.50	-1.50
22MM5	6.67416	205.9	201.95	-3.95
3M7	7.15472	212.5	211.74	-0.76
33MM6	7.44944	220.0	217.74	-2.26
234MMM5	7.98360	225.5	228.62	3.12
3M8	8.15472	231.94	232.11	0.17
3M4E7	9.09758	252.39	251.31	-1.08
22MM3E6	9.58718	253.31	261.29	7.98
4M9	9.15472	255.33	252.48	-2.85
3M5E7	9.15472	257.64	252.48	-5.16
5E2M7	9.30944	259.58	255.63	-3.95
25MM3E6	9.37816	261.72	257.03	-4.69
33MM8	9.44958	262.94	258.49	-4.45

Recall that the numbers $m_{x,y}$ are not mutually independent. In the case of alkanes, they are interrelated as

$$6m_{1,2} + 4m_{1,3} + 3m_{1,4} = 2m_{2,3} + 3m_{2,4} + 4m_{3,3} + 5m_{3,4} + 6m_{4,4} + 12$$

Furthermore, because the numbers of vertices of degree 2, 3 & 4 are given by

$$(m_{1,2} + 2m_{2,2} + m_{2,3} + m_{2,4})/2$$
$$(m_{1,3} + m_{2,3} + 2m_{2,3} + m_{3,4})/3$$
$$(m_{1,4} + m_{2,4} + m_{3,4} + 2m_{4,4})/4$$

respectively, these expressions must be integer-valued.

From Eq. (4) it follows that if two molecular graphs have equal $m_{x,y}$ - values for all edge types, then their ABC indices necessarily coincide. Then also their χ values coincide. In such cases we may say that the extents of branching of the respective two molecules are equal. However, the ABC indices may coincide also if the extents of branching are clearly different. In particular, edges of the type (1,2), (2,2), (2,3) & (2,4) have precisely same (unit) contributions to ABC. For instance, nonane, 3-ethylheptane, 4-ethylheptane and 3,3-diethylpentane, all have ABC = 8.00, inspite of the fact that nonane is unbranched, 3- & 4-ethylheptanes are moderately branched, whereas 3,3-diethylpentane is highly branched. This detail clearly illustrates the fact that the ABC index is by no means a measure of molecular branching. This property of ABC is a desired one, as far as we are intending to use it for predicting heats of formation and similar physico-chemical properties. It is worth noting that for the four above mentioned C_9-alkane isomers the standard enthalpies of formation have remarkably close values: -228.7, -231.8, -231.8 and -232.8 kJ/mol for nonane, 3- ethylheptane, 4-ethylheptane and 3,3-diethylpentane, respectively.

Describing and predicting heats of formation of alkanes

Among the most important attributes that a molecular descriptor needs to have is its ability to describe at least one experimental property of molecules[25]. This is a necessary attribute in order to consider a graph invariant as a molecular structure-descriptor. In this work we have selected the standard heats of formation of 64 alkanes to test the usability of the ABC index in QSPR studies. The quality of the correlation between heats of formation and ABC can be seen from Fig. 2, which should be compared with Fig. 1.

Alkanes are commonly employed for testing topological descriptors because they are the simplest organic compounds in which electronic and non-

Table 3 — Values serving as a test of the model, Eqs (5) and (6); H_n is the experimental gas phase standard enthalpy of formation (taken with opposite sign) of the normal alkane C_nH_{2n+2} ; recall that if our model were exact, then $H_{n+1} - H_n$ and $nH_n - (n-1)H_{n+1}$ would coincide with $b = 20.37$ and $a = 65.98$, respectively.

n	$H_{n+1} - H_n$	$nH_n - (n-1)H_{n+1}$
3	19.99	62.69
4	21.11	62.33
5	20.26	65.73
6	20.67	63.68
7	21.00	61.70
8	20.00	68.70
9	21.61	55.82
10	21.20	58.90

bonding interactions are minimised and most of their properties are directly dependent on topological features, such as the connectivity between their atoms. The series of alkanes was divided into two subsets, 48 compounds were used as a training set and the remaining 16 compounds, chosen at random, were placed in an external prediction set. Compounds in the external prediction set were never used in the development of regression model.

The best linear regression model obtained for the training set is:

$$-\Delta H^\circ_f(\text{kJ / mol}) = 65.98 + 20.37\ \text{ABC} \qquad \ldots (5)$$
$$N = 48 \qquad R = 0.9970 \qquad s = 3.12$$

These results should be compared with what earlier was given for the connectivity index. It is obvious that ABC is superior to χ as far as enthalpies of formation are concerned.

As can be seen from the above statistical parameters, the method for calculating the standard heats of formation of alkanes on the basis of the ABC index can be considered as a good QSPR model. In Table 1 we give the value of the ABC index as well as the experimental and calculated heats of formation for alkanes in the training set.

The most important aspect that a QSPR model needs to have is a good predictive ability for compounds not included in the training set. In order to test the predictive ability of our model we computed the standard heats of formation for the 16 alkanes in the external prediction set. The experimental and calculated values of $-\Delta H^\circ_f$ are shown in Table 2.

The correlation coefficient of experimental versus calculated $-\Delta H^\circ_f$ in the prediction set is equal to 0.9956 and the root mean square error for the prediction is 3.47 kJ /mol. These parameters show the very good predictive features of our model, e.g., it explains more than 99% of the variance of the heats of formation of alkanes not included in the training set, and the prediction error is only 11 % higher than the one obtained for the training set.

As already discussed, the ABC index does not have a high isomer-discriminating power, since there are pairs triples, etc., of isomers with the same ABC-value. However, if we examine Tables 1 and 2 we see that most alkanes with degenerate ABC indices have very close heats of formation. For instance, both 3,4,5-trimethylheptane and 2,3-dimethyl-4-ethylhexane have ABC = 9.34974 and their values of $-\Delta H^\circ_f$ differ by only 0.09 kJ/ mol. The example of nonane, 3-ethylheptane, 4-ethylheptane and 3,3-diethylpentane has been discussed previously. These examples corroborate the conclusion that isomer-discriminating power and ability to describe physicochemical properties are not necessarily related, i.e., highly discriminant indices need not produce good QSPR models, and vice versa .

Our novel graph invariant - the ABC index - is based on simple and immediately recognisable structural features of the molecular graph. Consequently, ABC is quite easy to compute. Bearing in mind that the index is based on vertex (atom) and edge (bond) degrees, it is not difficult to extend it to molecules containing heteroatoms. For instance, if we use the Kier-Hall valence degrees instead of vertex degrees, then we can calculate a valence ABC index by using the expression (3). The definition, analysis and applications of this valence ABC index will be elaborated in a forthcoming article.

Physical meaning of the model

The parameters a and b in our model, namely

$$-\Delta H^\circ_f = a + b\ \text{ABC} \qquad \ldots (6)$$

(cf. Eq. (5)) have the following physical meaning. For the sake of brevity denote by H_n the experimen-

tal value for the standard gas phase enthalpy of formation (taken with opposite sign) of the normal alkane with n carbon atoms.

If X_1 is an alkane possessing a CH_2 -group and if X_2 is obtained by formally replacing this CH_2-group by a CH_2CH_2-fragment, then the respective ABC values will differ by exactly 1 : $ABC(X_2) = ABC(X_1) + 1$. Consequently, according to the model, Eq. (6) the heat of formation of X_2 is predicated to differ by b (kJ /mol) from the heat of formation of X_1. In other words, the slope *b* is just the change of the enthalpy of formation when an alkyl chain is extended by one C-atom. In particular, the slope *b* should correspond to the difference H_{n+1}-H_n for any, sufficiently large, value of n. The respective values of H_{n+1}-H_n are found in Table 3, indicating the plausibility and quality of the model proposed in this paper, Eqs (5) and (6).

The physical meaning of the intercept *a* in Eq. (6) is seen from the following reasoning. As already mentioned, all edges possessing an end vertex of degree two have equal, unit, contributions to the ABC index. As a consequence, all edges of the molecular graph of a normal alkane have equal unit, contributions to ABC and if n is the number of carbon atoms, then ABC=n-1. The respective contribution to the enthalpy of formation (i.e., to quantity H_n) is then (n-1) *b* (kJ/mol). On the other hand, the actual enthalpy of formation differs from this value because the expression (n-1)*b* does not take into account the fact that the carbon-atom chain of the molecule has two ends. Consequently, the intercept *a* in our model, Eq. (6) may be understood as the correction term reflecting the terminal-group effects on the enthalpy of formation. In the shorthand notation introduced above,

$$a = H_n - (n-1)b = H_n - (n-1)(H_{n+1} - H_n) = n H_n - (n-1) H_{n+1} \quad ...(7)$$

Recall that the right-hand side of Eq. (7) is an experimental value, which may only approximately agree with the parameters *a*. In Table 3 are found these experimental values, calculated for various values of n.

Acknowledgement

One of the authors (IG) thanks the Centre of Chemical Bioactive Agents of the Central University in Santa Clara, Cuba for hospitality in November/December 1997 and to the Third World Academy of Science for financial support.

References

1 Mihalic Z & Trinajastic N, *J chem Educ*, 69 (1992) 701.
2 Randic M, *J chem Inf Comput Sci*, 37 (1997) 672.
3 Randic M, *J Am chem Soc*, 97 (1975) 6609.
4 Kier L B & Hall L H, *Molecular connectivity in chemistry and drug research* (Academic Press, New York), 1976.
5 Kier L B & Hall L H, *Molecular connectivity in structure-activity analysis* (Wiley-Research Studies Press, New York), 1986.
6 Gálvez J, García-Domenech R, de Julián-Ortiz J V & Soler R, *J chem Inf Comput Sci*, 35 (1995) 272.
7 Randic M, Hansen P J & Jurs P C, *J chem Inf Comput Sci*, 28 (1988) 60.
8 Kunz M, *Colln Czech chem Commun*, 55 (1990) 630.
9 Estrada E, *J chem Inf Comput Sci*, 35 (1995) 1022.
10 Balaban A T, *Chem Phys Lett*, 89 (1982) 399.
11 Estrada E, *J chem Inf Comput Sci*, 35 (1995) 31.
12 Estrada E, *J chem Inf Comput Sci*, 35 (1995) 701.
13 Estrada E & Ramrez A, *J chem Inf Comput Sci*, 36 (1996) 837.
14 Harary F, *Graph theory* (Addison-Wesley, Reading), 1969.
15 Gutman I & Estrada E, *J chem Inf Comput Sci*, 36 (1996) 541.
16 Estrada & Gutman I, *J chem Inf Comput Sci*, 36 (1996) 850.
17 Diudea M V, Horvath D & Bonchev D, *Croat chem Acta*, 68 (1995) 131.
18 Gutman I, Popovic L, Mishra B K, Kuanar M, Estrada E & Guevara N, *J Serb chem Soc*, 62 (1997) 1025.
19 Gutman I, Popovic L, Estrada E & Bertz S H, *ACH Models in Chem*, accepted for publication.
20 Estrada E, *J chem Inf Comput Sci*, 36 (1996) 844.
21 Estrada E, Guevara N & Gutman I, *J chem Inf Comput Sci*, submitted for publication.
22 Bertz S H, *J chem Soc chem Commun* (1981) 818.
23 Bertz S H, Discrete *appl Math*, 19 (1988) 65.
24 Bertz S H, in R B (Ed) *Chemical applications of topology and graph theory* (Elsevier, Amsterdam), 1983, 206.
25 Randic M, *J math Chem*, 7 (1990) 155.
26 *CRC Handbook of chemistry and physics*, 61th *Edn* (CRC Press, Boca Raton), 1981.