communications in mathematical
match
and in computer chemistry

no. 44, October 2001
MATCDY (44) 155 - 167 (2001)
ISSN 0340 - 6253

The Balaban *J* Index in the Multidimensional Space of Generalized Topological Indices. Generalizations and QSPR Improvements♣

Ernesto Estrada and Yaquelin Gutierrez
Faculty of Pharmacy, Department of Organic Chemistry, University of Santiago de Compostela, Santiago de Compostela 15706, SPAIN and GMIXON, 848 Chemin du Carreyrat, 8200 Montauban, France.

Abstract. A generalization of topological indices based on a vector-matrix-vector multiplication procedure is used to represent the differences between some "classical" topological indices and the Balaban *J* index in a radial graphic. The same approach is used to optimize the Balaban *J* index for describing the motor octane number (MON) of octane isomers. After this optimization the optimal Balaban *J*** index is obtained which give a correlation coefficient of 0.9829 with MON. A cubic model between MON and *J*** index produces an excellent correlation of 0.9919 and a standard deviation of 3.51.

INTRODUCTION

The history of topological indices can be traced back to the introduction of the Wiener *W* index in 1947 [1]. Further developments conducted to the introduction of several molecular descriptors, such as Zagreb group M_1 and M_2 indices [2,3]; Hosoya *Z* index [4]; Randić χ index [5] and so forth. These first topological indices were useful in structure-property and structure-activity relationship studies [6], especially the Randic index and its extensions carried out by Kier and Hall [7,8]. However, these molecular descriptors were not too much discriminant for isomers, which is an expected and desired property for an index describing the topology of a molecule [9]. In this context Prof. A. T. Balaban introduced that he called a "highly discriminant distance-based topological index" [10]. This index, which is now known

♣ This work is dedicated to Prof. Alexandru T. Balaban in occasion of his 70th anniversary.

as the Balaban J index, was defined on the basis of the Randić formula but using distance sums for vertices instead of vertex degrees [10].

The Balaban J index shows a good isomer discrimination ability and produces good correlations with some physical properties, such as the motor octane number (MON) [11]. This index appears in theoretical models for predicting and screening drug candidates in rational drug design strategies [12]. It has been modified to account for heteroatom differentiation [13] and very recently the strategy of "variable molecular descriptors" has been applied to it [14]. According to this strategy Randić and Pompe have improved the correlation of J index with MON from 0.9277 to 0.9537 and the use of an index $1/J^*$ produced a correlation of 0.9739 [14].

In a recent work, Estrada proposed a generalization of some of the most important topological indices described in the literature by using a vector-matrix-vector multiplication procedure [15]. This approach uses a novel variable graph theoretical matrix $\mathbf{\Gamma}$, and a couple of novel variable vectors, $\mathbf{y}$ and $\mathbf{z}$. The adjacency $\mathbf{A}$ and distance $\mathbf{D}$ matrices of a (molecular) graph are now particular cases of the new matrix $\mathbf{\Gamma}$, and Wiener W index [1], Zagreb group M_1 and M_2 indices [2, 3], Randic χ connectivity index [5], Harary H number [16-18] and the Balaban J index [10] are derived from the same invariant [15]. In fact, the variable molecular descriptors strategy of Randić is a particular case of the vector-matrix-vector multiplicative approach with the generalized graph matrix. Here we will use this approach to study the relation between the J index and some other topological indices as well as to generalize and improve the correlation ability of this index for describing MON.

SOME "CLASSICAL" TOPOLOGICAL INDICES

The classical topological indices that we will study in the current work can be grouped in two sets according to the matrix from they are derived. The first group is formed by indices related to the adjacency matrix of the molecular graph. These indices are the Zagreb group indices M_1 and M_2 defined on the basis of the vertex degrees, which are the sums of the rows or columns of the $\mathbf{A}$ matrix [2, 3]:

$$M_1 = \sum_i (\delta_i)^2 \tag{1}$$

$$M_2 = \sum_k (\delta_i \delta_j)_k \tag{2}$$

In M_2 the summation is carried out over all adjacent vertices (edges) of the graph. Intimately related to these indices is the Randić connectivity index [5], which is defined in a similar way than M_2 but using an exponent –0.5 in the invariant:

$$\chi = \sum_k \left(\delta_i \delta_j\right)_k^{-0.5} \tag{3}$$

The other group of descriptors are related to the distance matrix of the molecular graph. The first of them, the Wiener index, is simply the half-sum of all elements of such matrix [1]:

$$W = \frac{1}{2}\sum_{ij} d_{ij} = \frac{1}{2}\sum_i s_i \tag{4}$$

where s_i is the distance sum, that is the sum of the ith row or column of the distance matrix. The Balaban J index uses these distance sums instead of the vertex degrees in an expression analogous to (3) [10]:

$$J = C \cdot \sum_k \left(s_i s_j\right)_k^{-0.5} \tag{5}$$

where $C = m/(1+\mu)$; m is the number of edges and μ is the cyclomatic number.

Another distance-based topological index that will be studied here is the so-called Harary number H [16-18]:

$$H = \frac{1}{2}\sum_{ij} d_{ij}^{-k} \tag{6}$$

where d_{ij}^{-k} are the elements of the distance matrix for which the non-diagonal entries have been raised to the power $-k$. This index has been indistinctly defined by using k = -1 and k =-2 [16-18]. We will call these indices H_1 and H_2, respectively.

It is not hard to see that most of these topological indices appear unrelated to each other making generalizations and interpretations quite difficult.

VECTOR-MATRIX-VECTOR MULTIPLICATION PROCEDURE

In a couple of papers Estrada et al [19, 20] proposed a vector-matrix-vector multiplication procedure to extend and improve the QSPR quality of Wiener-like and MTI-like topological indices. According to this procedure Zagreb group indices are defined as follows [19]:

$$M_1 = \mathbf{v} \cdot \mathbf{A} \cdot \mathbf{u}^{\mathrm{T}} \tag{7}$$

$$M_2 = \frac{1}{2}\left(\mathbf{v} \cdot \mathbf{A} \cdot \mathbf{v}^{\mathrm{T}}\right) \tag{8}$$

where **A** is the adjacency matrix, **v** is a vector of vertex degrees and **u** is a unitary vector, and T is used for transpose. The Randić index is defined by using a vector $\mathbf{v}'$ of vertex degrees raised to –0.5, $\left(\delta_1^{-0.5} \quad \delta_2^{-0.5} \quad \cdots \quad \delta_n^{-0.5}\right)$ [15]:

$$\chi = \frac{1}{2}\left(\mathbf{v}' \cdot \mathbf{A} \cdot \mathbf{v}'^{\mathrm{T}}\right) \tag{9}$$

The distance-based indices are defined as follows [19, 15]:

$$W = \frac{1}{2}\left(\mathbf{u} \cdot \mathbf{D} \cdot \mathbf{u}^{\mathrm{T}}\right) \tag{10}$$

$$J = \frac{1}{2}C\left(\mathbf{s}' \cdot \mathbf{A} \cdot \mathbf{s}'^{\mathrm{T}}\right) \tag{11}$$

where $\mathbf{s}' = \left(s_1^{-0.5} \quad s_2^{-0.5} \quad \cdots \quad s_n^{-0.5}\right)$.

Finally, the Harary numbers are defined as [20]:

$$H = \frac{1}{2}\left(\mathbf{u} \cdot \mathbf{D}^{-k} \cdot \mathbf{u}^{\mathrm{T}}\right) \tag{12}$$

where $\mathbf{D}^{-k}$ is the distance matrix with non-diagonal entries raised to the power $-k$.

GENERALIZED TOPOLOGICAL INDICES

The generalization of topological indices proposed by Estrada is based on the introduction of a novel variable graph matrix called the generalized graph matrix and a couple of vectors defined as follows [15]:

DEFINITION 1: Let $\mathbf{\Gamma}(x,p) = \left|g_{ij}(x,p)\right|_{nxn}$ be the generalized graph matrix, which is a square symmetric matrix whose elements g_{ij} are defined as follows:

$$g_{ij} = \begin{cases} 1 \text{ iff } d_{ij} = 1 \\ \left(d_{ij} x^{d_{ij}-1}\right)^p \text{ iff } i \neq j; d_{ij} \neq 1 \\ 0 \text{ otherwise} \end{cases}$$

We can introduce a change of variable in $\mathbf{\Gamma}(x, p)$ so that x is substituted by y or z and p = 1, we obtain the following matrices: $\mathbf{\Gamma}(y,1)$ and $\mathbf{\Gamma}(z,1)$. Using these matrices two new graph-theoretical vectors that will be used in the generalization of TIs are introduced as follows [15]:

DEFINITION 2: Let $\mathbf{y}(w, y, q)$ and $\mathbf{z}(s, z, r)$ be two vectors of order n whose elements y_i and z_i are defined as follows:

$$y_i = \left(w_i + \sum_j g_{ij}(y,1) \right)^q \qquad z_i = \left(s_i + \sum_j g_{ij}(z,1) \right)^r$$

It is straightforward to show that the adjacency and distance matrices are particular cases of the current generalized matrix as well as the distance matrices with non-diagonal entries raised to –1 and –2, now called Harary matrices **H1** (also called reverse distance **RD**) and **H2**, respectively:

$$\mathbf{\Gamma}(0,1) = \mathbf{A}, \qquad \mathbf{\Gamma}(1,1) = \mathbf{D}, \ \mathbf{\Gamma}(1,-1) = \mathbf{RD} = \mathbf{H1}, \ \mathbf{\Gamma}(1,-2) = \mathbf{H2}$$

Another general matrix that has as particular case two graph matrices (**D** and **RD**) was recently introduced [21]. The vectors used in expressions (7)-(11) are also particular cases of **y** or **z**.

The following generalized graph invariant based on the vector-matrix-vector multiplication procedure has been used to generalize the topological indices studied in the current work [15]:

DEFINITION 3: Let $\Im$ be a generalized vector-matrix-vector invariant defined as follows:

$$\Im = C[\mathbf{y}(y, w, q)\mathbf{\Gamma}(x, p)\mathbf{z}(z, s, r)]$$

By using this expression we can obtain all adjacency-based and distance-based topological indices under study in the current work. The mathematical expressions for these "classical" indices are given below:

$$M_1 = [\mathbf{y}(0,0,1) \cdot \mathbf{\Gamma}(0,1) \cdot \mathbf{z}(0,0,0)] \tag{13}$$

$$M_2 = \frac{1}{2}[\mathbf{y}(0,0,1) \cdot \mathbf{\Gamma}(0,1) \cdot \mathbf{z}(0,0,1)] \tag{14}$$

$$\chi = \frac{1}{2}[\mathbf{y}(0,0,-0.5) \cdot \mathbf{\Gamma}(0,1) \cdot \mathbf{z}(0,0,-0.5)] \tag{15}$$

$$W = \frac{1}{2}[\mathbf{y}(1,0,0) \cdot \mathbf{\Gamma}(1,1) \cdot \mathbf{z}(1,0,0)] \tag{16}$$

$$J = \frac{1}{2}[\mathbf{y}(1,0,-0.5)\cdot\boldsymbol{\Gamma}(0,1)\cdot\mathbf{z}(1,0,-0.5)] \quad (17)$$

$$H_1 = \frac{1}{2}[\mathbf{y}(1,0,0)\cdot\boldsymbol{\Gamma}(1,-1)\cdot\mathbf{z}(1,0,0)] \quad (18)$$

$$H_2 = \frac{1}{2}[\mathbf{y}(1,0,0)\cdot\boldsymbol{\Gamma}(1,-2)\cdot\mathbf{z}(1,0,0)] \quad (19)$$

MULTIDIMENSIONAL REPRESENTATIONS

According to the generalized vector-matrix-vector multiplication procedure all topological indices studied here are points in a multidimensional space. A radial graphic representing any of these indices is composed by eight axis: *y*, *w*, *q*, *x*, *p*, *z*, *s*, and *r*, and an octagonal figure that represents the index. For instance, the Balaban *J* index is represented in Figure 1.

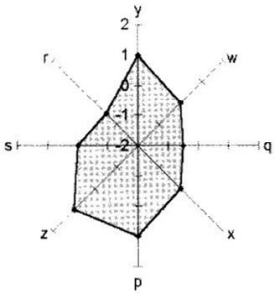

FIGURE 1. Radial graphic representation of the Balaban *J* index. The axis of the graphic are the parameters of the generalized vector-matrix-vector multiplication procedure.

In order to observe the main similarities and differences between the Balaban *J* index and the rest of topological indices studied here we will use the radial representation of the differences between *y*, *w*, *q*, *x*, *p*, *z*, *s*, and *r* of the indices and *J* index. In this way, the Balaban index will be represented as a regular octagon because all values of parameters will be equal to zero (a consequence of resting *J*-*J*). The rest of the radial graphics represent the differences between the corresponding index and the Balaban *J* index. In Figure 2 we show this representation for *J* index and adjacency-based descriptors studied here. As can be seen the

greater similarities are with the Randić connectivity index (it is the closest to a regular octagon of the three adjacency-based indices) due to the fact that both indices are based on the same graph invariant but the Balaban one using distance sums instead of vertex degrees.

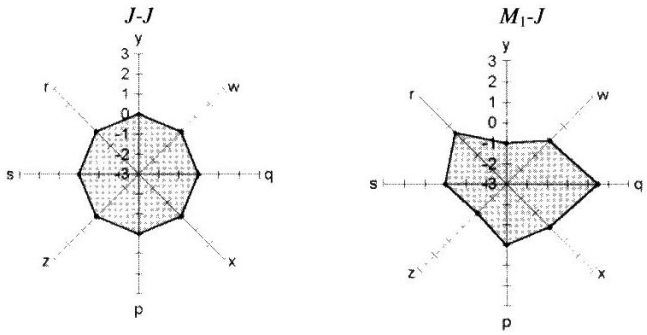

FIGURE 2. Radial graphic representation of the differences between adjacency-based topological indices and the Balaban *J* index.

In Figure 3 we represent the differences between Balaban *J* index and distance-based topological indices studied here. It is observed that no one of these indices is closely related to the Balaban index. These differences can be explained by the fact that graph-invariants used in the definition of Wiener and Harary numbers are based on the sum of all elements in the distance matrix while *J* index is based on the multiplication of vertex properties of adjacent vertices. In fact, we can say that the Balaban *J* index is more closely related to the connectivity index than to the Wiener-like topological indices based on the distance matrix.

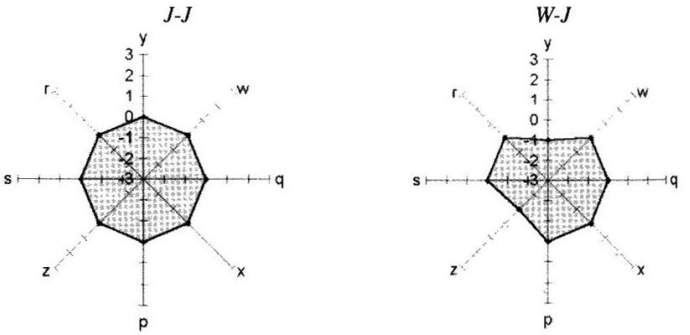

FIGURE 3. Radial graphic representation of the differences between distance-based topological indices and the Balaban *J* index.

INDEX OPTIMIZATION

In order to make an optimization of the Balaban *J* index we have selected the motor octane number (MON) of octane isomers. The values of MON for n-octane and 2,2,3,3,-tetramethylbutane are not available, thus the number of octane isomers is $n = 16$. Because for octane isomers the value of C in the expression for calculating J index is constant we will not consider it in calculating this index. The optimization of this index according to our approach of generalized vector-matrix-vector multiplication procedure consists simply in changing systematically the values of the eight parameters from which any of the studied indices depends on. For instance, if we want to derive the indices studied by Randić and Pompe as variable molecular descriptors [14] based on distance related matrices we have only to change the values of w and s from a minimal value to a maximal one with a given step. This procedure permits the definition of an infinite number of topological indices from which the known ones are particular cases. They are only single drops in a ocean of descriptors. However, this possibility represents a real challenge from the computational point of view. For instance, if we consider simultaneous variations in all eight axis from 0 to 0.9 with a step of 0.1 we generate 100 000 000 (!) indices. In order to avoid this numerical explosion we will consider the following strategy. For a given value of s and w we maintain constant x, p, q and r and change systematically y and z. After correlation with MON we will obtain the optimal

value of y and z for this value of s, w. Then, we change s and w and repeat the process. After we have determined the optimal values of s, w, y and z, we will proceed to optimize the rest of parameters. For instance, taken into consideration the previous work of Randić and Pompe [14], we select $s = w = 100$ and optimized y and z. The values of y and z were changed in the range –0.6 to 1.4 with a step of 0.2 (the range is equal to ± 0.6 the values of the original Balaban index [10]). For this value of s and w, the optimal values of y and z resulted $y = -0.2$ and $z = 0.6$. This set of parameters give a correlation coefficient with MON of $R = 0.9791$, an standard deviation of $sd = 5.19$ and a Fisher ratio of $F = 325.2$. This value represents an improvement respect to the best regression obtained by Randić and Pompe with J^* index (R = 0.9537, with the variable parameter of 100) [14]. Then, we changed $s = w$ to 90 and obtain the optimal values of y and z. After these values were obtained, s and w were changed with a step of 10, and the optimal values of y and z were determined. The results of these first sets of optimizations are given in Table 1.

TABLE 1. Correlation coefficients and standard deviations of the linear models for describing motor octane number of octane isomers versus modified Balaban J indices changing s and w parameters. In all cases the optimal values of $y = -0.2$ and $z = 0.6$ were found.

s, w	R	sd	F
100	0.9791	5.19	325.2
90	0.9794	5.16	328.8
80	0.9792	5.18	326.0
70	0.9794	5.16	329.8
60	0.9795	5.15	330.4
50	0.9797	5.12	334.4
40	0.9801	5.08	340.7
30	0.9806	5.01	349.7
20	0.9814	4.91	365.7
15	0.9817	4.86	373.1
10	0.9805	5.02	348.2

As can be seen in this table the optimal set of parameters is $s = w = 15$, $y = -0.2$, $z = 0.6$, which give a correlation coefficient with MON of 0.9817. Keeping these optimal values constant, we changed q and r in the range from –1 to 1 with step of 0.2. The 121 indices so-derived were correlated to MON determining the optimal values of q and r, which resulted to be $q = -1$ and $r = -0.8$. This optimal set of parameters produced a correlation coefficient with

MON of $R = 0.9829$ and a standard deviation of $sd = 4.70$. The values of J^{**} as well as J and MON are given in Table 2. Further attempts to optimize x and p maintaining constant the rest of optimal parameters do not produced any significant improvement in the correlation with MON. As a point of comparison the best correlation obtained by Randić and Pompe with the variable descriptor strategy give $R = 0.9739$ and $sd = 5.80$ with the index $1/J^*$ (x = -25) [14]. Thus, the current model represents an improvement of 19 % in the standard deviation respect to the best model derived in the previous optimization of J index.

By this means, the optimal J index (J^{**}) for describing MON derived with the generalized vector-matrix-vector multiplication procedure can be written as follows:

$$J^{**} = \frac{1}{2}\left[\mathbf{y}(-0.2,20,-1)\cdot\mathbf{\Gamma}(0,1)\cdot\mathbf{z}(0.6,20,-0.8)\right] \qquad (20)$$

A representation of this index in a radial 8-dimensional graphic is given in Figure 4 together with the original Balaban J index.

TABLE 2. Values of Balaban J index, optimal J^{**} index and motor octane number (MON) for octane isomers.

isomer	J	J^{**}	MON
n	0.36144	0.03653	-
2M	0.38798	0.03613	23.8
3M	0.40887	0.03596	35.0
4M	0.41709	0.03587	39.0
2,5M	0.41826	0.03572	55.7
3E	0.43919	0.03570	52.4
2,4M	0.44269	0.03555	69.9
2,2M	0.44454	0.03554	77.4
2,3M	0.45297	0.03553	78.9
3,4M	0.47035	0.03544	81.7
3,3M	0.48191	0.03535	83.4
2M3E	0.47927	0.03535	88.1
2,3,3M	0.52976	0.03507	99.4
2,2,4M	0.48413	0.03513	100.0
2,3,4M	0.49489	0.03519	95.9
3M3E	0.51189	0.03524	88.7
2,2,3M	0.51761	0.03508	99.9
2,2,3,3M	0.57434	0.03477	-

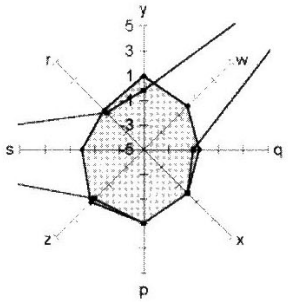

FIGURE 4. Radial graphic representation of the optimal *J*** index for describing MON and the original Balaban *J* index, which is shadowed.

The expression of the optimal Balaban index for describing MON of octane isomers can be expressed in terms of distance sums as the original *J* index by using the following expression:

$$J^{**} = \frac{1}{2}\sum_{k}\left[\left(s_i^{*}\right)^{q}\left(s_j^{**}\right)^{r} + \left(s_i^{**}\right)^{r}\left(s_j^{*}\right)^{q}\right]_k \qquad (21)$$

where

$$s_i^{**} = 15 + \sum_i d_{ij}(0.6)^{d_{ij}-1} \text{ and } s_i^{*} = 15 + \sum_i d_{ij}(-0.2)^{d_{ij}-1}$$

A plot of *J*** index versus MON shows a polynomial dependence between both variables. Consequently, we have fitted these variables by a polynomial function obtaining the best QSPR model for describing octane isomers MON with modified Balaban *J* index. The correlation coefficient of this regression model is $R = 0.9919$ and the standard deviation is $sd = 3.51$ (see Figure 5). The fit obtained here is with a polynomial of order three, while Randic and Pompe obtained $R = 0.9910$ and $s = 3.54$ with a polynomial of order two and their index $1/JJ^*$ [14].

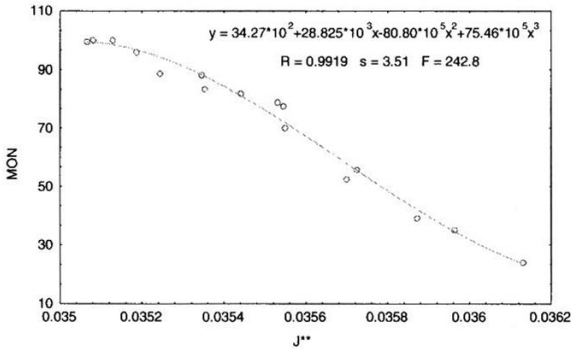

FIGURE 5. Plot of motor octane number (MON) versus the optimized Balaban J index (J^{**}) for octane isomers.

CONCLUSIONS

We have shown that most of the "classical" topological indices can be derived from the generalized vector-matrix-vector multiplication procedure. Here we have concentrated our attention on the Balaban J index. We have optimized this descriptor for improving its ability to correlate with octane isomers MON. As previously stated by Randic the optimization will in general depend on the property considered [14]. Thus, in case we were interested in describing boiling point or any other physicochemical property different from MON, we have to find the optimal values of y, w, q, x, p, z, s, and r. We believe that this procedure as well as the variable molecular descriptors proposed by Randić, which is a particular case of the current approach, will permit better description of physicochemical and biological properties of organic compounds with the existing topological indices.

We believe that the current approach of generalized vector-matrix-vector multiplication procedure will be meaningful in the fructification of Balaban's efforts in the search of novel topological indices as well as in their generalization and interpretation.

REFERENCES

[1] H. Wiener, *J. Am. Chem. Soc.* **69**, 17 (1947).

[2] I. Gutman, B. Ruscik, N. Trinajstić and C. F. Wilcox, *J. Chem. Phys.* **62**, 3399 (1975).

[3] I. Gutman and N. Trinajstic, *Chem. Phys. Lett.* **17**, 535 (1972).

[4] H. Hosoya, *Bull. Chem. Soc. Japan* **44**, 2332 (1971).

[5] M. Randic, *J. Am. Chem. Soc.* **97**, 6609 (1975).

[6] A. Sablijć and N. Trinajstić, *Acta Phram. Jugosl.* **31**, 189 (1981)

[7] L. B. Kier and L. H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.

[8] L. B. Kier and L. H. Hall, Molecular Connectivity in Structure-Activity Analysis, research Studies Press, Letchworth, 1986.

[9] M. Razinger, J. R. Chretien and J. E. Dubois, *J. Chem. Inf. Comput. Sci.* **25**, 23 (1985).

[10] A. T. Balaban, *Chem. Phys. Lett.* **89**, 399 (1982).

[11] A. T. Balaban, L. B. Kier and N. Joshi, *Commun. Math. Comput. Chem.* (*MATCH*), **28**, 13, (1992).

[12] G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floch and R. Buelow, Nature Biotech. **16**, 748 (1998).

[13] O. Ivanciuc and A. T. Balaban, *Commun. Math. Comput. Chem.* (*MATCH*), **30**, 117 (1994).

[14] M. Randić and M. Pompe, *J. Chem. Inf. Comput. Sci.* **41** (2001), in press.

[15] E. Estrada, *Chem. Phys. Lett.* (2001) in press.

[16] Z. Mihalić and N. Trinajstić, *J. Chem. Educ.* **69**, 701 (1992).

[17] D. Plavšić, S. Nikolić, N. Trinajstić and Z. Mihalić, *J. Math. Chem.* **12**, 309 (1993).

[18] O. Ivanciuc, T.-S. Balaban and A. T. Balaban, J. Math. Chem. **11**, 70 (1992).

[19] E. Estrada, L. Rodríguez, and A. Gutierrez, *Commun. Math. Comput. Chem.* (*MATCH*), **35**, 145 (1997).

[20] E. Estrada and L. Rodríguez, *Commun. Math. Comput. Chem.* (*MATCH*), **35**, 157 (1997).

[21] O. Ivanciuc, Rev. Roum. Chim. **44**, 519 (1999).