Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemical Physics Letters 410 (2005) 343–347
CHEMICAL PHYSICS LETTERS
www.elsevier.com/locate/cplett

Generalised topological indices: Optimisation methodology and physico-chemical interpretation

Adelio R. Matamala [a,*], Ernesto Estrada [b]

[a] Faculty of Chemical Sciences, QTC Group, Universidad de Concepción, Casilla 160-C, Concepción, Chile
[b] Complex Systems Research Group, X-Rays Unit, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, Santiago de Compostela 15782, Spain

Received 27 April 2005; in final form 20 May 2005
Available online 21 June 2005

Abstract

A new methodology that combines the generalized topological indices (GTI) and the down hill *simplex* optimisation procedure has been developed to search for optimised quantitative structure–property relationship models. The structural interpretation of the optimal molecular descriptors is attained analytically by means of the GTIs decomposition in terms of geodesics (shortest paths) in the molecular graph. This approach provides a clear explanation about the role of topological structure in the study of molecular physico-chemical properties.

1. Introduction

Molecular descriptors are the basic components of the quantitative structure–property (QSPR) and structure–activity (QSAR) relationships. The so-called topological indices (TIs) are molecular descriptors based on a graph-theoretical representation of the molecular structure [1]. They have proved to be very useful in real-world applications of QSPR and QSAR to molecular design and physico-chemical studies of molecular properties [2]. Basically, any QSPR/QSAR based on TIs assumes the existence of a correlation between a (molecular or molar) property/activity and the molecular structure described at a topological level, where the linear model is the simplest representative [3,4]. Within graph-based QSPR/QSAR studies, the essential information contained in the molecular graphs must be coded using numerical invariants. Of course, the adequate definition of these invariants is a critical point of the method. The introduction of topological indices for the study of each particular property remains as a heuristic process and each property requires an ad hoc graph-invariant definition. The lack of a general methodology for defining and applying TIs to QSPR/QSAR problems has generated an undesired proliferation of such molecular descriptors and has produced significant problems in their physico-chemical interpretation. Thus, the search for generalized approaches to TIs that permit their optimisation and better understanding of their meaning has become a challenging area in chemical graph theory [5–11].

In a previous Letter [12], a general approach that unifies many of the 'classical' topological indices into one theoretical framework has been introduced, opening new ways in both theoretical and applied aspects of graph invariant methods in computational physical chemistry [13–15]. Here, a general methodology for obtaining new topological indices and its interpretation is presented. Optimised invariants are obtained combining the simplex optimisation method [16–19] and the so-called generalized topological indices (GTI) [12]. The optimised molecular descriptors are then interpreted by mean of the

* Corresponding author. Fax: +56 41 245974.
E-mail address: amatamal@udec.cl (A.R. Matamala).

0009-2614/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2005.05.096

decomposition of the GTI in terms of geodesic (shortest paths) matrices [20] in the graph. The study of normal boiling points of octane isomers has been included as an example to develop the present methodology.

2. Structural interpretation of generalized topological indices

Let $G(V,E)$ be a molecular-graph with $|V| = n$ vertices and $|E| = m$ edges. Let d_{ij} be the entries of the $n \times n$ topological distance matrix of the graph $G(V,E)$. The GTI associated to the graph $G(V,E)$ is defined by the following vector–matrix–vector formula [1,12–15]:

$$\begin{aligned}\mathrm{GTI}[G] &= \frac{1}{2}\left(\begin{array}{ccc|c} p & q & r & \bar{s} \\ x & y & z & \bar{w} \end{array}\right)_G \\ &= \frac{1}{2}\mathbf{u}^{\mathrm{T}}(G;y,\bar{w},q)\Gamma(G;x,p)\mathbf{v}(G;z,\bar{s},r), \end{aligned} \tag{1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are column vectors whose components are given by

$$\mathbf{u}_i(G;y,q) = \left[w_i + \sum_{j=1}^{n} g_{ij}(y,1)\right]^q \text{ and}$$

$$\mathbf{v}_i(G;z,r) = \left[s_i + \sum_{j=1}^{n} g_{ij}(z,1)\right]^r. \tag{2}$$

The Γ matrix is the so-called *generalized molecular-graph matrix* whose $n \times n$ entries are expressed in terms of the topological distance through

$$g_{ij}(G;x,p) = \begin{cases} 1, & \text{if } d_{ij} = 1, \\ \left(d_{ij}x^{d_{ij}-1}\right)^p, & \text{if } i \neq j;\ d_{ij} \neq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The scalars $x,y,z,p,q,r,\bar{w} = (w_1,w_2,\ldots,w_n)$, and $\bar{s} = (s_1,s_2,\ldots,s_n)$, form a set of $2n + 6$ real parameters. For simplicity,

$$\left(\begin{array}{ccc|c} p & q & r & \bar{0} \\ x & y & z & \bar{0} \end{array}\right)_G = \begin{pmatrix} p & q & r \\ x & y & z \end{pmatrix}_G. \tag{4}$$

When $\bar{w} = (0,0,\ldots,0) = \bar{0}$ and $\bar{s} = (0,0,\ldots,0) = \bar{0}$. From (1), it is straightforward to obtain several of the well-known classical indices. For instance (see [12–14] for more examples), the Wiener index (W) [21] and the Randić connectivity index (χ) [22] are expressed as follows:

$$W(G) = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}_G, \quad \chi(G) = \frac{1}{2}\begin{pmatrix} 1 & -1/2 & -1/2 \\ 0 & 0 & 0 \end{pmatrix}_G. \tag{5}$$

Using the geodesic (shortest path) matrices [20] of different orders of the graph, $\Delta^{(k)}$, whose entries are defined by

$$\Delta_{ij}^{(k)}[G] = \begin{cases} 0, & \text{if } d_{ij} \neq k \text{ in graph } G, \\ 1, & \text{if } d_{ij} = k \text{ in graph } G, \end{cases} \tag{6}$$

and after some algebraic work, Eq. (1) becomes

$$\mathrm{GTI}[G] = \sum_{k=1}^{\mathrm{diam}(G)} C_k N^{(k)}[G]\xi^{(k)}[G], \tag{7}$$

where $\mathrm{diam}(G)$ is the 'diameter' of the graph G, i.e., the largest geodesic in the graph G,

$$C_k = C_k(x,p) = k^p x^{p(k-1)}, \tag{8}$$

$$N^{(k)}[G] = \sum_{i=1}^{n}\sum_{j=1+i}^{n} \Delta_{ij}^{(k)}[G] \tag{9}$$

and

$$\xi^{(k)}[G] = \frac{1}{N^{(k)}[G]}\sum_{i=1}^{n}\sum_{j=1+i}^{n}\langle i,j\rangle_G \Delta_{ij}^{(k)}[G]. \tag{10}$$

Each $\xi^{(k)}$ term is a function on $y,z,q,r,\bar{w} = (w_1,w_2,\ldots,w_n)$, and $\bar{s} = (s_1,s_2,\ldots,s_n)$, parameters. The quantity $N^{(k)}[G]$ is the number of pair vertices at distance k in the graph G and each bracket $\langle i,j\rangle_G$ is defined by

$$\langle i,j\rangle_G = \frac{1}{2}[\mathbf{u}_i(G;y,q)\times\mathbf{v}_j(G;z,r) + \mathbf{u}_j(G;y,q)\times\mathbf{v}_i(G;z,r)]. \tag{11}$$

For example, by simple inspection of Fig. 1, the $\xi^{(k)}$ for the 2,4-dimethylhexane molecule are:

$$\begin{aligned} \xi^{(1)} &= \frac{1}{7}[\langle 1,2\rangle + \langle 2,3\rangle + \langle 2,4\rangle + \langle 4,5\rangle \\ &\quad + \langle 5,6\rangle + \langle 5,7\rangle + \langle 7,8\rangle], \\ \xi^{(2)} &= \frac{1}{8}[\langle 1,3\rangle + \langle 1,4\rangle + \langle 2,5\rangle + \langle 3,4\rangle + \langle 4,6\rangle \\ &\quad + \langle 4,7\rangle + \langle 5,8\rangle + \langle 6,7\rangle], \\ \xi^{(3)} &= \frac{1}{6}[\langle 1,5\rangle + \langle 2,6\rangle + \langle 2,7\rangle + \langle 3,5\rangle + \langle 4,8\rangle + \langle 6,8\rangle], \\ \xi^{(4)} &= \frac{1}{5}[\langle 1,6\rangle + \langle 1,7\rangle + \langle 2,8\rangle + \langle 3,6\rangle + \langle 3,7\rangle], \\ \xi^{(5)} &= \frac{1}{2}[\langle 1,8\rangle + \langle 3,8\rangle]. \end{aligned} \tag{12}$$

Eq. (7) shows that any GTI can be separated in term of the contributions of pairs of vertices at the same distance in the graph. Each $\xi^{(k)}$ term defines the average

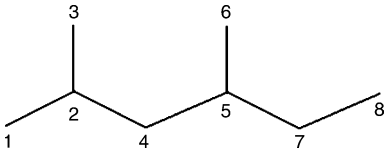

Fig. 1. Molecular-graph for the 2,4-dimethylhexane showing the labels used to work the example in the text.

contribution to GTI of all those pairs of vertices whose vertices are separated at distance k in the graph. These contributions are scaled by the number of pairs of vertices of each type and by the x and p scale parameters through the C_k coefficients.

The C_k coefficients separate the several contributions in (7), i.e., the x and p parameters define the relative importance of each term in the expansion with respect to how far the currently vertices are in the graph. The presence of the diameter, diam(G), and the numbers $N^{(k)}$ in the expansion (7) introduce information about the branching of the graph in the GTI definition. On the other hand, from Eq. (2), if $q = 1$ and $y = 0$ then the components $\mathbf{u}_i$ reduce to the topological vertex degree, i.e., the number of edges attached to the vertex. Similar analysis is valid for the $\mathbf{v}_i$ components. So, each $\mathbf{u}_i$ or $\mathbf{v}_i$ generalise the concept of the classical concept of vertex degree permitting to assign local characteristics to each vertex; and each bracket $\langle,\rangle_G$ in (7) introduces local information about the vertices in the GTI definition. Therefore, each GTI codes global and local information about the graph structure on the set of vertices.

3. Optimisation of generalised topological indices

Every GTI is obtained by settling the value of each parameter in (1). In what follows, a minimization criterion is introduced to define any GTI for practical purposes. The study will be restricted to the six-dimensional parameter subspace formed by all 6-tuples (x, y, z, p, q, r) only.

Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_g\}$ be a set of any physicochemical property data for a collection of molecules represented by the set $\hat{G} = \{G_1, G_2, \ldots, G_g\}$ of molecular-graphs. By hypothesis, it is assumed the existence of a 6-tuple $(x_0, y_0, z_0, p_0, q_0, r_0)$ that minimizes the six-dimensional scalar function Q,

$$Q_{xyz}^{pqr} = 1 - |R(\mathrm{GTI}[G] \leftrightarrow \Theta)|, \tag{13}$$

where the symbol $|\cdot|$ means absolute value and $R(A \leftrightarrow B)$ denotes the linear correlation coefficient between the data A and B. Now, to find the minimum (local or global) of the function (13), a multidimensional optimisation procedure is required. The so-called downhill simplex method of optimisation [16–19] has been selected here for the sake of simplicity, because this method requires only function evaluations, not derivatives.

A simplex is a geometrical object consisting of $n + 1$ points and all their interconnecting line segments, where n is the number of parameter to be optimised. The downhill simplex method starts from an initial simplex. Then, by successive reflections, expansions and contractions operations, the algorithm moves the simplex in the direction of the best point (minimum). Several different start simplexes are required in order to study the local or global nature of any minimum. Reference [23] contains a detailed exposition about this method and its implementation on computers.

All the previous ideas (generalized topological indexes, linear correlation function and downhill simplex method), have been coded in a computer program termed GTI-Simplex where each simplex involves seven 6-tuples of the form (x, y, z, p, q, r). Each 6-tuple defines a GTI response point and GTI-Simplex explores the GTI-space to find a minimum of function (13).

4. Application of GTI methodology

In this section, the practical use of GTI methodology is illustrated through the study of the normal boiling point of the eighteen octane isomers [24,25]. Octane isomers represent a challenging data set for QSPR models using topological descriptors [26,27]. In this case the QSPR models are not 'falsified' by the effect of the molecular weight. After the optimisation process through the GTI-Simplex methodology, Table 1 shows the linear regression coefficients and statistical parameters obtained for the optimised GTI. Classical Wiener index and Randić index have been included for comparison. It is noteworthy the very good improvement in the fitting process achieved using GTI.

The linear models for the boiling points (in Celsius degrees) on octane isomers can be written as follows:

$$\mathrm{BP}[G] = 78.06 + 0.512 \times \frac{1}{2}\begin{pmatrix} 1.0000 & 0.0000 & 0.0000 \\ 1.0000 & 1.0000 & 1.0000 \end{pmatrix}_G, \tag{14}$$

$$\mathrm{BP}[G] = 2.535 + 30.49 \times \frac{1}{2}\begin{pmatrix} 1.0000 & -0.5000 & -0.5000 \\ 0.0000 & 0.0000 & 0.0000 \end{pmatrix}_G, \tag{15}$$

$$\mathrm{BP}[G] = -309.80 + 707.91 \times \frac{1}{2}\begin{pmatrix} -1.0000 & -0.4998 & -0.4998 \\ -3.2223 & 0.7537 & 0.7537 \end{pmatrix}_G. \tag{16}$$

Or equivalently, using Eq. (7):

$$\mathrm{BP}[G] = 78.06 + 0.512 \sum_{k=1}^{\mathrm{diam}(G)} kN^{(k)}[G], \tag{17}$$

$$\mathrm{BP}[G] = 2.535 + 30.49N^{(1)}[G]\xi^{(1)}[G], \tag{18}$$

$$\mathrm{BP}[G] = -309.80 + 707.91 \sum_{k=1}^{\mathrm{diam}(G)} (-1)^{k-1} \times \left(\frac{0.3103}{k}\right)^{k-1} N^{(k)}[G]\xi^{(k)}[G]. \tag{19}$$

In the first model, the scale x parameter does not appear and all the $\xi^{(k)}$ terms are equal to the unit. These two strong restrictions reduce drastically the flexibility

Table 1
Statistical and regression results for the normal boiling point of the octane isomers[a]

	W	χ	GTI
A	78.06	2.535	−309.80
B	0.512	30.49	707.91
N	18	18	18
R	0.541	0.824	0.990
SD	5.320	3.587	0.891
RCV	–	–	0.988
SDCV	–	–	0.983

[a] A: Intercept; B: Slope; N: Sample size; R: Correlation coefficient; SD: Standard deviation; RCV: Cross-validated correlation coefficient; SDCV: Cross-validated standard deviation; W: Wiener index; χ: Randić index; GTI: Generalized topological index, whose optimum parameter values are: $x = -3.2223$, $y = 0.7537$, $z = 0.7537$, $p = -1.0000$, $q = -0.4998$, $r = -0.4998$.

of the Wiener index and, consequently, the quality of the fit. In this model all pair of vertices have the same contribution to the index without any consideration about how far or close the vertices are in the graph. And, the fact that all the $\xi^{(k)}$ terms are equal to the unit is consequence that all $\mathbf{u}_i$ and $\mathbf{v}_i$ (see Eq. (2)) are equal to the unit too. Therefore, within this model, it is impossible to distinguish any vertex in the graph with respect to its effect on the property.

In the second model, the Randić index permits to distinguish the vertices in the graph according to the degree of them. From Eq. (2), it is easy to show that each $\mathbf{u}_i$ or $\mathbf{v}_i$ term is equal to the reciprocal of the square root of the degree of vertex i. So, the above result introduces more flexibility in the graph-invariant description of the molecule. But, unfortunately, Randić index only involves contributions due to those vertices separate at unit distance, cutting the expansion (7) at first order.

Finally, the third model (the optimised GTI) permits to distinguish the vertices in the graph through the y, z, q and r parameters in agreement with the fact that the neighbour atoms affect the local environment of each atom. Moreover, the several contributions to the property are separated in terms of the distance among vertices in agreement with the idea that the interactions inside the molecule are strongly dependent on the separation of the interacting sites. Eq. (19) shows the well-known reciprocal dependence between the boiling point of octane isomers and the branching of the molecule. The scaling factor: $(0.3103/k)^{k-1}$, in Eq. (19) decreases drastically the terms associated to large distances. According to our calculations (data not shown), it is clear that the first three terms in (19) are the leading terms in the current GTI evaluations. The first and the third contributions in (19) are positive whereas the second is negative. The first contribution is practically constants for all isomers, so the current contribution value depends mainly on the second and third terms. Molecules with high branching tend to have a high number of atom pairs separated at distance two in contrast with more linear molecules. Therefore, molecules with high branching degree increase the second term in the expansion (19), which decreases the boiling point.

A strong long-range intermolecular force increases the boiling point of liquids. In general, the intermolecular forces are the result of the electrostatic potential around the molecule. This electrostatic potential is a function on the molecular charge distribution, so it is easy to understand the role that plays the molecular branching for explaining the concentration of charge. However, it is important to stress the inhomogeneous nature of the electric field around the molecule. In Section 2, the importance of the $\xi^{(k)}$ terms to introduce local information on the vertices in the graph was discussed. Our calculations about the distribution of $\xi^{(k)}$ along the isomers (data nor shown) reveals small variations in these contributions that are responsible for the good fit obtained in the optimisation process. The inhomogeneous contribution of vertices in the graph to the property is coded by the inhomogeneous values in the $\mathbf{u}$ function. As an example of the above, Fig. 2 shows how the function $\mathbf{u}$ changes from model 1 to model 3 in 2,4-dimethylhexane. It can be seen the important role that the local atomic environment plays in the description of the boiling point of octane isomers. In this way, GTI offers a method to study the effect of the local environment of each atomic site in the molecule.

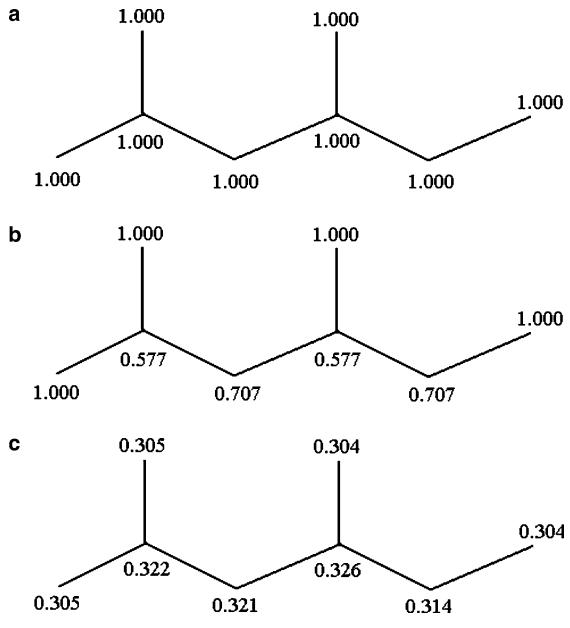

Fig. 2. Molecular-graph for the 2,4-dimethylhexane showing the values of $\mathbf{u}$ function on the vertices: (a) Wiener index; (b) Randić index; and (c) GTI index after the optimisation process.

5. Conclusions

We have introduced a new graph-theoretical approach that combines the robustness of optimisation methods with the transparency of simple topological approaches. GTI-Simplex approach permits the design of an index for describing quantitatively a physicochemical property instead of using an ad hoc descriptor that could not be optimal for such property. GTI descriptors were written in terms of geodesics encoding global and local information about the graph structure on the set of vertices, which permits to study the role of individual topology of single molecules in understanding the properties of molecular ensembles. The understanding of how the properties of single molecules result into the properties of molecular ensembles has been claimed as an important step in the comprehension of molecular complexity [28].

Acknowledgements

A.R.M. thank to FONDECYT (CHILE) the support under Grant No. 1040463. E.E. thanks 'Ramón y Cajal' program, Spain and members of the QTC group for warming hospitality during two visits in 2002 and 2003.

References

[1] M. Randić, in: P.v.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Looman, H.F. Schaefer III, P.R. Schreimer (Eds.), The Encyclopedia of Computational Chemistry, Wiley, Chichester, UK, 1998, p. 3018.
[2] J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, Amsterdam, 2001.
[3] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.
[4] A.T. Balaban (Ed.), Chemical Applications of Graph Theory, Academic Press, London, 1976.
[5] M. Randić, D. Mills, S.C. Basak, Int. J. Quantum Chem. 80 (2000) 1199.
[6] M. Randić, New J. Chem. 24 (2000) 165.
[7] M. Randić, M. Pompe, J. Chem. Inf. Comput. Sci. 41 (2001) 631.
[8] M. Randić, M. Pompe, J. Chem. Inf. Comput. Sci. 41 (2001) 575.
[9] M. Randić, S. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 614.
[10] M. Pompe, Chem. Phys. Lett. 404 (2005) 296.
[11] B. Lucic, A. Milicevic, S. Nikolic, N. Trinajstic, Indian J. Chem. 42A (2003) 1279.
[12] E. Estrada, Chem. Phys. Lett. 336 (2001) 248.
[13] E. Estrada, Y. Gutiérrez, MATCH 44 (2001) 55.
[14] E. Estrada, J. Phys. Chem. A 107 (2003) 7482.
[15] E. Estrada, J. Phys. Chem. A 108 (2004) 5468.
[16] J.A. Nelder, R. Mead, Comput. J. 7 (1965) 308.
[17] L.A. Yarbro, S.N. Deming, Anal. Chim. Acta 73 (1974) 391.
[18] C.L. Shavers, M.L. Pearsons, S.N. Deming, J. Chem. Educ. 56 (1979) 307.
[19] D.J. Leggett, J. Chem. Educ. 60 (1983) 707.
[20] F. Harary, Graph Theory, Addison-Wesley, Reading, 1969.
[21] H. Wiener, J. Am. Chem. Soc. 69 (1947) 17.
[22] M. Randić, J. Am. Chem. Soc. 97 (1975) 6609.
[23] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in Fortran 77 (Section 10.4), Cambridge University Press, New York, 1986.
[24] A.A. Gakh, E.G. Gakh, B.G. Sumpter, D.W. Noid, J. Chem. Inf. Comput. Sci. 34 (1994) 832.
[25] American Petroleum Institute Research Project 44 at the National Bureau of Standards, 1947–1991, Physical and Thermodynamical Properties of Hydrocarbons.
[26] M. Randić, N. Trinajstić, J. Mol. Struct. (Theochem) 284 (1993) 197.
[27] M. Randić, N. Trinajstić, J. Mol. Struct. (Theochem) 300 (1993) 551.
[28] G.M. Whitesides, R.F. Ismagilov, Science 284 (1999) 89.