Simplex Optimization of Generalized Topological Index (GTI−Simplex): A Unified Approach to Optimize QSPR Models

Adelio R. Matamala*[,†] **and Ernesto Estrada**[‡]

Facultad de Ciencias Químicas, Grupo QTC, Universidad de Concepción, Casilla 160-C, Concepción, Chile, and Complex Systems Research Group, X-rays Unit, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, Santiago de Compostela 15782, Spain

Received: June 23, 2005; In Final Form: August 31, 2005

GTI−simplex is a new methodology that combines the generalized topological indices and the down hill simplex optimization procedure to search for optimized quantitative structure−property relationship models (*Chem. Phys. Lett.* **2005**, *410*, 343). In this study, the fundamental role of the graph topological distance inducing a local shell structure on vertexes and a detailed derivation of the GTI-decomposition in terms of the so-called "geodesic-brackets" , i.e., functions that mix the local shell structure for different vertexes are presented. Applications of the GTI−simplex to a set of physicochemical properties covering those depending on intramolecular and/or intermolecular interactions are included. GTI−simplex has showed to be a very effective methodology for the description of different properties from a unified point of view. No ad hoc definition for topological index is required to each property as in the traditional use of topological indices or other molecular descriptors to QSAR/QSPR studies.

1. Introduction

The central role that graph−theoretical invariants play in the study of quantitative structure−property relationships (QSPR) and quantitative structure−activity relationships (QSAR) is well-known.[1] Far from the days on which graph-based methods were considered as a numerical curiosity only, topological indices have become nowadays a vigorous branch of computational chemistry.[2] However, the myriad of descriptors and the apparent disconnection among them make a claim for an underlying general theory. Previously, serious efforts have been devoted to construct new topological indices. One of these strategies has been the empirical and semiempirical modification of topological indices, such as the molecular connectivity indices, to account for more structural information.[3−8] On the other hand, an approach that allows the optimization of topological indices to improve their performance in QSAR/QSPR studies has also been developed.[9−14] A generalized approach that unifies many of the "classical" topological indices into one theoretical framework permitting the development of optimal QSPR and QSAR models, and a better understanding of their meaning has been recently introduced.[15−18]

Any QSPR/QSAR based on graph−theoretical descriptors assumes the existence of a correlation between a (molecular or molar) property/activity and the molecular structure described at a topological level. Although a formal demonstration does not exist for the above assumption, regression process had shown during decades the validity of it through building analytical models, where the linear model is the simplest representative.[19,20]

Basically, 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 way and each property requires an ad hoc graph-invariant definition.

We have developed a general methodology for obtaining optimized topological indices[21] that combines the *simplex* optimization method[22−25] and the so-called generalized topological index (GTI).[15] The main objective of this approach is to obtain optimized invariants for each property under study. The great merit of this methodology is that it permits to study different properties from a unified point of view deriving molecular descriptors, which are optimal to describe the studied properties and not by using an ad hoc series of descriptors that could not be optimized for describing such properties.

In this paper, we report the theoretical aspects about the GTI-decomposition in terms of the so-called "geodesic-brackets" and the fundamental role of the graph topological distance inducing a local shell structure on vertexes. This local structure on vertexes is the responsible for coding into any GTI the specific features of the molecular topology present in the molecules. Finally, we discuss the application of the simplex method to the optimization of GTIs, permitting us to model four physicochemical properties with a very good accuracy. The properties have been chosen in such a way that they account for intramolecular and/or intermolecular forces.

2. Generalized Topological Index (GTI)

In this section, the theoretical aspects concerning with the generalized topological indices (GTI) introduced in previous works[15−18,21] are discussed. Let $G(V,E)$ be a molecular-graph with $|V| = n$ vertexes and $|E| = m$ edges. Let $D = (d_{ij})_{n\times n}$ be the topological distance matrix of the graph $G(V,E)$. Using the kth order geodesic (shortest path) matrix of the graph G, whose entries are defined as follows

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

---

* To whom correspondence should be addressed. E-mail: amatamal@udec.cl.

[†] Universidad de Concepción.

[‡] University of Santiago de Compostela.

10.1021/jp053408x CCC: $30.25

the generalized topological index (GTI) associated with the graph $G(V,E)$ can be written as follows:

$$GTI[G] = \frac{1}{2}\begin{pmatrix} p_0 & p_1 & p_2 \\ x_0 & x_1 & x_2 \end{pmatrix}\!\Bigg\|\begin{matrix}\bar{s} \\ \bar{w}\end{matrix}\Bigg)_G = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{\mathrm{diam}(G)} C_k\langle i,j\rangle_G \Delta_{ij}^{(k)}[G] \quad (2)$$

where

$$C_k = C_k(x_0,p_0) = k^{p_0} x_0^{\,p_0(k-1)} \quad (3)$$

and each "*geodesic-bracket*" $\langle i,j\rangle_G$ is defined by

$$\langle i,j\rangle_G = \frac{1}{2}(u_i v_j + v_i u_j) \quad (4)$$

where

$$u_i(G;x_1,p_1,\bar{w}) = [w_i[G] + \delta_i[G] + \sum_{k=2}^{\mathrm{diam}(G)} k x_1^{\,k-1} N_i^{(k)}[G]]^{p_1} \quad (5)$$

$$v_i(G;x_2,p_2,\bar{s}) = [s_i[G] + \delta_i[G] + \sum_{k=2}^{\mathrm{diam}(G)} k x_2^{\,k-1} N_i^{(k)}[G]]^{p_2} \quad (6)$$

and

$$N_i^{(k)}[G] = \sum_{j=1}^{n} \Delta_{ij}^{(k)}[G] \quad (7)$$

The scalars x_0, x_1, x_2, p_0, p_1, p_2, $\bar{w} = (w_1, w_2, ..., w_n)$, and $\bar{s} = (s_1, s_2, ..., s_n)$, form a $(2n + 6)$-dimension real space of parameters. The first six parameters x_0, x_1, x_2, p_0, p_1, and p_2 are free parameters whereas the parameters $\bar{w}$ and $\bar{s}$ are predefined quantities that could permit one to introduce a priori nonequivalency of the atoms in the molecule (i.e., the presence of heteroatoms in the molecule). The quantity $\delta_i[G]$ is the classic degree of the ith vertex in the graph G, whereas the quantity $N_i^{(k)}[G]$ is the number of vertexes at distance k from the ith vertex. For simplicity, when $\bar{w} = (0, 0, ..., 0) = \bar{0}$ and $\bar{s} = (0, 0, ..., 0) = \bar{0}$, the following compact symbol

$$\begin{pmatrix} p_0 & p_1 & p_2 \\ x_0 & x_1 & x_2 \end{pmatrix}\!\Bigg\|\begin{matrix}\bar{0} \\ \bar{0}\end{matrix}\Bigg)_G = \begin{pmatrix} p_0 & p_1 & p_2 \\ x_0 & x_1 & x_2 \end{pmatrix}_G \quad (8)$$

is used in order to abbreviate the notation. Now, by the introduction of the following definition

$$\eta^{(k)}[G] = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle i,j\rangle_G \Delta_{ij}^{(k)}[G] \quad (9)$$

eq 2 can be written in the following compact form:

$$GTI[G] = \sum_{k=1}^{\mathrm{diam}(G)} C_k \eta^{(k)}[G] \quad (10)$$

Equation 10 shows that any GTI can be separated in terms of the contributions of pair of vertexes at the same topological distance in the graph. Each $\eta^{(k)}$ term defines the contribution to GTI of all those *interactions* due to the pairs of vertexes separated at distance k in the graph. These contributions are scaled by two real parameters through the C_k coefficients. For example, by simple inspection of Figure 1, the $\eta^{(k)}$ for the 2,4-dimethylhexane molecule are

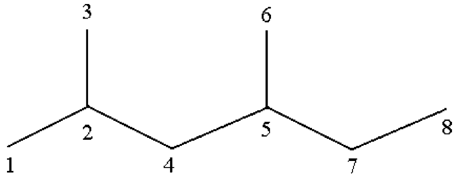

Figure 1. Molecular graph labeling for the 2,4-dimethylhexane.

$$\left.\begin{aligned}
\eta^{(1)} &= \langle 1,2\rangle + \langle 2,3\rangle + \langle 2,4\rangle + \langle 4,5\rangle + \langle 5,6\rangle + \langle 5,7\rangle + \langle 7,8\rangle,\\
\eta^{(2)} &= \langle 1,3\rangle + \langle 1,4\rangle + \langle 2,5\rangle + \langle 3,4\rangle + \langle 4,6\rangle + \langle 4,7\rangle + \langle 5,8\rangle + \langle 6,7\rangle,\\
\eta^{(3)} &= \langle 1,5\rangle + \langle 2,6\rangle + \langle 2,7\rangle + \langle 3,5\rangle + \langle 4,8\rangle + \langle 6,8\rangle,\\
\eta^{(4)} &= \langle 1,6\rangle + \langle 1,7\rangle + \langle 2,8\rangle + \langle 3,6\rangle + \langle 3,7\rangle,\\
\eta^{(5)} &= \langle 1,8\rangle + \langle 3,8\rangle.
\end{aligned}\right] \quad (11)$$

The diameter of the graph is a global descriptor and its presence in definition (10) reveals the dependence of any GTI on the "size" of the molecular-graph. On the other hand, the molecular-connectivity relationship among atoms in the molecule defines the "shape" of the molecular graph. This "shape" is coded by the so-called *geodesic-brackets* (see eqs 5 and 6). The functions u and v are the generalization of the "classic" vertex degree notion. Through these functions u and v and by settling the x and p parameters, a pair of weights is assigned to each vertex in the graph. On each vertex these weights code the topological environment around it. From the previous analysis, it is clear that the codification of the topological complexity of any molecular-graph relies on the $N_i^{(k)}[G]$ quantities, defined by eq 7. The current x and p real parameters are responsible for the connection of this topological graph-complexity with the physicochemical property. From any distance matrix, it is very easy to obtain the $N_i^{(k)}[G]$. For example, the process for the 2,4-dimethylhexane molecule (see Figure 1) reads as follows:

$$\begin{array}{c|cccccccc}
d_{ij} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline
1 & 0 & 1 & 2 & 2 & 3 & 4 & 4 & 5 \\
2 & 1 & 0 & 1 & 1 & 2 & 3 & 3 & 4 \\
3 & 2 & 1 & 0 & 2 & 3 & 4 & 4 & 5 \\
4 & 2 & 1 & 2 & 0 & 1 & 2 & 2 & 3 \\
5 & 3 & 2 & 3 & 1 & 0 & 1 & 1 & 2 \\
6 & 4 & 3 & 4 & 2 & 1 & 0 & 2 & 3 \\
7 & 4 & 3 & 4 & 2 & 1 & 2 & 0 & 1 \\
8 & 5 & 4 & 5 & 3 & 2 & 3 & 1 & 0
\end{array}
\quad \Rightarrow \quad
\begin{array}{cc|cccccccc}
 & & & & & i & & & & \\
 & N_i^{(k)} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline
 & 1 & 1 & 3 & 1 & 2 & 3 & 1 & 2 & 1 \\
 & 2 & 2 & 1 & 2 & 4 & 2 & 2 & 2 & 1 \\
k & 3 & 1 & 2 & 1 & 1 & 2 & 2 & 1 & 2 \\
 & 4 & 2 & 1 & 2 & 0 & 0 & 2 & 2 & 1 \\
 & 5 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 2
\end{array}$$

Within the GTI method, the $N_i^{(k)}[G]$ quantities are the basic units of information describing the graph-topology associated with molecules. So, the distance matrix has a central role in the present theoretical approach. For a given distance k, the set of all $N_i^{(k)}[G]$ defines a k-shell structure centered on the ith vertex. For example, the k-shell structure on the vertex 4 of the molecular-graph for the 2,4-dimethylhexane (see Figure 1) is shown in Figure 2. Every vertex in the graph has its shell distance-induced coding locally the topological global features of the graph.

3. GTI−Simplex Methodology

The methodology used here is the same reported in a previous work.[21] By the combination of the GTI approach and the simplex optimization method, which is described below, a general methodology for obtaining optimized GTI has been developed.

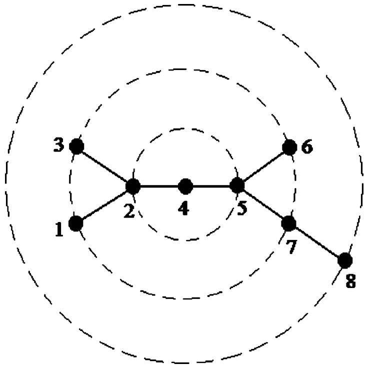

Figure 2. Distance *k*-shell structure on vertex 4 of the molecular graph for the 2,4-dimethylhexane.

Let $\Pi = \{\pi_1, \pi_2, ..., \pi_g\}$ be a set of any physical-chemical property data for a collection of molecules represented by the set $\hat{G} = \{G_1, G_2, ..., G_g\}$ of molecular-graphs. In this section, the concept of GTI will be used for obtaining a general linear regression procedure that fits the property data and the molecular graph structures. In what follows, the study will be restricted to the six-dimensional parameter subspace formed by all 6-tuples (x_0, x_1, x_2; p_0, p_1, p_2); i.e., we will consider GTIs of the form shown in (4) only.

Let Q be the following scalar function

$$Q = 1 - |R(GTI[G] \leftrightarrow \Pi)| \qquad (12)$$

where the symbol $|\cdot|$ means the absolute value and $R(A \leftrightarrow B)$ denotes the linear correlation coefficient between data A and B. By hypothesis, it is assumed the existence of a 6-tuple that minimizes the six-dimensional scalar function Q. Now, to find the minimum (local or global) of the function Q, we used the so-called downhill simplex method of optimization.[22−25] The main goodness of this method is that it 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 optimized. The downhill simplex method starts from an initial simplex. Then, by successive reflection, expansion, and contraction operations, the algorithm moves the simplex in the direction of the best point (minimum). Several different starting simplexes are required in order to study the local or global nature of any minimum. A detailed exposition about this method and its implementation on computers can be found elsewhere.[26]

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_0, x_1, x_2; p_0, p_1, p_2). Each 6-tuple defines a GTI response point and GTI−simplex explores the GTI space to find a minimum of function.[21]

Two important constrains on the simplex optimization procedure should be imposed for negative values of the parameters: x_0, x_1, x_2. First, for $x_0 < 0$ the real parameter p_0 only adopts the discrete values 0, ±1, ±2, ... The second constraint is less restrictive than the first. From eqs 5 and 6, some continuum intervals of negative values for the parameters x_1 and x_2 are compatible with noninteger values of the parameters p_1 and p_2. By a trial and error process, these intervals are easily found. Then, the simplex algorithm works within these intervals under the boundary condition $Q = 1$ for pathological cases (fractional power of a negative number).

TABLE 1: Selected Properties of Octane Isomers[a]

octane isomers[a]		CAS registry no.	P_1	P_2	P_3	P_4
1	8	111-65-9	−208.4	39.194	41.48	398.8
2	2M7	592-27-8	−215.5	39.234	39.68	390.8
3	3M7	589-81-1	−212.6	39.102	39.84	392.1
4	4M7	589-53-7	−212.1	39.119	39.68	390.9
5	3E6	619-99-8	−210.9	38.946	39.65	391.7
6	22MM6	590-73-8	−224.7	39.255	37.30	380.0
7	23MM6	584-94-1	−213.9	38.983	38.79	388.8
8	24MM6	589-43-5	−219.4	39.132	37.78	382.6
9	25MM6	592-13-2	−222.6	39.261	37.87	382.3
10	33MM6	563-16-6	−220.1	39.011	37.54	385.1
11	34MM6	583-48-2	−213.0	38.864	38.98	390.9
12	23ME5	584-94-1	−211.2	38.838	38.53	388.8
13	33ME5	563-16-6	−215.0	38.719	37.99	391.4
14	223MMM5	564-02-3	−220.1	38.927	36.93	383.0
15	224MMM5	540-84-1	−224.1	39.264	35.15	372.4
16	233MMM5	560-21-4	−216.4	38.764	37.23	387.9
17	234MMM5	565-75-3	−217.4	38.870	37.71	386.6
18	2233MMMM4	594-82-1	−225.9	38.634[c]	35.19	379.4

[a] P_1: standard heat of formation in the gas phase (kJ/mol). P_2: molar refraction (cm^3/mol). P_3: standard heat of vaporization (kJ/mol). P_4: normal boiling point (K). Data were mainly taken from ref 19. [b] M: methyl. E: ethyl. [c] From ref 27.

4. Results and Discussions

In this section, we apply the GTI−simplex methodology to a set of four properties for the eighteen octane isomers: standard heat of formation in gas phase (property P_1), molar refraction (property P_2), standard heat of vaporization (property P_3), and normal boiling point (property P_4). Table 1 shows the values used as the input data in the present work. Data were taken from Kier and Hall,[19] except one taken from CRC Handbook.[27]

The goal to chose the above properties consist of covering properties depending on the intramolecular forces (the first two) and properties depending on the intermolecular forces (the last two). On the other hand, octane isomers represent a challenging data set for QSPR models using topological descriptors.[28,29] In this case the descriptors cannot account for the effect of molecular weight on the physicochemical properties studied and the QSPR models are not "falsified" by this effect. At last but not least we have selected this set of molecules as the simplest representative of organic molecules. It is straightforward to realize that the results obtained for these data set can be extended to other series of molecules in a similar way as the topological indices, first tested for alkanes, have been extended to the whole chemistry.

After optimization process and using the notation in (4), any GTI−simplex model can be written in the following standard form:

$$P[G] = A + B \times \frac{1}{2}\begin{pmatrix} p_0 & p_1 & p_2 \\ x_0 & x_1 & x_2 \end{pmatrix} \qquad (13)$$

where A and B are linear correlation constants that connects the graph−theoretic approach with the physicochemical property under study. Table 2 shows the results obtained for the four selected properties of the octane isomers. For comparative purposes, Table 3 shows the use of the well-known Wiener index[30] and Randić index[31] to model the same four properties. The stability of the models was studied by cross validation experiments using the leave-one-out approach. Every time the GTI parameters were reoptimized for every new data in which one structure was hidden, and this hidden datum was predicted using these new parameters.

TABLE 2: Statistical and Linear Regression Results for the Four Selected Properties of the Eighteen Octane Isomers after the GTI−Simplex Optimization Process (Parameters Shown Below)[a]

	P_1	P_2	P_3	P_4
A	396.0	8.262	61.97	−36.52
B	−3.855	0.3922	−0.00581	707.7
R	−0.973	0.982	−0.995	0.990
S	1.242	0.038	0.170	0.891
CVR	0.965	0.945	0.991	0.987
CVS	1.357	0.062	0.224	1.009
x_0	0.0577	0.0899	0.8111	−3.2223
x_1	0.8110	1.0773	0.6919	0.7537
x_2	1.2271	0.6130	−0.0643	0.7537
p_0	0.7247	1.4349	0.7205	−1.0000
p_1	1.2374	0.3031	1.6555	−0.4998
p_2	−0.0379	0.7144	1.9503	−0.4998

[a] *A*: intercept. *B*: slope. *R*: correlation coefficient. *S*: standard deviation. CVR: cross-validated correlation coefficient. CVS: cross-validated standard deviation.

TABLE 3: Statistical and Linear Regression Results for the Four Selected Properties of the Eighteen Octane Isomers Using the Wiener Index (*W*) and the Randić Index (χ)[a]

		P_1	P_2	P_3	P_4
W	A	−245.3	37.42	24.06	351.1
	B	0.4094	0.0227	0.2031	0.5138
	R	0.5066	0.7385	0.8200	0.5423
	S	4.6501	0.1385	0.9462	5.3123
	CVR	0.3335	0.6610	0.7798	0.3824
	CVS	2.4299	0.1190	0.8320	2.9845
χ	A	−314.7	37.52	4.381	275.5
	B	26.84	40.85	9.272	30.54
	R	0.8502	0.3396	0.9581	0.8250
	S	2.8394	0.1932	0.4734	3.5727
	CVR	0.8195	−0.1234	0.9373	0.7480
	CVS	2.6939	0.0660	0.5655	3.6744

[a] *A*: intercept. *B*: slope. *R*: correlation coefficient. *S*: standard deviation. CVR: cross-validated correlation coefficient. CVS: cross-validated standard deviation.

From Tables 2 and 3, we observe three remarkable facts. First, GTI−simplex method is a very effective methodology for the description of different properties from a unified point of view. No ad hoc definition for topological index is required for each property as in the traditional use of topological indices or other molecular descriptors to QSAR/QSPR studies. Second, the improvements in the correlation with respect to the "classical" indices are quite remarkable. For instance, the relative percentage of improving in R for the properties P_1, P_2, P_3, and P_4 are, respectively, 92%, 33%, 21%, and 83%, compared to those correlations obtained using the Wiener index. On the other hand, the relative percentage of improving in R for the properties P_1, P_2, P_3, and P_4 are, respectively, 14%, 190%, 4%, and 20%, compared to those correlations obtained using the Randić index. Finally, the cross-validation experiment using the leave-one-out technique shows that this regression models are very stable to the inclusion−exclusion of data points, which is indicative of the lack of overfitting for these models.

In the section 2, we discussed the central role of the graph diameter and the distance-induced shell structure on vertexes for defining the GTIs. Of course, the graph diameter is a relevant descriptor for the molecular branching, i.e., any decrease in the graph diameter means to pass from less starlike to more starlike graphs. A star with N vertexes, $S_{1,N-1}$, is a graph having one central vertex with degree $N - 1$ and $N - 1$ vertexes with degree one. Examples are propane ($S_{1,2}$), 2-methylpropane ($S_{1,3}$), and 2,2-dimethylpropane ($S_{1,4}$). However, the complexity in the description of molecular branching relays on the "diversity" of graphs with the same diameter. Therefore, topological indices capable of describing very accurately this "diversity" are the candidates for building a good theory about the application of graph theory in chemistry.

The distance-induced shell structure on vertexes is given for each graph G by the quantities $N_i^{(k)}[G]$, i.e., the number of vertexes at distance k from the ith vertex in G. Every graph defines a distribution $N_i^{(k)}[G]$ on its vertexes, intrinsic information contained in the graph distance matrix. It is clear that within the present theoretical framework these distributions of non-negative integers on a sample of graphs are the responsiblity of the "fine" graph-based description of any physicochemical property.

Figure 3 shows the plots of each selected property for octane isomers vs its optimized GTI (left-hand side) and the plots of the average values of the current optimized GTI for graphs with the same graph diameter vs the diameter (right-hand side). In these plots, the dispersion on the average values has been included with 95% of confidence (*t*-Student distribution).

The analysis of all plots in Figure 3 clearly shows a common behavior, that is as follows: any increase in the graph diameter produces an increase in the value of the property. However, the plots of the dispersions of the average GTI vs diameter exhibit a relevant difference between those properties depending on the intramolecular forces from those depending on intermolecular forces.

Let us concentrate our analysis on graphs with diameters 4, 5, and 6, which are the only subsets with any possibility of "diversity" in the current sample. For all properties the "diversity" in graph is the same but the relative dispersion on the plots in Figure 3 vary from one case to another. On one hand, the optimization process for the properties P_3 and P_4 reduces drastically the dispersion for graphs from diameter 4 through 6. On the other hand, the optimization process for properties P_1 and P_2 maintains practically constant the dispersion for graphs with diameter 4 through 6, this is particularly obvious for the property P_2. What do these dispersions try to tell us about the discriminative power of any GTI for recognizing molecular branching?

Having in mind that the number of octane isomers with diameter 4 is 6, with diameter 5 is 7, and with diameter 6 is 3, the intimate relationships between the dispersion on plots and the molecular branching are obvious. In fact, property P_1 is the heat of formation, a property strongly bond-additive. Thus, for this property, the particular topology is less relevant than the number and type of bonds broken or formed from a reference compound. Property P_2 is the molar refraction, a property that represents the volume occupied by the molecules per unit mole. So, this is a property closest to the molecular "size" than to the molecular "shape" . Properties P_3 and P_4, the heat of vaporization and boiling point, respectively, are strongly dependent on the intermolecular forces, i.e., they depend on the electric field surrounding the molecule, where the lack of homogeneities in the field are extremely relevant in defining the current value on the property. Thence, the extreme increment in the dispersion on plots when the "diversity" is high is a direct consequence of the fact that during the optimization process, to obtain a good fit between the property and the index, the GTI−simplex algorithm recognizes the main role of the branching on the graph sample for these P_3 and P_4 properties. A more accurate analysis for the case of property P_1 indicates that the corresponding GTI is sensitive to molecular branching too. This is because we are

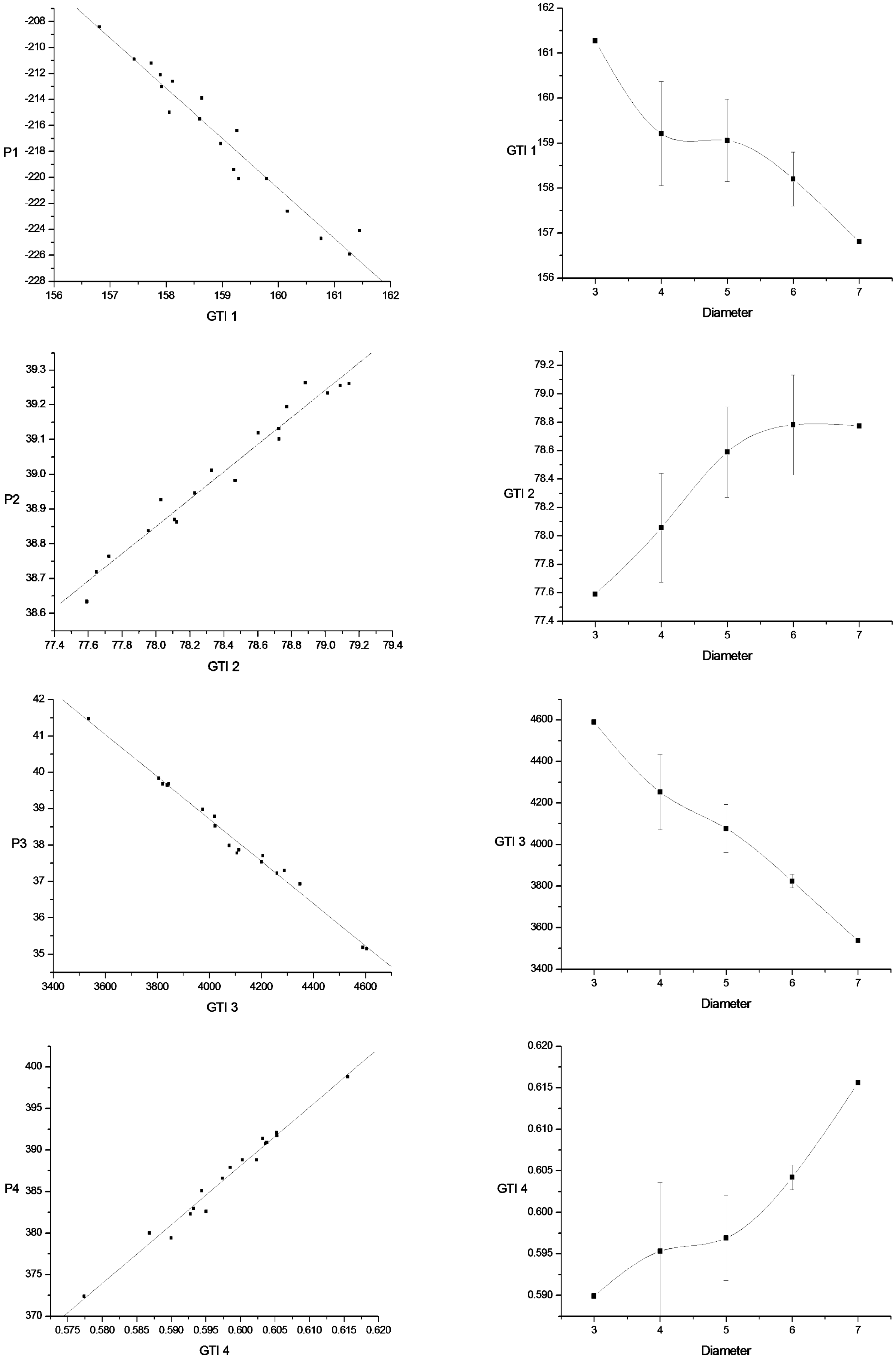

Figure 3. Plot of each property vs its optimized GTI (left-hand side plots) and plot of the average values of the current GTI for graphs with the same diameter vs the graph diameter (right-hand side plots). The 95% of confidence (*t*-Student distribution) has been included to show the dispersion about each point due to the graph branching effect.

considering heat of formation in the gas phase, and a vaporization process is implicit.

5. Conclusions

We have developed a basic methodology for the optimal development of QSPR/QSAR models using graph−theoretical molecular descriptors. This approach, named GTI−simplex, represents a unifying framework that avoids the ad hoc use of topological indices in a nonoptimal way to describe physicochemical or biological properties. Instead, GTI−simplex finds the optimal descriptor for each property by optimizing several parameters that describe the global and local topology of a molecule. The advantages of this approach are not only those related to the practical results obtained by using it, i.e., the great improvement of the quality of QSAR/QSPR models. This approach encloses several of the most important topological indices into one graph−theoretical invariant, which open new possibilities to generalize and interpret this type of molecular descriptors as a whole. As we have shown here, this approach also permits a clear structural interpretation of the QSAR/QSPR models developed using it by means of their decomposition into the geodesic brackets. These results point to the direction that a generalized approach to the topological description of molecular structure by using graph theory is a necessary step forward the definition of a *molecular complexity theory* due to "*networks are the prerequisite for describing any complex system, indicating that complexity theory must inevitably stand on the shoulders of network theory.*"[32] And network theory is not anything other than graph theory.

It is obvious that the graph−theoretical approach to chemistry is telling us nothing about the nature of the atoms and bonds forming the molecule, characteristics which are well described by quantum mechanics. GTI and other topological approaches are mainly based on a description of the organization, i.e., the topological structure, of atoms in a molecule. However, as Mark Buchanan has written: "*Some of the deepest truths of our world may turn to be truth about organization, rather than about what kinds of things make up the world and how those things behave as individuals.*"[32] This organization of atoms in a molecule, which defines concepts such as those treated in this work, e.g., diameter, geodesics, branching, etc., is accounted for by a graph−theoretical representation of a molecule. Because graph theory "*is the branch of mathematics that deals with questions concerning the various ways that a group of things can be connected together, and the theory applies no matter what these "things" might be.*"[33]

Acknowledgment. A.R.M. thanks the FONDECYT (Chile) for the support under Grant 1040463. E.E. thanks the "Ramón y Cajal" program, Spain, for partial financial support.

References and Notes

(1) Randić, M. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Looman, P. A., Schaefer, H. F., III, Schreimer, P. R., Eds.; Wiley: Chichester, U.K., 1998; p 3018.
(2) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, 2001.
(3) Pogliani, L. *Chem. Rev.* **2000**, *100*, 3827.
(4) Pogliani, L. *J. Phys. Chem.* **1994**, *98*, 1494.
(5) Pogliani, L. *J. Phys. Chem.* **1995**, *99*, 925.
(6) Pogliani, L. *J. Phys. Chem.* **1996**, *100*, 18065.
(7) Pogliani, L. *J. Phys. Chem. A* **1999**, *103*, 1598.
(8) Pogliani, L. *J. Phys. Chem. A* **2000**, *104*, 9029.
(9) Randić, M.; Mills, D.; Basak, S. C. *Int. J. Quantum Chem.* **2000**, *80*, 1199.
(10) Randić, M. *New J. Chem.* **2000**, *24*, 165.
(11) Randić, M.; Pompe, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631.
(12) Randić, M.; Pompe, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575.
(13) Randić, M.; Basak, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614.
(14) Pompe, M. *Chem. Phys. Lett.* **2005**, *404*, 296.
(15) Estrada, E. *Chem. Phys. Lett.* **2001**, *336*, 248.
(16) Estrada, E.; Gutiérrez, Y. *MATCH* **2001**, *44*, 55.
(17) Estrada, E. *J. Phys. Chem. A* **2003**, *107*, 7482.
(18) Estrada, E. *J. Phys. Chem. A* **2004**, *108*, 5468.
(19) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(20) *Chemical applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(21) Matamala, A. R.; Estrada, E. *Chem. Phys. Lett.* **2005**, *410*, 343.
(22) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308.
(23) Yarbro, L. A.; Deming, S. N. *Anal. Chem. Acta* **1974**, *73*, 391.
(24) Shavers, C. L.; Pearsons, M. L.; Deming, S. N. *J. Chem. Educ.* **1979**, *56*, 307.
(25) Leggett, D. J. *J. Chem. Educ.* **1983**, *60*, 707.
(26) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran 77*; Cambridge University Press: New York, 1986; Sect. 10.4.
(27) *CRC Handbook of Chemistry and Physics*, 84th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 2003.
(28) Randić, M.; Trinajstić, N. *J. Mol. Struct. (THEOCHEM)* **1993**, *284*, 197.
(29) Randić, M.; Trinajstić, N. *J. Mol. Struct. (THEOCHEM)* **1993**, *300*, 551.
(30) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(31) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(32) Buchanan, M. *Nexus*; W. W. Norton & Co.: New York, 2002.
(33) Barabási, A. L. *Linked*; Plume: New York, 2003.