Chemistry Explained by Topology: An Alternative Approach

Jorge Galvez[1], Vincent M. Villar[2], María Galvez-Llompart[3] and José M. Amigó[*,4]

[1]*Molecular Connectivity and Drug Design Unit, University of Valencia, Valencia, Spain*

[2]*Department of Physiology, Pharmacology and Toxicology, CEU Cardenal Herrera University, Moncada, Valencia, Spain*

[3]*Department of Pharmaceutical Technology, University of Valencia, Valencia, Spain*

[4]*Operation Research Center, Miguel Hernández University, Elche, Alicante, Spain*

Abstract: Molecular topology can be considered an application of graph theory in which the molecular structure is characterized through a set of graph-theoretical descriptors called topological indices. Molecular topology has found applications in many different fields, particularly in biology, chemistry, and pharmacology. The first topological index was introduced by H. Wiener in 1947 [1]. Although its very first application was the prediction of the boiling points of the alkanes, the Wiener index has demonstrated since then a predictive capability far beyond that. Along with the Wiener index, in this paper we focus on a few pioneering topological indices, just to illustrate the connection between physicochemical properties and molecular connectivity.

Keywords: Molecular topology, topological indices, wiener index, hosoya index, kier, hall indices.

INTRODUCTION

Molecular topology (MT), also called mathematical chemistry or chemical graph theory, is a mathematical approach to the profiling of chemical structures, which has demonstrated to be an excellent tool for a quick and precise prediction of many physicochemical and biological properties [2, 3]. One of the most interesting advantages of MT is the straightforward calculation of molecular descriptors which are later on correlated with the experimental properties, thus leading to quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) analyses.

Within this framework a molecule is assimilated to a graph —the *molecular graph*—, where each vertex represents an atom and each edge, a bond. Usually, all the hydrogen atoms are removed from the molecule before producing the corresponding molecular graph. Hereafter it will be assumed that molecular graphs refer to hydrogen-suppressed molecules even when not explicitly stated. Given a graph G with N vertices and the corresponding edges, a so-called *adjacency* or *connectivity* $N\times N$ *matrix* $\mathbf{A} = \mathbf{A}(G) = (A_{ij})_{1\le i,j\le N}$ can be defined in such a way that its entry A_{ij} takes the value either 1 or 0, depending on whether or not the vertex i is connected to the vertex j, respectively. Other fundamental $N\times N$ matrix associated to a graph G is the *distance matrix* $\mathbf{d} = \mathbf{d}(G) = (d_{ij})_{1\le i,j\le N}$, where d_{ij} is the minimal number of edges connecting the vertices i and j. The entry d_{ij} is called the *topological distance* between i and j. To complete the graph-theoretical basics needed below, let us add that the *degree* or *topological valence* of the vertex i, denoted $\deg_i$, is the number of edges going in or out of i. By means of these and similar algebraic constructs, one can introduce a number of topological indices — or graph theoretical descriptors — which characterize each graph and hence can be used in QSPR and QSAR studies [4, 5].

Graph theory is a fine instance of pure mathematics that has found a variety of applications in course of time. Created in the works of L. Euler (1707-1783) in the eighteenth century, it was developed in the nineteenth century by A. Cayley (1821-1895) and J.J. Sylvester (1814-1897) who, by the way, coined the word *graph*. During the twentieth century it became an essential tool in any area of science and technology where connectivity plays a role. For instance, the optimization of communication and transport networks, the design of electrical circuits (e.g., in computers), the synchronization of interacting oscillators with different topologies, the analysis of social networks, etc. [6]. Interestingly, it was Cayley who pointed out the correspondence between certain chemical constitutions and graph-theoretical trees.

The objective of the topological indices is to codify information on the molecule structure in a purely numerical fashion. The information contained in those indices is of topological nature, thus independent of the numbering of the vertices, Euclidean distances between atoms, and deformations which do not change the connectivity of the molecule. Furthermore, their numerical format facilitates enormously the automatic search of other molecules with similar structural properties, hence strong candidates to share the physicochemical, biological, and/or pharmacological properties sought. The relation between graphs and topological indices is not one-to-one, though. This means that given the value of one index or the values of several indices, there are in general more than one molecular graph with the same or close values; this is called the *degeneracy problem*. It is precisely this multiplicity that allows to

*Address correspondence to this author at the Operation Research Center, Miguel Hernández University, Elche, Alicante, Spain; Tel: +34 96 665 8911; Fax: +34 96 665 8715; E-mail: jm.amigo@umh.es

1386-2073/11 $58.00+.00

identify groups of molecules with hypothetical common properties *via* topological indices [7]. The fact that two very different compounds in terms of structure may share the same values of some topological indices, so as they might belong to the same, say, therapeutic group, is exploited in the applications of molecular topology.

Molecular topology is very successful in chemical graph data mining, and in particular in the discovery and design of new drugs. These topics have already been addressed by some of the present authors in, for example [2, 7]. In this paper we rather wish to stress that there is more to it, namely, the fact that molecular topology offers also an explanatory setting for physicochemical properties based on the topological features of the molecules. As a tool for screening a database in search of chemical compounds, the topological indices have got over time some interesting competitors. Perhaps to mention the most recent ones: hashed fingerprints, Maccs keys, extended connectivity fingerprints, frequent subgraphs, and bounded-size graph fragments. We refer the interested reader to [8] for details. But these are mainly numerical techniques. Beyond this instrumental aspect, molecular topology has also a connection to reality that has not been fully surveyed yet.

CONCEPTS AND DEFINITIONS

The first topological index ever was the Wiener index. The *Wiener index W* was originally defined as the number of all chemical bonds between pairs of (in general, non-hydrogen) atoms in an acyclic molecule [1]. The current definition of *W*, based on the distance matrix **d**, was proposed by Hosoya [9]:

$$W(G) = \frac{1}{2}\sum_{i,j=1}^{N} d_{ij} = \sum_{i<j} d_{ij} = \sum_{i>j} d_{ij},$$

(since d_{ii} = 0), where N is the number of vertices in the molecular graph G. In words, W is the row-wise (or equivalently, column-wise) sum of the entries of the $N \times N$ distance matrix **d**.

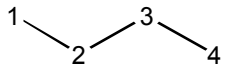

Fig. (1). Graph of butane.

$$\mathbf{d} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{bmatrix} \begin{matrix} \text{dij} \\ 6 \\ 4 \\ 4 \\ 6 \end{matrix}$$

Fig. (2). Distance matrix **d** for for the butane molecule.

As way of illustration, consider the butane molecule (Fig. **1**). Here N = 4. In order to calculate the value of the Wiener index for this molecule, we need first to build up its distance matrix, shown in Fig. (**2**). For instance, $d_{24} = d_{42} = 2$, because there are two edges between the vertices 2 and 4. Also shown in Fig. (**2**) are the row-wise sums of the entries of the distance matrix. Therefore, the Wiener index of butane is

$$W(G) = \frac{1}{2}\sum_{i,j=1}^{N} d_{ij} = \frac{1}{2}(6+4+4+6) = 10.$$

In spite of its simplicity, the Wiener index correlates very well with some physical properties like, for example, the boiling point of the alkane series [1]. Yet, its degeneracy is comparatively high, and this calls for other indices to be jointly used.

Hosoya [9] proposed also the topological index now known as the *Hosoya index*:

$$Z(G) = \sum_{k=0}^{\lfloor N/2 \rfloor} n(G,k),$$

where $\lfloor N/2 \rfloor$ denotes the integer part of $N/2$, and $n(G, k)$ is the number of ways in which k nonadjacent edges of the molecular graph G can be chosen. By definition, $n(G, 0) \equiv 1$, and $n(G, 1)$ is the number of edges in G.

Below we shall also refer to the *Kier and Hall indices* ${}^{l}\chi$ [10, 11]. In order to define ${}^{l}\chi$, the molecular graph is decomposed into all possible paths P_l of length l. A path P_l of length l of a graph G is a subgraph consisting of a sequence of l+1 vertices of G together with the l edges between consecutive vertices. Then,

$${}^{l}\chi(G) = \sum \left(\deg_{i_0} \cdot \deg_{i_1} \cdot \ldots \cdot \deg_{i_l} \right)^{-1/2},$$

where the sum is over all paths P_l in the molecular graph G, and i_0, i_1,…, i_l are the vertices of G belonging to P_l. Note that

$${}^{0}\chi(G) = \sum_{i=1}^{N} \deg_i^{-1/2}.$$

The index ${}^{l}\chi$ is also called the *l-th order connectivity index*. For l = 1, the index is referred to as the *Randić index* [12], which was the first connectivity index based on the notion of topological degree instead of distance. Kier and Hall generalized the Randić index to higher values of l.

Lastly, by replacing $\deg_i$ in the definition of ${}^{l}\chi$ by

$$\deg_i^v = Z_i^v - H_i,$$

where Z_i^v is the number of valence electrons of atom i, and H_i is the number of hydrogen atoms suppressed at atom i, we obtain the so-called *valence connectivity indices*, denoted by ${}^{l}\chi^v$. Numerical examples of calculations of these and other topological indices can be found in [2, 7].

SOME RELATIONSHIPS BETWEEN TOPOLOGY AND PHYSICAL CHEMISTRY

Topological indices yield excellent results in the prediction of physicochemical properties although, in general, no definitive statements can be made as to their physical meaning. In the following subsections we are going to substantiate this claim with a few examples involving the Wiener, Hosoya, and the Kier and Hall indices.

a) The Wiener Index

We already mentioned the excellent correlation of the Wiener index with the boiling points of the alkanes. Furthermore, Gutman and Zenkevich [13] showed the relationship between *W* and the internal (mainly vibrational) energy of organic molecules. The Wiener index has been shown to correlate with the van der Waals surface area of the molecule [14], and this explains why it correlates with numerous physicochemical properties of organic compounds.

In a different context, the minimum atom to atom separation in terms of minimum total distance of the graph, as expressed by the Wiener number, is able to predict the most compact (the most stable) structure of crystals [15]. In this case it is the topology that determines the equilibrium forms of crystals and atomic clusters, and not the other way around.

Directly related to the Wiener index is the *hyper-Wiener index WW* [16], which was the result of efforts to improve correlation with physicochemical properties. Its definition is as follows:

$$WW(G) = \frac{1}{2}\sum_{i<j}\left(d(G)^2 + d(G)\right)_{ij} .$$

Klein and collaborators [17] made this new index applicable to graphs that contain cycles. The application of this index in spiro-graphs containing 3 to 6 membered rings was proposed by Diudea and coworkers [18]. These formulas are derived on the basis of Hosoya's and Klein's formulas for evaluating *W* and *WW* respectively, in cycle-containing graphs.

b) The Hosoya Index

Wiener's and Hosoya's indices are correlated. According to Wagner [19], for rooted ordered trees with *n* vertices, the correlation coefficient of their Wiener indices W_n and *Hosoya indices* Z_n is given asymptotically by

$$r(W_n,Z_n) \sim (0.40351) \cdot (0.99637)^n.$$

But the Hosoya index has also interest on its own. Thus Fischermann *et al*. [20] showed that the chemical trees minimizing the energy agree with those minimizing the Hosoya index for a small number of vertices. Furthermore, the Hosoya index for large conjugated hydrocarbons such as polyenes and polycyclic aromatic hydrocarbons, have a good correlation with their π-electronic properties as the bond order and π-electronic energy [21].

c) The Kier and Hall Indices

The first-order connectivity index $^1\chi$ may be related to both electronic and vibrational energies of alkanes and conjugated hydrocarbons, E_π, according to the formula:

$$E_\pi = N_\pi\alpha + 4\beta^1\chi$$

Here α and β stand for the Coulomb and resonance integrals, respectively, as defined in the original Hückel Molecular Orbital theory and N_π is the number of π electrons of the conjugated hydrocarbon [22]. This equation allows for a very good fitting to, for example, the Hückel results for resonance energies. Moreover, the equation can be refined into a more accurate version if we include the influence of the overlap integral.

On the other hand, it is easily seen that vibrational energy may also be expressed as a function of topological valences. Thus, the values ranging from 1634 to 1675 cm^{-1} of the vibrational frequencies for all substituted ethylene derivatives follow surprisingly well the following relation:

$$\nu\ (cm^{-1}) = 1780.6 - 147.2[(1/\deg_1) + (1/\deg_2)]^{1/2},$$

where the subscripts 1 and 2 refer to the ethylene carbons.

Extending some of the earlier work done by Estrada [23] and based on the concept of accessibility introduced by Kier and Hall [24], a prediction of the molecular volume and surface of alkanes was done in our research group [25]. The formalism is purely geometrical and is based on (i) the concentric spheres representing the covalent and Van der Waals volumes, (ii) the relation between molecular volume and surface area in alkanes, and (iii) connectivity indices. The demonstration is based on the loss of accessible volume per atom as two atoms bind each other. The representation as concentric spheres of the covalent and van der Waals volumes as well as the loss of accessible volume given by the intersection of the Van der Waals volumes when the two covalent spheres are in contact, provide a measure of the loss of accessibility. According to these results, the molecular volume and surface area of alkanes can be expressed as a function of $^0\chi$ and $^1\chi$, i.e., the zero-order and first-order connectivity indices, respectively. The linear regression equations yield coefficients r = 0.987 and r = 0.991 for volume and surface area with $^0\chi$ and $^1\chi$, respectively. These results fit well with the known fact that with an increasing degree of branching, i.e., number of tertiary or quaternary carbons, the molecular volume increases while the molecular surface area decreases. The key role played by the molecular volume and surface area is well-known for experimental properties ranging from molecular polarizibility to boiling temperatures and, in general, for all intermolecular forces and interactions. The molecular accessibility, *A*, is an important concept which aids in deriving the given model for the connectivity indices. Alternative definitions of this concept can be given in terms of connectivity indices, as, for example, in the following equation:

$$A = 4(^1\chi^v) - {}^0\chi^v.$$

Despite the simplicity of this definition, it is possible to predict the increment of the standard free energy (and hence the spontaneity of the process) for positional isomerisation between pairs of hydrocarbons and even more complex molecules. For instance, the equilibria between n-butane and isobutane, between 2-methylpentane and 3-methylpentane, between propylamine and isopropylamine, and between o-methylaniline and m-methylaniline may all be predicted. All these features reinforce the idea that at least some important physical and geometrical molecular parameters can be expressed as functions of topological indices and thus bypass the need for cumbersome calculations.

To conclude, we borrow from Ref. [26] a practical example of how Molecular Topology is applied to predict physicochemical properties. In this particular case, the study dealt with the parameter BOD_5 —Biochemical Oxigen

Demand—, defined by the U.S. Environmental Protection Agency (EPA) as the amount of dissolved oxygen consumed in five days by biological processes breaking down organic matter, as a way to measure the biodegradability of compounds in water. The methodology explained, for instance, in [2, 7], lead to the following topological model with three variables:

$$\text{Log } (1/BOD_5) = -6.753 - 0.638\ Dz + 0.038\ CENT + 4.384\ piPC05$$

$$N = 26,\ r^2 = 0.8929,\ Q^2 = 0.7796,\ SEE = 0.127,\ F(3,22) = 61.1,\ p < 0.00001,$$

where N is the sample size, r is the regression coefficient, Q is the predictive correlation coefficient, *SEE* is the standard error of the estimate, F is the Fisher-Snedecor parameter, and p is the p-value. In this case, the indices employed in the model were descriptors evaluating both topological (CENT and piPC05) and electronic (Dz) characteristics of the molecule (see [26] for more details). Fig. (**3**) shows the plot experimental versus calculated values. As it can be appreciated, there is a good concordance between both.

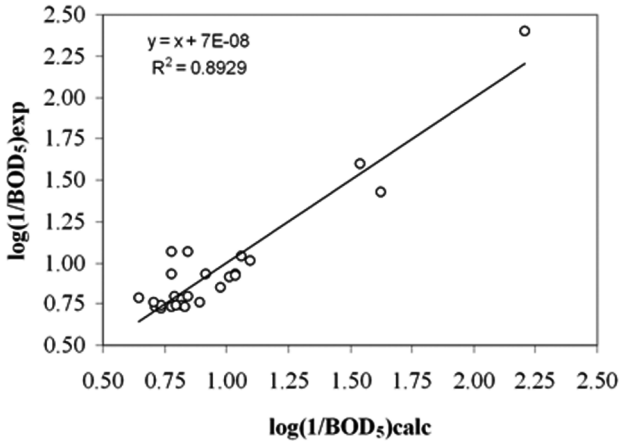

Fig. (3). Graphic representation of log $(1/BOD_5)_{exp}$ versus $\log(1/BOD_5)_{calc}$ for the topological model selected.

The predictability quality and robustness of this model was verified by means of different validation criteria. Among them are the internal validation or cross-validation with leave-one-out, the external validation employing external data, and the randomization test [26].

CONCLUSION AND OUTLOOK

The application of topology to understanding different aspects of chemistry is becoming increasingly important both in theory and applications. In our excursion through the basic aspects of molecular topology, we have revisited a few topological indices and pointed out their relevance for chemistry. The application of graph-theoretical methods to chemistry is a growing field, one masterful exponent being Randić's review of chemical phenomena otherwise explained by quantum methods [27]. This clearly shows the importance of this intriguing approach to chemistry, whose possibilities have begun to be explored just in the last two decades.

Related to this, let us mention that a major research area in molecular topology is drug discovery and design *via* topological descriptors. In this way, new lead compounds can be obtained in such diverse therapeutical fields as analgesics, antibacterials, hypolipidemics, hypoglycemics, bronchodilators, antivirals, antineoplastics, antihystaminics, antimalarials, and so on [2]. The mean level of accuracy in the different therapeutical scopes is over 80%. Other examples of new lead compounds discovered *via* topology include acaricides and organic semiconductors. Altogether these results demonstrate the efficiency of topological molecular design, in such a way that it could be considered an excellent tool to describe molecular structures. It makes possible the selection of new lead drugs without a need to know their mechanism of action, which usually is implicit in the topological fingerprint.

In summary, molecular topology is not only an alternative but also an independent approach to describe molecules as compared to conventional methods based on quantum or classical mechanics [22]. In this regard, its potential is still far from being fully exploited.

ACKNOWLEDGEMENTS

We are thankful to our referees for their constructive criticism. We also thank Dr. Gerardo Antón (Chemistry Department, CEU Cardenal Herrera University of Valencia) for helpful discussions. J.M.A. was supported by the Spanish Ministry of Science and Innovation, grant MTM2009-11820.

ABBREVIATIONS

MT = Molecular topology

QSAR = Quantitative structure-activity relationship

QSPR = Quantitative structure-property relationship

REFERENCES

[1] Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc*., **1947**, *69*, 17-20.

[2] Garcia-Domenech, R.; Galvez, J.; De Julian-Ortiz, J.V.; Pogliani, L. Some new trends in chemical graph theory. *Chem. Rev*., **2008**, *108*, 1127-1169.

[3] Dudek, A; Arodz, T.; Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Combin. Chem. & High Throughput Screen*., **2006**, *3*, 213-228.

[4] Devillers, J.; Balaban, A.T. In *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers: Singapore, **1999**.

[5] Karelson, M. *Molecular Descriptors in QSAR/QSPR*, J Wiley & Sons: New York, **2000**.

[6] Newman, M.; Barabási, A.L.; Watts, D.J. *The Structure and Dynamics of Networks*, Princeton University Press: Princeton, **2006**.

[7] Amigó, J.M.; Galvez, J.; Villar, V.M. A review on molecular topology: applying graph theory to drug discovery and design. *Naturwissenschaften*, **2009**, *96*, 749-761.

[8] Wale, N.; Ning, X.; Karypis, G. Trends in Chemical Graph Data Mining. In: Aggarwal, C.C., and Wang, H. (ed.), *Managing and Mining Graph Data* (Chapter 19), Springer Verlag, 2010.

[9] Hosoya, H. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Japan*, **1971**, *44*, 2332-2339.

[10] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drugs Research*, Academic: London, **1976**.

[11] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure–Activity Analysis*, Research Studies: Letchworth, **1986**.
[12] Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc*., **1975**, *97*, 6609-6615.
[13] Gutman, I.; Zenkevich, I.G. Wiener Index and Vibrational Energy. *Z. Naturforsch*., **2002**, *57*, 824-828.
[14] Gutman, I.; Körtvélyesi, T. Wiener indices and molecular surfaces. *Z. Naturforsch*., **1995**, *50*, 669-671.
[15] Bonchev, D. My Life–Long Journey in Mathematical Chemistry. *Internet Electronic J. Molec. Design,* **2005**, *4*, 434-490.
[16] Randić, M. Novel molecular descriptor for structure–property studies. *Chem. Phys. Lett*., **1993**, *211*, 478-483.
[17] Klein, D.J.; Lukovits, I.; Gutman, I. On the definition of the hyper-Wiener index for cycle-containing structures. *J. Chem. Inf. Comput. Sci*., **1995**, *35*, 50-52.
[18] Diudea, M.V.; Katona, G.; Minailiuc, O.M.; Parv, B. Molecular topology 24. Wiener and hyper-Wiener indices in spiro-graphs. *Rus. Chem. Bul*., **1995**, *44*, 1606-1611.
[19] Wagner, S.G. Correlation of graph-theoretical indices. *SIAM J. Discr. Math*., **2007**, *21*, 33-46.
[20] Fischermann, M.; Gutman, I.; Hoffmann, A.; Rautenbach, D.; Vidović, D.; Volkmann, L. Extremal chemical trees. *Z. Naturforsch*., **2002**, *57*, 49-52.
[21] Hosoya, H.; Hosoi, K.; Gutman, I. A topological index for the total π-electronic energy. Proof of a generalised Hückel rule for an arbitrary network. *Theoretica Chimica Acta*, **1975**, *38*, 37-47.
[22] Galvez, J. On a topological interpretation of electronic and vibrational molecular energies. *J. Mol. Struc*., **1998**, *429*, 255-264.
[23] Estrada, E. Characterization of 3D molecular structure. *Chem. Phys. Lett*., **2000**, *319*, 713-718.
[24] Kier, L.B.; Hall, L.H. Intermolecular accessibility: The meaning of molecular connectivity. *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 792-795.
[25] Galvez, J. Prediction of molecular volume and surface of alkanes by molecular topology. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 1231-39.
[26] Gálvez, J.; Parreño, M.; Pla, J.; Sanchez, J.; Gálvez-Llompart, M.; Navarro, S.; García-Domenech, R. Application of molecular topology to the prediction of water quality indices of alkylphenol. *Int. J. Cheminform. Chem. Eng*., **2011**, *1*(1), 1-11.
[27] Randić, M. Aromaticity of polycyclic conjugated hydrocarbons. *Chem. Rev*., **2003**, *103*, 3449-3606.

Received: November 15, 2010 **Revised: January 26, 2011** **Accepted: February 18, 2011**