Naturwissenschaften (2009) 96:749–761
DOI 10.1007/s00114-009-0536-7

REVIEW

A review on molecular topology: applying graph theory to drug discovery and design

José María Amigó · Jorge Gálvez · Vincent M. Villar

Received: 16 October 2008 / Revised: 13 March 2009 / Accepted: 1 April 2009 / Published online: 10 June 2009

Abstract Molecular topology is an application of graph theory and statistics in fields like chemistry, biology, and pharmacology, in which the molecular structure matters. Its scope is the topological characterization of molecules by means of numerical invariants, called topological indices, which are the main ingredients of the molecular topological models. These are statistical models that are instrumental in the discovery of new applications of naturally occurring molecules, as well as in the design of synthetic molecules with specific chemical, biological, or pharmacological properties. In this review, we focus on pharmacology, which is a novel field of application of molecular topology. Besides summarizing some recent developments, we also seek to bring closer this interesting biomedical application of mathematics to an interdisciplinary readership.

Keywords Topological indices · Molecular topological models · Discovery of new drugs

J. M. Amigó (✉)
Operation Research Center, Miguel Hernández University,
Elche, Alicante, Spain
e-mail: jm.amigo@umh.es

J. Gálvez
Head of Molecular Connectivity and Drug Design Unit,
University of Valencia,
Valencia, Spain
e-mail: jorge.galvez@uv.es

V. M. Villar
Department of Physiology, Pharmacology and Toxicology,
CEU Cardenal Herrera University,
Moncada, Valencia, Spain
e-mail: vmvillar@uch.ceu.es

Introduction

Mathematics has a quite impressive record of biomedical applications. To mention just a few: growth and propagation of tumors, computational neuroscience, design of implantable devices and drug delivery mechanisms, genetics, computerized tomography, expert systems, clinical analyses, and epidemiology. The objective of this review, which is based on (Amigó et al. 2007) and (García-Domenech et al. 2008), is to introduce the reader to an application of graph theory and statistics which is being used for the discovery of new drugs and molecule design. This application, which goes by the name of *molecular topology* (or *connectivity*), is still young, its origin dating from the 1970s, when L.B. Kier and L.H. Hall (1976), and other researchers started using "indices" based on graph theory to study some physicochemical properties of organic compounds, like formation heat and boiling temperature. They found that those properties can be expressed as linear combinations of a few such indices. The application of molecular topology to the pharmacological research was only a matter of time, the pioneering work being done in the mid-1980s (Arviza 1985) and the first papers appearing at the beginning of the 1990s (Gálvez et al. 1991). Whatever the field, the interest on molecular topology is clear: Predicting with confidence some specific activity of a molecule saves time and money. Although the applications of molecular topology are manifold, we will focus on the pharmacological ones because of their novelty value and social impact.

Basically, molecular topology builds on the somewhat surprising correlations existing between a given physical, chemical, or biological property of a substance and the corresponding molecular characterization provided by some numerical descriptors generically called topological indices.

So to speak, these indices encapsulate structural information at the molecular level which is pertinent to the property in question. Other subsequently proposed indices, called topological charge indices, incorporate also physicochemical information in the form of the number of valence electrons. To illustrate the *modus operandi* of molecular topology, suppose that a new drug with a specific activity is sought. Once an "optimal" suite of topological indices has been selected with the aid of known active molecules, a classification function is produced [usually via linear discriminant analysis (LDA) or neuronal networks] to distinguish between active and inactive molecules. This classification function is then used to filter potentially active candidates from a chemical data base. If the data base contained naturally occurring molecules, and the activity of the selected molecule was unknown before, the result is the discovery of a new drug. If the data base consisted of synthetic (say, computer-generated) molecules, we are dealing with the inverse task: design of new drugs. Last but not the least, the predicted activity of the candidates is put to test in vivo or in vitro.

Besides molecular topology, there are other techniques for molecular design (not least, quantum-mechanical methods), but they are not so straightforward nor are they always applicable. In particular, in the case of design ex novo, i.e., design of an entirely new drug, these techniques, unlike molecular topology, require information on the biological receptor.

It is worth highlighting that graph theory is a fine instance of pure mathematics that has found a variety of applications in the course of time. Created in the works of L. Euler (1707–1783) in the eighteenth century, and developed by A. Cayley (1821–1895) and J.J. Sylvester (1814–1897) in the nineteenth century, graph theory became in the twentieth century an essential tool in any area of science and technology where connectivity plays a role. Think, for instance, of 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. (Newman et al. 2006). Interestingly enough, it was Cayley who pointed out the correspondence between certain chemical constitutions and graph-theoretical trees.

This paper is organized as follows. In the next section, we introduce a few basic concepts of graph theory that are necessary to understand the subsequent exposition, along with a selection of some important topological indices, both for illustration purposes and further references. The two following sections are devoted to present the statistical tools needed for the application of the graph-theoretical information contained in the topological indices: linear regression equations (also called predictive equations), and discriminant (or classification) equations; both kinds of equations comprise what is called a model. Our exposition continues with different applications of molecular topology to the selection and design of molecules in pharmacology, together with model instances. Lastly, we will present one real example of discovery of new drugs via molecular topology. Of course, we could have addressed many other topological indices and applications, say, to physical chemistry. But the purpose of the present review is rather to convey a general picture of both theory and praxis of molecular topology, in particular of its high degree of "connectivity" with other areas of science.

The basis of molecular topology: topological indices

Molecular topology deals with the application of graph theory (Bollobás 1998) to the description of molecular structures. To fix the basic concepts and the notation, let us recall that a graph G is a set of points, called *vertices*, along with a set of links, called *edges*, joining some pairs of vertices. The set of vertices will be denoted by $V=V(\mathrm{G})$, and the set of edges by $E=E(\mathrm{G})$. Formally, a graph is an ordered pair of sets, $G=(V, E)$, where E is a subset of unordered pairs of V. We consider only finite graphs, that is, graphs with a finite number of vertices and edges. In this case, $|G|$ denotes the *order* or number of vertices of G, which can be thought to be numbered in some convenient way. We say that two vertices i, j are *adjacent* if they are joined by an edge e (that is, $e=\{i, j\} \in E$), in which case, we write $e = e_{ij} = e_{ji}$; alternatively, we say that $i, j \in V$ are the *endvertices* of $e_{ij} \in E$. Furthermore, the number of adjacent vertices to a given vertex $i \in V$ will be called the *degree* of i and denoted by $\deg_i$.

A *path* P is a graph of the form

$$V(P) = \{i_0, i_1, \ldots, i_l\},\ E(P) = \{e_{i_0 i_1}, e_{i_1 i_2}, \ldots, e_{i_{l-1} i_l}\}. \quad (1)$$

The vertices i_0, i_l are the *endvertices of P*, and l is the *length of P*. Sometimes, we want to consider P as going from i_0 to i_l, and then call i_0 the *initial* and i_l the *terminal vertex* of P. A graph is *connected* if for every pair $\{i, j\}$ of distinct vertices, there is a path from i to j. Usually, paths appear as subgraphs of a given graph G, i.e., $P \subset G$. There are also other notions related to that of a path in a graph. In particular, a *walk W of length l* in a graph is an alternating sequence of vertices and edges of the form: $i_0, e_{i_0 i_1}, i_1, e_{i_1 i_2}, \ldots, e_{i_{l-1} i_l}, i_l$. If a walk W is such that $l \geq 3$, $i_0 = i_l$, and the vertices i_k, $0 \leq k \leq l-1$, are distinct from each other, then W is said to be a *cycle* or a *circuit* of length l. A path of length l will be denoted by P_l, and a cycle of length l (an *l-cycle*) by C_l. In particular, we call C_3 a *triangle*, C_4 a

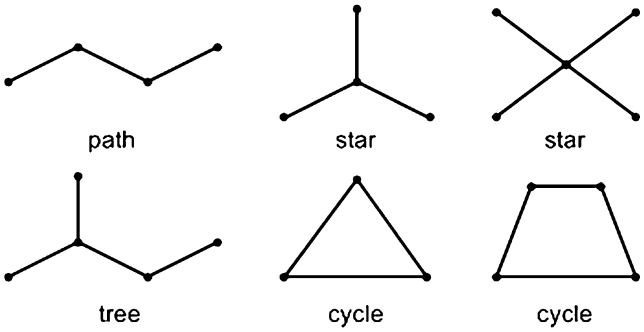

Fig. 1 Different kinds of subgraphs

quadrilateral, and so on. Sometimes, an *l*-cycle is also called an *l-gon*. A graph without any cycles is called *acyclic*.

Beside paths and cycles, we will also meet below subgraphs of the type tree and star. A *tree* of length *l* is a connected, acyclic graph with *l* edges. A *star* consists of $|G|-1$ vertices of degree 1 linked to a "central" vertex of degree $|G|-1$. The number of edges, namely, $|G|-1$, is called the length of the star. Figure 1 illustrates the different kinds of graphs we have presented.

When graph theory is applied to molecules, the vertices correspond to atoms and the edges to chemical bonds—usually covalent bonds, since molecular topology has found its major field of application in the organic chemistry. The resulting graph, which depicts how atoms are bonded with each other and which path or paths connect one atom with another one in the molecule, is called a *molecular graph*. Thus, in a molecular graph, an edge between two nodes signalizes the existence of a chemical bond between the atoms represented by those nodes, no matter what the valence of the chemical bond is. Sometimes it is needed to take into account the valence of the chemical bonds in order to deal with molecules with the same number of atoms and topology but different kind of chemical bonds—simple, twofold, triple, or quadruple covalent bonds. The result is called a molecular *pseudograph* or *multigraph*. We will not consider pseudographs in this introductory paper, so there will be no multiple edges between connected nodes.

Thus, suppose that we want to structurally characterize an organic compound. In the usual procedure, one starts removing all the hydrogen atoms from the molecule. The remaining atoms build the vertices of the molecular (hydrogen-suppressed) graph. Henceforth, all molecular graphs are meant to be hydrogen-suppressed, although not explicitly stated (For simplicity, we will not consider here more general procedures in which the hydrogen atoms are retained.). The structural information contained in the molecular graph can be now codified by means of different mathematical objects, like matrices, numerical indices, polynomials, spectra, groups, and operators.

Next, we will give a flavor of some of the simplest graph-theoretical tools used in molecular topology.

Matrices associated to a molecular graph

The *adjacency matrix* is perhaps the simplest graph-theoretical tool of molecular topology, since it gives only information about which vertices/atoms are joined/bonded in the graph/molecule. If the given hydrogen-suppressed molecule has N atoms, then the molecular graph G has order $|G|=N$, and its adjacency matrix $\mathbf{A}=\mathbf{A}(G)$ is an $N\times N$ symmetric matrix with entries

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if the atoms } i, j \text{ are bonded,} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Since $A(G)$ is real and symmetric, its distinct eigenvalues, $\lambda_{\min} < \ldots < \lambda_{\max}$, are also real. It is easy to prove that $\lambda_{\min} < 0 < \lambda_{\max}$ (Bollobás 1998). The eigenvalues of the adjacency matrix are related to several properties of G. In particular, $\delta(G) \le \lambda_{\max}(G) \le \Delta(G)$, where $\delta(G)$ and $\Delta(G)$ are the minimal and maximal degree of the vertices of G, respectively. For this reason, $\lambda_{\max}(G)$ is usually taken as a measure of the number of vertices of G with degree 3 and higher, a property that we will refer to as *branching* or *ramification* of G.

Figure 2 shows the molecular graph of the molecule 2-methyl, 3-aminopropane, along with its adjacency matrix. This simple molecule is going to be our "workhorse" in the sequel.

The sum of all entries on the ith row of $\mathbf{A}$, $\sum_{j=1}^{N} \mathbf{A}_{ij}$, as well as the sum of all entries on its ith column, $\sum_{j=1}^{N} \mathbf{A}_{ji}$, result in the number of edges going into (or out of) the vertex i, that is, the degree (or *topological valence*) of vertex i, $\deg_i$. If $\mathbf{DEG} = \mathbf{DEG}(G) = \left(\mathbf{DEG}_{ij}\right)_{1\le i,j\le N}$ is the diagonal matrix with entries $\mathbf{DEG}_{ij}=\deg_i\delta_{ij}$ (where δ_{ij} is Kronecker's delta, i.e., $\delta_{ij}=0$ if $i\neq j$, and $\delta_{ii}=1$), then the (combinatorial) *Laplacian matrix* of graph G, $\mathbf{L} = \mathbf{L}(G) = \left(\mathbf{L}_{ij}\right)_{1\le i,j,N}$, is defined as

$$\mathbf{L}(G) = \mathbf{DEG}(G) - \mathbf{A}(G). \quad (3)$$

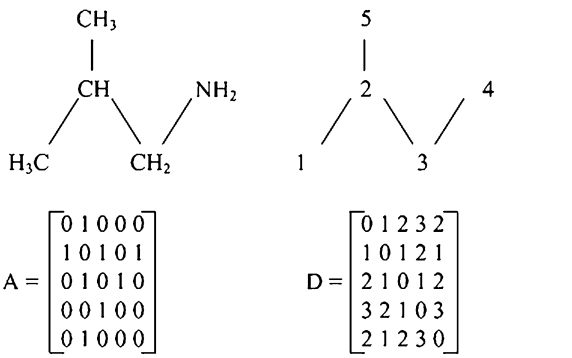

Fig. 2 The graph of the 2-methyl, 3-aminopropane molecule, and its adjacency and distance matrices

The eigenvalues of $\mathbf{L}(G)$ are also related to properties of G (Bollobás 1998).

The *distance matrix*, $\mathbf{D}=\mathbf{D}(G)$, is an $N\times N$ symmetric matrix whose entries are the *topological distances*,

$$\mathbf{D}_{ij}=\begin{cases}\text{minimal length of the paths joining node i with node } j, & \text{if } i\neq j,\\ 0, & \text{if } i=j\end{cases} \tag{4}$$

Therefore, $\mathbf{D}$ provides a qualitative picture of the proximity relation between pairs of vertices in the molecular graph. The sum of the topological distances between a given vertex i and all other vertices of the graph, is called the *distance sum* of vertex i:

$$\mathbf{DS}_i=\sum_{j=1}^{N}\mathbf{D}_{ij}=\sum_{j=1}^{N}\mathbf{D}_{ji} \tag{5}$$

Observe that the distance matrix can be obtained from the adjacency matrix. Figure 2 shows also the distance matrix for the 2-methyl, 3-aminopropane molecule.

The *path layer matrix* (also known as the atomic path code) of a graph G is the matrix $\tau=\tau(G)=(\tau_{i,j})$, where $\tau_{i,j}$ is the number of simple paths $P\subset G$ with initial vertex i and length j (Randić 1979, 1990). Other matrices used in molecular topology include the matrix $\chi=\chi(G)=\left(\chi_{ij}\right)_{1\leq i,j\leq N}$ (Randić 1992), where

$$\chi_{ij}=\begin{cases}\left(\deg_i\cdot\deg_j\right)^{-1/2} & \text{if } e_{ij}\in E(G),\\ 0 & \text{otherwise,}\end{cases} \tag{6}$$

the *matrix of inverse distances* (Ivanciuc et al. 1993), the *detour matrix* (Trinajstić et al. 1997), the *resistance matrix* (Klein and Randić 1993), etc.

Topological indices

The objective of the *topological indices* is to codify information on the molecule structure in a purely numerical fashion. Moreover, 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 that value or those values; this is called the *degeneration problem*. It is precisely this degeneration that allows to identify groups of molecules with hypothetical common properties via topological indices.

Out of the manifold of topological indices discussed in the literature (see, for instance, (Balaban et al. 1984; Trinajstić 1992; Ivanciuc 1998)), we present next a selection of those that have proved to be specially useful. In particular, the indices of Hosoya, Kier and Hall, and Randić are among the first ever proposed. Moreover, we restrict our selection to topological indices based on the adjacency and distance matrix.

Discrete invariants

The following, so-called discrete indices will be used below.

1. Already appeared, N is the number of non-hydrogen atoms, i.e., the number of vertices in the molecular graph.
2. V_k is the number of vertices of degree k, i.e., the number of atoms having k bonds in the hydrogen-suppressed molecular graph.
3. PR_k, $k\geq 0$, is the number of pairs of ramifications (i.e., pairs of vertices with degree ≥ 3) at topological distance k.
4. L is the *diameter* of the graph G, which is defined as $L=\max_{i,j\in V(G)}\mathrm{D}_{ij}$.
5. E is a so-called *form factor*. Its definition is $E=S/L^2$, where S is the molecular surface parameter. S is calculated as the sum of contributions from simple subgraphs whose S values can be found, e.g., in Table 5 of (Gálvez et al. 1994b).

Connectivity indices

Randić introduced in (Randić 1975) a connectivity number χ_R, now called *Randić index*, to characterize the branching of a molecular graph G. If f: $E(G)\rightarrow\mathbf{R}$ ($\mathbf{R}$ is the set of real numbers) is the map given by

$$f\left(e_{ij}\right)=\left(\deg_i\cdot\deg_j\right)^{-1/2}, \tag{7}$$

then the index χ_R is defined as

$$\chi_R(G)=\sum_{e_{ij}\in E(G)}f\left(e_{ij}\right). \tag{8}$$

As a result, the greater the branching of the molecule, the smaller the value of χ_{R}.

Later, Randić introduced a second connectivity number, called *identification number ID* (Randić 1984). Given P_l, a path of length l in G, define the map f^*: $E(P_l)\rightarrow\mathbf{R}$ as

$$f^*(P_l)=\prod_{e_{ij}\in E(P_l)}f\left(e_{ij}\right), \tag{9}$$

Then,

$$ID(G) = N + \sum_{P_l} f^*(P_l), \tag{10}$$

where the summation goes over all distinct paths in G.

The Randić index was generalized by Kier and Hall (1976, 1986). In order to define the *Kier and Hall indices* ${}^l\chi$, the molecular graph is decomposed into all possible paths P_l of length l. Then,

$${}^l\chi(G) = \sum \left(\deg_{i_0} \cdot \deg_{i_1} \cdot \ldots \cdot \deg_{i_l}\right)^{-1/2}, \tag{11}$$

where the sum is over all paths $P_l \subset G$, and $\{i_0, i_1, \cdots, i_l\} = V(P_l)$. Note that

$${}^0\chi(G) = \sum_{i=1}^{N} \deg_i^{-1/2}, \; {}^1\chi(G) = \chi_R(G). \tag{12}$$

Example 1 Figure 3 depicts the topological valences $\deg_i$ of the 2-methyl, 3-aminopropane molecular graph and its decomposition into length-2 paths. This is all the information needed to calculate ${}^2\chi$ for this molecule:

$$\begin{aligned} {}^2\chi &= \sum \left(\deg_{i_0} \cdot \deg_{i_1} \cdot \deg_{i_2}\right)^{-1/2} \\ &= \frac{1}{\sqrt{1\cdot 3\cdot 1}} + \frac{1}{\sqrt{1\cdot 3\cdot 3}} + \frac{1}{\sqrt{1\cdot 3\cdot 3}} + \frac{1}{\sqrt{3\cdot 3\cdot 1}} + \frac{1}{\sqrt{1\cdot 3\cdot 1}} = 2.488. \end{aligned}$$

In turn, the Kier and Hall indices (Eq. 11) can be further generalized in two steps:

- By taking in Eq. 11 connected subgraphs other than paths. The following notational remark is in order here: The subgraphs of type star, tree, and cycle (see Fig. 1) are traditionally called *cluster*, *path-cluster*, and *chain*, respectively, in the context of the generalized Kier and Hall indices. Thus, the indices

  $${}^l\chi_c, \; {}^l\chi_{pc}, \; {}^l\chi_{ch}, \tag{13}$$

  stand for those Kier and Hall indices obtained using subgraphs with l edges of type star, tree, and cycle, respectively. In agreement with this notation, sometimes ${}^l\chi$ is written ${}^l\chi_p$, where the subindex p stands for "path".

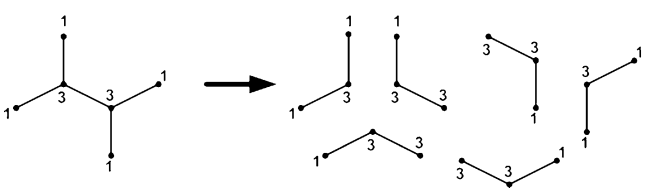

Fig. 3 Decomposition of the 2-methyl, 3-aminopropane molecular graph into paths of length 2

- By replacing $\deg_i$ in Eq. 13 by

  $$\deg_i^v = Z_i^v - H_i, \tag{14}$$

  where Z_i^v is the number of valence electrons (i.e., the electrons in the outermost electron shell) of atom i, and H_i is the number of hydrogen atoms suppressed at atom i. The resulting indices, called *valence Kier and Hall indices*, are written

  $${}^l\chi_p^v \; {}^l\chi_c^v, \; {}^l\chi_{pc}^v, \; {}^l\chi_{ch}^v. \tag{15}$$

Closely related to these generalized Kier and Hall indices are the *difference of connectivity indices*,

$${}^lD_t = {}^l\chi_t - {}^l\chi_t^v, \tag{16}$$

where $l \geq 0$ and t=p, c, pc, ch, etc., and the *quotient of connectivity indices*,

$${}^lC_t = \frac{{}^l\chi_t}{{}^l\chi_t^v}. \tag{17}$$

The *Wiener index* W was originally defined as the number of all chemical bonds between pairs of atoms in an acyclic molecule (Wiener 1947). The current definition of W, based on the distance matrix $\mathbf{D}$, was proposed by Hosoya (1971):

$$W(G) = \frac{1}{2}\sum_{i,j=1}^{N} \mathbf{D}_{ij} = \sum_{i<j} \mathbf{D}_{ij} = \sum_{i>j} \mathbf{D}_{ij}, \tag{18}$$

(since $\mathbf{D}_{ii}$=0). In spite of its simplicity, the Wiener index correlates very good with some physical properties like, for example, the boiling point of the alkane series (methane, ethane, propane,...). Yet, its degeneration is comparatively high, and this calls for other indices to be jointly used.

Hosoya (1971) proposed also the topological index now known as the *Hosoya index*:

$$Z(G) = \sum_{k=0}^{\lfloor N/2 \rfloor} n(G,k), \tag{19}$$

where $n(G, k)$ is the number of ways in which k non-adjacent edges of G can be chosen. By definition, $n(G, 0) \equiv 1$ and $n(G, 1) = |E(G)|$.

The *Balaban index* (Balaban 1982) of the molecular graph G is calculated by the formula

$$J(G) = \frac{|E(G)|}{\mu + 1} \sum_{e_{ij} \in E(G)} \left(\mathbf{DS}_i \cdot \mathbf{DS}_j\right), \tag{20}$$

where $\mathbf{DS}_i$ and $\mathbf{DS}_j$ denote the distance sums (5) of the endvertices of the edge $e_{ij} \in E(G)$, and μ denotes the number of cycles of G. It has been proven analytically and computationally that $J(G)$ has a low degeneracy.

It was also Balaban who suggested to replace the function $f^*(P_l)$ appearing in the definition of Randić's identification number, Eqs. (7–10), by the function

$$g^*(P_l) = \prod_{e_{ij} \in E(P_l)} \left(\mathbf{DS}_i \cdot \mathbf{DS}_j\right)^{-1/2}, \tag{21}$$

in order to define the *selective identification number SID*:

$$SID(G) = N + \sum_{P_l} g^*(P_l), \tag{22}$$

where the sum goes over all distinct paths in G. As advertised by its name, it has been found that the index *SID* is highly selective (Balaban 1987).

To wrap up this short list, we include also the *molecular topological index MTI* (Ivanciuc and Balaban 1999), in whose definition both matrices **A** and **D** enter explicitly. First of all, calculate the vector $\mathbf{E}=\mathbf{E}(G)=(\mathbf{E}_1,\cdots,\mathbf{E}_N)$ of structural descriptors for the vertices,

$$\mathbf{E}(G) = \mathbf{Deg}(G)(\mathbf{A} + \mathbf{D}), \tag{23}$$

where the *degree vector* of G, $\mathbf{Deg}(\mathrm{G})=(\deg_1,\cdots,\deg_N)$, multiplies as a row vector (i.e., as a $N\times N$ matrix) the $N\times N$ matrix $\mathbf{A}+\mathbf{D}$. Then

$$MTI(G) = \sum_{i=1}^{N} \mathbf{E}_i. \tag{24}$$

Table 1 shows some of the indices discussed so far for the 2-methyl, 3-aminopropane molecule.

Topological charge indices

The topological charge indexes G_l and J_l evaluate the charge transfers between pairs of atoms, and hence, the global charge transfers in the molecule (Gálvez et al. 1994a). Since many physical, chemical, and biological properties are related to the charge distribution, the introduction of topological indexes to characterize this property is convenient.

Define the $N\times N$ *matrix*

$$\mathbf{M} = \mathbf{AD}^*, \tag{25}$$

Table 1 Some topological indices for the 2-methyl, 3-aminopropane molecule

${}^0\chi$	${}^1\chi$	${}^2\chi$	${}^3\chi_c$	${}^4\chi_{pc}$	W	Z	J
5.155	2.643	2.488	0.667	1	18	10	1,785

where **A** is the adjacency matrix and $\mathbf{D}^* = \left(\mathbf{D}^*_{ij}\right)_{1\le i,j\le N}$ is the *inverse square distance matrix*,

$$\mathbf{D}^*_{ij}(G) = \begin{cases} \mathbf{D}_{ij}^{-2} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{26}$$

The matrix **M** gives rise to the charge term matrix **CT** as follows:

$$\mathbf{CT}_{ij}(G) = \begin{cases} \mathbf{M}_{ij} - \mathbf{M}_{ji} & \text{if } i \neq j, \\ \deg_i & \text{if } i = j. \end{cases} \tag{27}$$

Therefore, for $i=j$, the charge terms $\mathbf{CT}_{ij}$ represent the topological valence of the vertex i. For $i\neq j$ the charge terms $\mathbf{CT}_{ij}$ are a measure of the net charge transferred from atom j to atom i. Hence, if $\mathbf{CT}_{ij}<0$, this means that atom i will transfer a net charge to atom j.

Next, for each path of length l, we define the so-called *topological charge index* as

$$G_l = \sum_{i=1}^{N-1} \sum_{j=1+1}^{N} \left|\mathbf{CT}_{ij}\right| \delta\left(l, \mathbf{D}_{ij}\right), \tag{28}$$

where $\delta(l, \mathbf{D}_{ij})$ is Kronecker's delta. These new descriptors evaluate the total charge transfer between atoms placed at topological distance l. Thus, for a linear molecule, there are $N-1$ indexes G_l: $G_1,\cdots,G_{N-1}$.

Lastly, we introduce the *average topological charge index*

$$J_l = \frac{1}{N-1} G_l, \tag{29}$$

and the *total topological charge index*

$$J = \sum_{l=1}^{L} J_l, \tag{30}$$

The *valence topological charge indices* G_k^v and J_k^v are defined in a similar way to G_k and J_k, respectively, but using $\mathbf{A}^v$, the "electronegativity-modified adjacency matrix," instead of **A**. The entries of **A** and $\mathbf{A}^v$ are identical except for the main diagonal, where **A** has zeros and $\mathbf{A}^v$ the corresponding Pauling electronegativity values[1], weighted by 2 for each atom which is neither carbon nor hydrogen.

QSAR/QSPR models

The quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) are

[1] Electronegativity explains the fact that the covalent bond between different atoms (A–B) is stronger than would be expected by taking the average of the strengths of the A–A and B–B bonds. Pauling proposed in 1932 an empirical formula for its calculation (Pauling 1960).

approaches in pharmacology and physical chemistry based on the assumption that the activity or, more generally, the physicochemical properties of a compound depend on their molecular structure (Tropsha 2006a, b). In molecular topology (and this is what differentiates molecular topology from other methods), the information on the molecular structure takes the form of topological indices; see (Devillers and Balaban 2000) for an excellent collection of research papers on this topic. Thus, according to the QSAR/QSPR approach, given a property y (say, boiling temperature or analgesic activity), there is a set of topological indices $x_1,\cdots,x_q$ such that y is a mathematical function of them, $y=f(x_1,\cdots,x_q)$. Unfortunately, the function f and the relevant topological indices for y cannot be derived from first principles; hence, it has to be determined phenomenologically.

The simplest and hence best choice when looking for an unknown functional relationship is a linear one both for the parameters and the input variables with an offset (or bias):

$$y = c_0 + c_1 x_1 + \ldots + c_q x_q, (q \geq 1). \tag{31}$$

After all, any function over a small domain can be approximated by an affine function (or, geometrically, the graph of a differentiable function can be locally approximated by an hyperplane). This means that the effective QSAR/QSPR models will be valid only within certain ranges of the variables, called *validity domain* (Tropsha and Golbraikh 2007). If $\mathbf{y}=(y_1,\cdots y_n)^{\mathrm{T}} \in \mathbf{R}^n$ (the upper index T denotes, as usual, transposition) denotes the column vector made up by n experimental values of y as measured at n distinct compounds, $x_0=(1,1,\cdots,1) \in \mathbf{R}^n$, and $\mathbf{x}_1,\cdots,\mathbf{x}_q \in \mathbf{R}^n$ are the corresponding data of the topological indices (i.e., the ith component of the row vector $\mathbf{x}_j$ is the ith experimental value of the topological index x_j), then we can write a linear system of equations for the coefficients c_i:

$$\mathbf{y} = \mathbf{X}\mathbf{c}, \tag{32}$$

where $\mathbf{X}$ is the $n\times(q+1)$ matrix whose jth column is the vector $\mathbf{x}_j^{\mathrm{T}}$, $0\leq j\leq q$, and $\mathbf{c}=(c_0,c_1,\cdots,c_q)^{\mathrm{T}}\in\mathbf{R}^{q+1}$. The system 32 is the starting point for the fitting of the linear model and for the discriminant analysis. The optimal topological indices are normally chosen either by trying all combinations of the best suited indices or, more efficiently, by means of a greedy algorithm, i.e., a forward selection technique which at each step introduces the next best index (more on this, in the next section).

Of course, one can also use a nonlinear *ansatz* in the input variables, like $y = c_0 + c_1 x_1^{\alpha_1} + \ldots + c_q x_q^{\alpha_q}$, and then find the "nonlinear parameters" $\alpha_1,\cdots,\alpha_q$ with optimization procedures. For brevity, we will consider only linear expressions.

The quality of the regression Eq. 31 is usually assessed by a variety of statistical parameters, which include the correlation parameter of the regression r, the standard deviation of the estimate s, and the Fisher ratio F. Sometimes, a few outliers worsen the statistical quality of a predictive equation. In such cases, we recommend to compare experimental with predicted property plots.

When fitting whether linear or nonlinear models to data, it is good methodology to divide the data in two sets: (1) a *training* (or *calibration*) *set*, used to derive the fit-model equation (as we did above with n hypothetical observations) and to fix its range of validity, and (2) an *evaluation* (or *hold-out*) *set*, used to compare the predictions of the model with the experimental observations and hence evaluate its reliability. This should be also done to avoid overfitting, i.e., the use of models or methods that include more terms or procedures than are necessary (Hawkins 2004, Bishop 2006). An overfitted model can be more flexible than it needs to be (this usually happens with neural networks) or include irrelevant components that make it more complicated than needed. There are a number of reasons that make overfitting undesirable, like wasting resources or being detrimental to portability. In drug discovery—the case we are interested in—a wrong decision to use a certain molecular feature in a QSAR model when this feature is actually irrelevant, might entail the lost of valuable leads.

Linear regressions (but also neural networks) can produce satisfactory predictions for some data but grossly fail for others because the original data were overfitted. When data are short in number so as all of them are needed to set up the model, one can use the method called "leave-one-out," "cross-validation," or "jackknifing" to check for overfitting. In this method, one piece of the training data (or two pieces for the second-order jackknife, etc.) is removed, the training is performed on the remaining samples, and then the model so obtained is used to predict the sliced-out data. This procedure is repeated until each data point has been removed at some stage and predicted. The quality of the leave-one-out method is measured by the so-called prediction coefficient (Carbó-Dorca et al. 2000). It is also good policy to keep the number of indices (q) well below the number of observational data of the property y in order to avoid chance correlations that may occur whenever there are more variables than points to be fitted. A decreasing Fisher ratio F, with all other statistics increasing in quality, is a clear sign that the new, additional index is useless.

Example 2 Yaffe et al. (2001) describe the use of a classification procedure to predict aqueous solubility of organic compounds. In a calibration set of size 437, the resubstitution errors have a standard deviation of 0.0045. An independent hold-out data set of size 78 gives prediction errors with a standard deviation of 0.16, which is larger by a factor of 35. The vast difference between these two estimates of error is an indication that the

modeling method has close to one parameter per observation. This is a good example raising the suspicion that the model overfits.

Example 3 In a study on terpenoids, based on 20 experimental values taken from (Wang et al. 2008), terpenoids #8–#20 were used as a training set, while terpenoids #1–#7 were reserved for evaluation. The fit model equation was

$$\mathrm{Log(CCR)} = -0.96I_1 - 0.63I_2 + 2.219,$$

where CCR is shorthand for "corrected repellent ratio at 1.5 h" and I_1, I_2 are two topological indices. Figure 4 shows the excellent fit of the calculated Log(CCR) values with respect to the experimental ones (the correlation coeficient is 0.9487). But the same plot for the evaluation set results in an extremely poor prediction (see Fig. 5). This reveals again an overfitting problem.

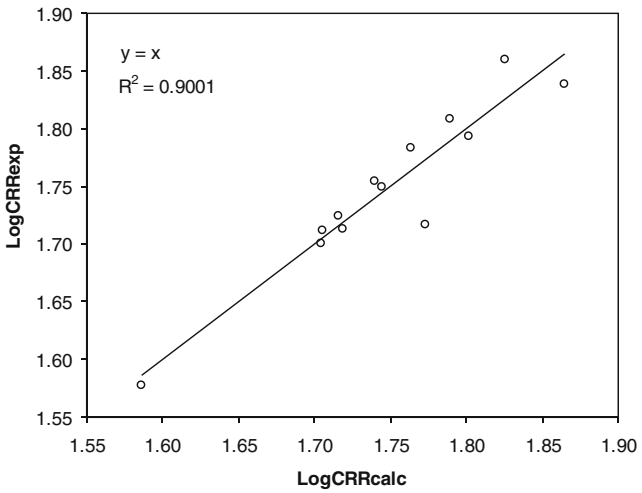

Fig. 4 Log(CRR) experimental vs Log(CRR) calculated for terpenoids #8–#20 (Wang et al. 2008). *R* here is the correlation coefficient (0.9487)

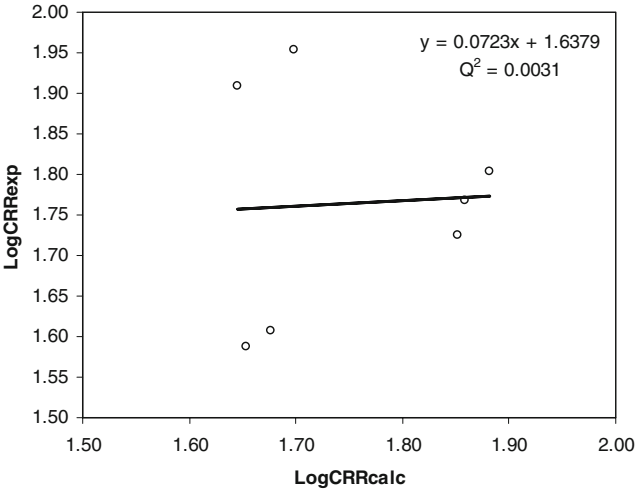

Fig. 5 Log(CRR) experimental vs Log(CRR) calculated for terpenoids #1–#7 (Wang et al. 2008). *Q* denotes the correlation coefficient of the test set (0.0557)

Molecular topological models

Molecular topological models are used to find new active compounds. For this, two different kinds of equations are generally needed, both containing topological indices: (1) *linear regression equations* (LEs) to predict quantitative properties and (2) *discriminant equations* (DEs) to recognize to which category (usually called "good" and "bad" in the two-category case) the compound belongs to. The LEs are also referred to as *predictive equations* for obvious reasons. A *model* consists of several such equations of either type, together with the corresponding validity domains, and thresholds to discriminate between categories. In sum, a model filters potentially good or active new compounds; only those satisfying the LEs and complying with the thresholds are selected for further scrutiny. The more equations the model consists of, the more selective it will be in principle, although a trade-off between modelization effort and number of potentially "good" compounds has to be made on a case by case basis.

Linear regression analysis

The objective of linear regression analysis is to find a linear combination of variables $x_1,\cdots,x_q$, see Eq. 31, which correlates with the physicochemical, biological, or pharmacological property y of interest. The Furnival–Wilson algorithm (Furnival and Wilson 1974) is used to obtain subsets of variables and equations with the least Mallows' parameter (Mallows 1973),

$$C_p = \frac{\mathrm{RSS}}{s^2} + 2p - n, \tag{33}$$

where RSS is the residual sum of squares based on the selected independent variables, s^2 is the residual mean square based on the regression using all independent variables, p is one plus the number of independent variables in the selected subset, and n is the number of data. The Furnival–Wilson algorithm combines two methods of computing the RSS for all possible regressions, into a simple leap-and-bound technique for finding the best subsets without examining all possible ones. The result is a reduction by several orders of magnitude in the number of operations required to find the best subsets.

Linear discriminant analysis

The objective of LDA is to find a linear combination of variables (topological indices in our case) that allows to discriminate between two or more categories or classes of objects. In practice, two classes of compounds are considered in the analysis: The good one, comprising a set of compounds with proven pharmacological activity, and the bad one, built by a set of compounds known to be inactive. The selection of the descriptors is based on the Fisher parameter, and the classification criterion is the shortest Mahalanobis distance (i.e., the distance of the observation from the mean of the good and bad classes used in the regression). Variables used in the linear classification functions are chosen stepwise. At each step, either the variable that contributes the most to the separation of the groups joins the discriminant function, or the variable that contributes the least is removed from the discriminant function. The quality of the discriminant function is measured by Wilks' λ, which is the bigger, the greater the overlap of the good and bad classes ($\lambda=0$ when there is no overlap). The descriptors in the discriminant function are selected or deleted so as to minimize Wilks' parameter.

There are several methods to asses the discriminant ability of a selected function. The simplest one uses an external validation set. The number of cases classified in each class and the percentage of correct classification are shown then in the so-called classification matrix. Needless to say, the higher the percentage of correct classification, the better the discriminant function.

Pharmacological-activity distribution diagrams

Beside predictive and discriminant equations, pharmacological-activity distribution diagrams (PDDs) is an auxiliary tool that is very useful in practice. These are histogram-like plots (see Fig. 6) in which the compounds are grouped into intervals of the predicted value of the property $y=f(x_1,\cdots,x_q)$. Thus, over each interval I of y-values, the number of compounds exhibiting those values is represented by a bar.

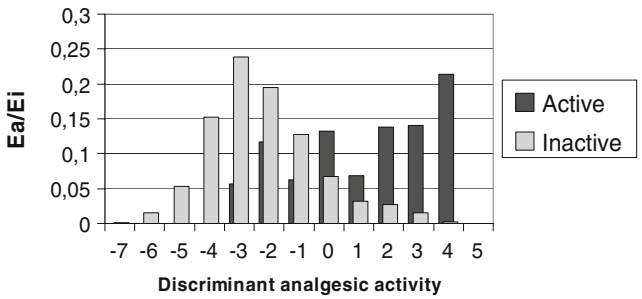

Fig 6 An example of a pharmacological–activity distribution diagram (PDD)

For each interval I, the *expectancy of activity* is defined as $E_a=a/(i+1)$, where a is the ratio of the number of active compounds in I to the total number of active compounds, and i is the corresponding ratio for the inactive compounds. The *expectancy of inactivity* is analogously defined as $E_a=i/(a+1)$. Given the step-functions E_a and E_i corresponding to the PDD of a given property y, it is in general a simple matter to decide whether the function $f(x_1,\cdots,x_q)$ is useful in molecular design, depending on whether the overlap of E_a and E_i is small or not. This also allows to determine those y-intervals where the probability of finding new active compounds is greatest compared to the choice of a false active.

Selection and design of molecules

Molecular topological models can be applied with a variety of purposes, depending on the data base used. If the data base contains naturally occurring molecules, we can find compounds with so far unknown properties, i.e., we can discover new drugs. If the data base consists of synthesized molecules, the discovery of new drugs amounts to the inverse task: design of new drugs. In the following subsections, we will present some variations on this basic theme. Needless to say, the molecular models will select from the data bases those active compounds used to build the model and, eventually, other molecules with similar structures. If this were its only performance, molecular topology would be nothing more than a complicated method of recognizing structural similarity. The magic of molecular topology is precisely its capability of catching structural similarities that the eye cannot catch.

Data base search

A mathematical model consisting of one or more equations with the corresponding thresholds is used to filter a structural database, and the so selected compounds are then checked for the sought activity in the database bibliography. Compounds found in the bibliography to be active validate the model, while those not listed are proposed as new potentially active compounds, pending from the experimental and/or clinical verdict.

Example 4 Anticonvulsant activity (Bruno-Blanch et al. 2003). The model obtained from LDA was

$$\begin{aligned} \mathrm{DF}_{AC} = &-28.88 - 1.94^4\chi^v_{pc} - 0.21G^v_1 + 4.64G_5 + 20.11J^v_3 \\ &- 45.87J_4 - 3.42^0D + 40.65^0C_p - 10.47^3C_p \\ &+ 2.79^4D_p + 1.32\mathrm{PR}_0 \end{aligned}$$

with F=10, λ=0.54 and n=128. The selection criterion was: a compound is active as anticonvulsant if

$$DF_{AC} > 0 \tag{34}$$

The 10,330 compounds comprising the Merck index were screened using this model to predict the anticonvulsant activity of compounds not considered in its derivation. As a result, 108 different compounds were selected as potential anticonvulsant drugs. It was found in the literature that 41 of them had been already reported to have anticonvulsant activity, which shows the predictibility power of the model. The rest were potential leads from molecules known to show a different pharmacological profile.

Molecular selection

A mathematical model consisting of one or more equations with the corresponding thresholds is used to filter a structural database, and the selected compounds are tested in vitro for the sought activity. Compounds selected as active but not showing activity in vitro, i.e., false positives, as well as the activities found for true positives are used to refine the model.

Example 5 Bronchodilator activity (Rios-Santamarina et al. 2002). The model consists this time of two discriminant functions: the first one, DF_1, was built using more than 300 bronchodilator drugs as well as additional structurally heterogeneous drugs showing some extent of bronchodillator activity. A second discriminant function DF_2 was also introduced to improve the discriminant efficiency of DF_1. DF_2 was obtained with a set of 70 drugs selected from every family of bronchodilators. The discriminant functions were the following:

$$DF_1 = 3.07^1\chi_p^v - 3.58G_1 + 15.32J_2 + 55.50J_4 - 1.68PR_1 + 0.879PR_2 - 11.71 \tag{35}$$

with F=287, λ=0.271 and n=339, and

$$DF_2 = 17.40^3D_p - 12.27^4D_p - 6.61 \tag{36}$$

with F=129, λ=0.315 and n=70. A compound is classified as a potentially active bronchodilator if

$$-1 < DF_1 < 10 \text{ or } 0 < DF_2 < 17. \tag{37}$$

Application of DF_1 and DF_2 to databases resulted in the selection of 20 hypothetically active molecules. The experimental tests showed the existence of compounds, such as fisetin and hesperetin, with in vitro relaxation rates on guinea pig trachea of above 80%.

Virtual combinatorial syntheses and computational screening

A mathematical model consisting of one or more equations with the corresponding thresholds is applied to a virtual combinatorial library of molecular structures resulting from a given synthesizing scheme, and the structures selected are then actually synthesized and tested. In this way, new drugs are designed.

Example 6 Anti-herpes activity (De Julián-Ortiz et al. 1999). The predictive equations were the following:

$$IC_{50} = -17.36^4\chi + 41.39^4\chi_{pc}^v + 21.71 \tag{38}$$

with F=38, r=0.914, s=0.6, n=18 and C_p=6.0,

$$\log_e(ID_{50}) = -1.42^0\chi + 4.81^0\chi_p^v - 11.41^3\chi_p^v - 1.32^3\chi_c^v + 4.17^4\chi_{pc} - 8.42 \tag{39}$$

with F=24, r=0.929, s=3.3, n=25 and C_p=5.04,

$$\log_e(UDU) = -4.67^1\chi_p^v + 8.70^2\chi - 3.64^3\chi + 3.15^3\chi_p^v - 8.05^3\chi_c - 9.23 \tag{40}$$

with F=41, r=0.957, s=12.4, n=25 and C_p=3.03. Here, IC_{50} means *inhibitory concentration 50* (i.e., the concentration inhibiting 50% of the virus growth), ID_{50} means *inhibitory dose 50* (i.e., dose leading to an inhibition of 50%), and UDU means *unchanged drug in urine*. The discriminant function was

$$DF = -1.17^0\chi_p^v + 2.11^3\chi + 2.79 \tag{41}$$

with F=23.4, λ=0.28, n=81. The selection criterion was: active if

$$-10 < IC_{50} < 20, -5 < \log_e(ID_{50}) < 3, \quad -4 < \log_e(UDU) < 4 \tag{42}$$

(the validity ranges of the predictive equations), and

$$-1 < DF < 5. \tag{43}$$

This model made possible the discovery of several synthetic active and highly active anti-herpes compounds, like the 1,2,3 triazole-4,5 dicarboxylic acid and the 2-(2,3,4-trifluorophenylcarbamoyl)-1-cyclopentene-1-carboxylic acid.

General applications

As already said in the introduction, molecular topology grew out of the study of the physicochemical properties of organic compounds. As expected, the application of molecular topology in biology and pharmacology is not so clear-cut as in chemistry due to the great variability of the test animals and patients. Moreover, many biological and pharmacological properties are less specific than physical and chemical properties in the sense that the former may depend only weakly on the molecular structure and strongly on other factors such as molecular size or the presence of functional chemical groups. Yet, molecular topology has been used successfully in different areas to predict parameters and properties. Let us mention next some representative achievements.

Prediction of physicochemical parameters Indeed, properties such as viscosity (García-Domenech et al. 1999), surface tension, and thermal conductivity (García-Domenech et al. 2003), refractive indices, and glass transition temperatures (García-Domenech and de Julián-Ortiz 2002), can be expressed as linear expressions of topological indices (including some that we have not presented above).

Prediction of pharmacological properties Pharmacological properties, such as antihistaminic (Duart et al. 2006), antifungal (García-Domenech et al. 2002), and carcinogenicity (García-Domenech et al. 2001) are predicted.

Mathematical models for the selection and design of new active compounds We have already presented some of them (anticonvulsant, bronchodilator, anti-herpes) in the previous section. Also, antibacterials (De Gregorio-Alapont et al. 2000) and antimalarials (Gálvez et al. 2005; Mahmoudi et al. 2006) belong to this group.

New biological activities discovered through virtual screening and molecular design As way of illustration of this last application of molecular topology, we will explain with some detail the discovery of new non-narcotic analgesics, as it happened in the praxis (Gálvez et al. 1994b). The model consisted of the predictive equation

$$\log_e \mathrm{IC}_{50} = 0.32G_1^v + 6.34J_1 - 0.68V_4 + 1.4E - 2.25 \quad (44)$$

with F=18, r=0.908, s=0.49, n=20, and the classification function

$$\mathrm{DF} = -1.32^0\chi + 4.67^1\chi + 1.96^1\chi_p^v - 6.56^2\chi_p^v - 4.25^3\chi - 4.11^3\chi_c + 2.68^3\chi_p^v + 13.31^3\chi_c^v + 1.28^4\chi + 11.75^4\chi_c + 1.22^4\chi_{pc} - 0.04$$

with F=9.3, λ=0.198, n=82. The selection criterion was: active if

$$0 < \log_e \mathrm{IC}_{50} < 3.5 \quad (45)$$

(validity range of the predictive equation) and

$$\mathrm{DF} < 0.686. \quad (46)$$

This model selected 17 molecules from a chemical database. Some of them were well-known for their analgesic activity, like the acetylsalicylic acid (marketed as Aspirin) and the pirazolones, but others were novel as analgesics, thus potential leads for a whole new line of analgesics.

The next step was to perform pharmacological tests (1) to confirm the analgesic activity of the novel compounds; (2) to determine their *efficient dose 50* (ED50), i.e., the optimal dose for the 50% of the animals tested; and (3) to determine their *lethal dose 50* (LD50), i.e., the dose that results lethal for the 50% of the animals which were administered the drug (Miller and Tainter 1944). These tests were performed with a statistically significant sample of rats weighing 20 to 30 g. The tests of analgesia followed the Witkin protocol (Witkin et al. 1961). Out of the 17 molecules tested, ten exhibited a clear analgesic activity. The most interesting result was that 2-(1-propenyl) phenol, one of the novel molecules, has an analgesia percentage almost twice the analgesia percentage of the acetylsalicylic

Table 2 Some pharmacological parameters of known and novel analgesics (see text)

Compound	Analgesia (%)	ED50 (mg/kg)	LD50 (mg/kg)	TI
Acetylsalicylic acid	49±1	100±8	500±20	5
2-(1-Propenyl) phenol	85±1	34±5	720±10	21
2,4-Dimethylacetophenone	80±1	45±5	700±10	16
p-Methyl-propiophenone	56±1	100±3	590±20	6
Sulfadiazine	43±1	112±10	2,000	18

acid. Furthermore, 2′,4′-dimethylacetophenone, the other of the novel molecules tested, obtained marks similar to the previous one. Both molecules were patented.

But this is not the end of the story. The so-called *therapeutic index* (TI), which is the number most commonly used to measure innocuity, is defined as the ratio

$$\mathrm{TI} = \frac{\mathrm{LD50}}{\mathrm{ED50}}. \tag{47}$$

To certify a medicament as safe, TI≥10 must hold. It was found that TI=21 for 2-(1-propenyl) phenol, and TI=16 for 2′,4′-dimethylacetophenone (Gálvez et al. 1994b), to be compared with TI=5 for the acetylsalicylic acid. As for the sulfadiazine, the other of the novel molecules, it was found that TI=18 (Gálvez et al. 1994b), which shows again the acceptable degree of innocuity of all these new analgesics, discovered thanks to the methods of molecular topology. Table 2 summarizes these facts.

Conclusion

Molecular topology has widely demonstrated its high performance in the discovery and design of new drugs, which is a major objective of both academia and the pharmaceutical industry. With this review, we seek to contribute to a better knowledge of molecular topology in the scientific community. Bearing this modest scope in mind, we explained the basics of its conceptual framework and reviewed the statistical machinery needed in applications (linear regression and discriminant analysis). The emphasis was laid on the applications to pharmacology, not only because of the scientific bias of the authors but mainly because of their novelty and social impact. In regard with this, let us mention that molecular topology has already achieved breakthroughs in the treatment of malaria, lung cancer, etc. (Jasinski et al. 2008a, 2008b; Mahmoudi et al. 2008), and it is currently being used in the biomedical research on AIDS, Alzheimer, and other major diseases of contemporary medicine. Even more is true: molecular topology can be employed to look for drugs healing, in principle, any disease, based on the structural information provided by known active compounds. One could argue that such "universality" follows from its semiempirical character. But the quantum-mechanical methods are also semiempirical in practice, with different phenomenological parameters having to be fine-tuned to fit the models, while much less applicable and efficient than molecular topology. Once more, "pure" mathematics comes to the rescue in practical problems, this time in the form of graph theory. If the book of nature is written with numbers, as Galileo said, then molecular topology is certainly a way of reading some chapters.

Acknowledgments We are very grateful to the referees for their constructive criticism. We thank very much Ramón García-Domenech (Chemistry Department, University of Valencia) for helpful discussions. Thanks are also due to Óscar Martínez Bonastre (Miguel Hernández University) for assistance with some figures.

References

Amigó JM, Falcó A, Gálvez J, Villar V (2007) Topología molecular. Boletín de la Sociedad Española de Matemática Aplicada 39:137–151

Arviza MP (1985) Predicción e interpretación de algunas propiedades fisicoquímicas y biológicas de un grupo de barbitúricos y sufonamidas por el método de conectividad molecular. PhD Thesis (supervised by J Gálvez) Universidad de Valencia

Balaban AT (1982) Highly discriminating distance-based topological index. Chem Phys Lett 89:399–404

Balaban AT (1987) Numerical modelling of chemical structures: local graph invariants and topological indices. Stud Phys Theor Chem 51:159–176

Balaban AT, Motoc I, Bonchev D, Mekenyan O (1984) Topological indices for structure–activity correlations. Top Curr Chem 114:21–71

Bishop CM (2006) Pattern recognition and machine learning. Springer Verlag, New York

Bollobás B (1998) Modern graph theory. Springer Verlag, New York

Bruno-Blanch L, Gálvez J, García-Domenech R (2003) Topological virtual screening: a way to find new anticonvulsant drugs from chemical diversity. Bioorg Med Chem Lett 13:2749–2754

Carbó-Dorca R, Robert D, Amat L, Gironés X, Besalú E (2000) Molecular quantum similarity in QSAR and drug design. Springer Verlag, Berlin

De Gregorio-Alapont C, García-Domenech R, Gálvez J, Ros MJ, Wolski S, García MD (2000) Molecular topology: a useful tool for the search of new antibacterial. Bioorg Med Chem Lett 10:2033–2036

De Julián-Ortiz JV, Gálvez J, Muñoz-Collado C, García-Domenech R, Jimeno-Cardona C (1999) Virtual combinatorial syntheses and computational screening of new potential anti-herpes compounds. J Med Chem 42:3308–3314

Devillers J, Balaban AT (eds) (2000) Topological indices and related descriptors in QSAR and QSPR. CRC, Boca Raton

Duart MJ, García-Domenech R, Gálvez J, Alemán P, Martín-Algarra RV, Antón-Fos GM (2006) Application of a mathematical topological pattern of antihistaminic activity for the selection of new drug candidates and pharmacology assays. J Med Chem 49:3667–3673

Furnival GM, Wilson RW (1974) Regressions by leaps and bounds. Technometrics 16:499–511

Gálvez J, García-Domenech R, Bernal J, García-March F (1991) Drug design based upon molecular topology: application to non-narcotic analgesics. Anales de la Real Academia de Farmacia 57:533–546

Gálvez J, García R, Salabert MT, Soler R (1994a) Charge indexes - new topological descriptors. J Chem Inf Comput Sci 34:520–525

Gálvez J, García-Domenech R, de Julián-Ortiz JV, Soler R (1994b) Topological approach to analgesia. J Chem Inf Comput Sci 34:1198–1203

Gálvez J, de Julián-Ortiz JV, García-Domenech R (2005) Diseño y desarrollo de nuevos fármacos contra la malaria. Enfermedades Emergentes 7:44–51

García-Domenech R, de Julián-Ortiz JV (2002) Prediction of indices of refraction and glass transition temperatures of linear polymers by using graph theoretical indices. J Phys Chem B 106:1501–1507

García-Domenech R, Villanueva A, Gálvez J, Gozalbes R (1999) Application de la topologie moléculaire a la prédiction de la viscosité liquide des composés organiques. J Chim Phys 96:1172–1185

García-Domenech R, de Julián-Ortiz JV, Duart MJ, García-Torrecillas JM, Antón-Fos GM, Ros-Santamarina I, de Gregorio-Alapont C, Gálvez J (2001) Search of a topological pattern to evaluate toxicity of heterogeneous compounds. SAR & QSAR Environ Res 12:237–254

García-Domenech R, Catalá AI, García-García A, Soriano A, Pérez-Modejar V, Gálvez J (2002) QSAR by molecular topology of 2,4-dihydroxythiobenzanilides: a virtual screening approach to optimize the antifungal activity. Indian J Chem 41B:2376–2384

García-Domenech R, Muñoz-Esp R, Roda-Fenollosa G, Villanueva-Montesinos A, Gálvez J (2003) Predicción de la tensión superficial y la conductividad térmica de disolventes orgánicos mediante la topologa molecular. Afinidad 60:161–168

García-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L (2008) Some new trends in chemical graph theory. Chem Rev 108:1127–1169

Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44:1–12

Hosoya H (1971) A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull Chem Japan 44:2332–2339

Ivanciuc O (1998) Canonical numbering and constitutional symmetry. In: Schleyer PVR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds) The encyclopedia of computational chemistry. John Wiley, Chichester, pp 167–183

Ivanciuc O, Balaban AT (1999) The graph description of chemical structures. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science, The Netherland, pp 59–167

Ivanciuc O, Balaban TS, Balaban AT (1993) Design of topological indices. Part IV: Reciprocal distance matrix, related local vertex invariants and topological indices. J Math Chem 12:309–318

Jasinski P, Welsh W, Gálvez J, Land D, Zwolak P, Ghandi L, Terai K, Dudek AZ (2008a) A novel quinoline, MT477: suppresses cell signalling through Ras molecular pathway, inhibits PKC activity, and demonstrates in vivo anti-tumor activity against human carcinoma cell lines. Invest New Drugs 26:223–232

Jasinski P, Zwolak P, Terai K, Dudek AZ (2008b) Novel Ras pathway inhibitor induces apoptosis and growth inhibition of K-ras-mutated cancer cells in vitro and in vivo. Transl. Res. 152:203–212

Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drugs research. Academic, London

Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Research Studies, Letchworth

Klein DJ, Randić M (1993) Resistance distance. J Math Chem 12:81–95

Mahmoudi N, de Julián-Ortiz JV, Ciceron L, Gálvez J, Mazier D, Danis D, Derouin F, García-Domenech R (2006) Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. J Antimicrob Chemother 57:489–497

Mahmoudi N, García-Domenech R, Gálvez J, Farhati K, Franetich JF, Sauerwine R, Hannoun L, Derouin F, Danis M, Mazier D (2008) New active drugs against liver stages of plasmodium predicted by molecular topology. Antimic Agents and Chem 52:1215–1220

Mallows CL (1973) Some comments on Cp. Technometrics 15:661–675

Miller LC, Tainter ML (1944) Calculation of ED50 and LD50. Proc Soc Exp Biol Med 57:261–264

Newman M, Barabási AL, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Pauling L (1960) The nature of the chemical bond. Cornell University Press, Ithaca

Randić M (1975) On characterization of molecular branching. J Am Chem Soc 97:6609–6615

Randić M (1979) Characterizations of atoms, molecules, and clases of molecules based on path enumerations. Comm Math Chem (MATCH) 7:5–64

Randić M (1984) On molecular identification numbers. J Chem Inf Comput Sci 24:164–175

Randić M (1990) Design of molecules with desired properties. In: Johnson MA, Maggiora GM (eds) Concepts and application of molecular similarity. John Wiley, New York, pp 77–145

Randić M (1992) Similarity based on extended basis descriptors. J Chem Inf Comput Sci 32:686–692

Rios-Santamarina I, García-Domenech R, Cortijo J, Santamaria P, Morcillo EJ, Gálvez J (2002) Natural compounds with bronchodilator activity selected by molecular topology. Internet Electron J Mol Design 1:70–79

Trinajstić N (1992) Chemical graph theory, 2nd edn. CRC, Boca Raton

Trinajstić N, Nikolić S, Lučić B, Amić D, Mihalić Z (1997) The detour matrix in chemistry. J Chem Inf Comput Sci 37:631–638

Tropsha A (2006a) Predictive QSAR modeling. In: Mason J (ed) Comprehensive medicinal chemistry II, V. 4. Elsevier

Tropsha A (2006b) Varible selection QSAR modeling, model validation, and virtual screening. In: Martin Y (ed) Ann Rev Comp Chem. Elsevier, pp 113–126

Tropsha A, Golbraikh A (2007) Predictive QSAR modeling workflow, model applicability domains, and screening. Curr Pharm Des 13:3494–3504

Wang Z, Song J, Chen J, Song Z, Shang S, Jiang Z, Han Z (2008) QSAR study of mosquitoes repellents from terpenoid with a six-member-ring. Bioorg and Med Chem Lett 18:2854–2859

Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69:17–20

Witkin LB, Heubner CF, Galdi F, O'Keefe E, Spitaletta P, Plummer AJ (1961) Pharmacology of 2-amino-indane hydrochloride (Su-8629): a potent non-narcotic analgesic. J Pharmacol Exp Ther 133:400–408

Yaffe D, Cohen Y, Espinosa G, Arenas A, Giralt F (2001) A fuzzy ARTMAP based on quantitative structure–property relationships (QSPRs) for predicting aqueous solubility of organic compounds. J Chem Inf Comput Sci 41:1177–1207