ELSEVIER

Journal of Molecular Structure (Theochem) 582 (2002) 39–51

THEO CHEM

www.elsevier.com/locate/theochem

QSAR for dihydrofolate reductase inhibitors with molecular graph structural descriptors

Ovidiu Ivanciuc[a,*], Teodora Ivanciuc[a], Daniel Cabrol-Bass[b]

[a]*Department of Marine Sciences, Texas A & M University at Galveston, Fort Crockett Campus, 5007 Avenue U, Galveston, TX 77551, USA*
[b]*Aromes, Synthèses et Interactions, Equipe Chimiométrie et Modélisation, University of Nice-Sophia Antipolis, Parc Valrose, 06108 Nice cedex 2, France*

Received 4 September 2001; accepted 2 October 2001

Abstract

Molecular graph descriptors are used, together with a large diversity of geometric, electrostatic, and quantum indices, to model physical, chemical, or biological properties with quantitative structure–property relationships and quantitative structure–activity relationships. The interest of developing new graph descriptors for organic compounds was stimulated in recent years by their use in virtual screening of combinatorial libraries, database mining, similarity and diversity assessment. Recently, we have extended topological indices by defining a series of molecular graph operators, providing an effective systematization and generalization of these structural descriptors. A graph operator uses a mathematical equation to compute a family of related molecular graph descriptors with different molecular matrices and various sets of parameters for atoms and bonds. In this paper we use structural descriptors computed with molecular graph operators to develop quantitative structure–activity relationships (QSAR) models for the dihydrofolate reductase inhibition with diaminopyrimidines. The molecular descriptors are derived from five molecular matrices, namely adjacency **A**, distance **D**, reciprocal distance **RD**, distance-path $\mathbf{D}_p$, and reciprocal distance-path $\mathbf{RD}_p$. The QSAR models are obtained by selecting descriptors with a genetic algorithm, and the best models are validated with the leave-one-out cross-validation method. The QSAR models with the highest prediction power are comparable with those obtained with substituent constants and neural networks, but they use a much lower number of parameters.

Keywords: Structural descriptors; Molecular graph operators; Dihydrofolate reductase inhibitors; Genetic algorithm; Quantitative structure–activity relationships

1. Introduction

A large number of constitutional, topological, geometric, electrostatic, and quantum indices were introduced in theoretical chemistry with the aim to express in a numerical form the chemical structure. Such structural descriptors can be used to model physical, chemical, or biological properties with quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR). The most efficient software used in QSPR or QSAR studies integrate the computation of structural descriptors with the generation of structure–property models. Several programs from this category, such as ADAPT [1–4], OASIS [5,6],

* Corresponding author. Tel.: +1-409-740-4947; fax: +1-409-740-4787.
E-mail address: ivanciuc@netscape.net (O. Ivanciuc).

0166-1280/02/$ - see front matter
PII: S0166-1280(01)00772-2

SciQSAR,[1] CODESSA [7–13], Cerius[2],[2] were used with success in developing a large number of QSPR and QSAR models; these programs compute more than one thousand structural descriptors from the above five classes, a significant fraction of them being derived from the molecular graph. A survey of the QSPR and QSAR models developed with these programs shows that molecular graph descriptors and topological indices are used with success to model various properties, and demonstrates that they are valuable descriptors of the chemical structure.

Molecular graphs are non-directed chemical graphs that represent, in different conventions, molecules. In the graph representation of molecules the geometrical features, like bond lengths or bond angles, are not considered and the chemical bonding of atoms is regarded as being their most important characteristic. In molecular graphs vertices correspond to atoms and edges represent covalent bonds between atoms. Using molecular graphs the chemical structure of an organic compound can be expressed by means of various graph matrices, polynomials, spectra, spectral moments, sequences counting distances, paths, and walks, or topological indices [14–23]. When compared with other classes of structural descriptors, such as geometric, quantum, or grid (field) descriptors, topological indices (TIs) have some important advantages because they can be easily computed from the molecular graph and they offer a simple way of measuring molecular branching, shape, size, complexity, and molecular similarity. On the other hand, because TIs are global descriptors of the molecular graph, they do not contain explicit information regarding the number of functional groups, pharmacophores, volume, surface area, interatomic distances, charge distribution, orbital energy, or electrostatic potential; for the generation of QSPR and QSAR models such information must be provided by other structural descriptors. Considering the advantages of graph invariants, TIs represent valuable descriptors that complement (and do not substitute) the structural information encoded in other classes of descriptors.

Recently we have introduced molecular graph operators as an extension of topological indices; a graph operator uses a mathematical equation to compute a family of related molecular graph descriptors with different molecular matrices and various sets of parameters for atoms and bonds [19,24,25]. The large majority of the topological indices proposed in the literature were derived from the adjacency and the distance matrix. In the last years, several new molecular matrices were defined and used to compute new structural descriptors [19,20]. The use of molecular graph operators introduces a systematization of topological indices by putting together all descriptors computed with the same mathematical formula or algorithm. As a consequence, when new molecular matrices are introduced there is no need to invent new names and symbols for the topological indices derived from them; the notation of graph operators is simple and general, and can accommodate new matrices, weighting schemes, and any parameter used in the definition of a family of topological indices. In this paper we use structural descriptors computed with molecular graph operators to develop QSAR models for the dihydrofolate reductase inhibition with diaminopyrimidines. The molecular descriptors are derived from five molecular matrices, namely adjacency **A**, distance **D**, reciprocal distance **RD** [26–29], distance-path $\mathbf{D}_p$ [30,31], and reciprocal distance-path $\mathbf{RD}_p$ [30,31].

2. Method

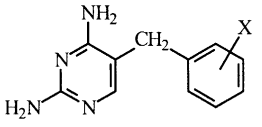
Data. The QSAR models were developed for the inhibition of dihydrofolate reductase by a set of 68 2,4-diamino-5-(substituted-benzyl) pyrimidines whose general formula, substituents, and experimental activity are presented in Table 1. This set of compounds has been extensively investigated by Hansch [32,33]. The proposed QSAR was formulated with a multi-linear regression (MLR) model and used empirical substituent constants [34] in an equation with eight adjustable coefficients:

$$\log(1/K) = 6.65 + 0.95\mathrm{MR}'_5 + 0.89\mathrm{MR}'_3 + 0.80\mathrm{MR}_4 - 0.21\mathrm{MR}_4^2 + 1.58\pi'_3 - 1.77\log(\beta 10^{\pi 3} + 1) \quad (1)$$

[1] SciQSAR, SciVision, Inc., 200 Wheeler Road, Burlington, MA 01803, USA. Tel.: +1-781-272-4949; fax: +1-781-272-6868, e-mail: scivision@delphi.com, www http://www.scivision.com.

[2] Cerius2 3.0 QSAR + , Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752. Fax: +1-619/458-0136, 1997.

Table 1
Structures of the pyrimidines congeners, observed dihydrofolate reductase inhibitory activities, calibration and prediction and residuals computed with the QSAR model represented by Eq. (17) from Table 2



No.	X	log (1/*K*)		
		exp	res_{MLR} [a]	res_{LOO} [b]
1	4-$O(CH_2)_5CH_3$	6.07	−0.25	−0.28
2	4-$O(CH_2)_6CH_3$	6.10	−0.09	−0.12
3	H	6.18	0.04	0.03
4	4-NO_2	6.20	−0.16	−0.16
5	3-F	6.23	−0.24	−0.24
6	3-$O(CH_2)_7CH_3$	6.25	−0.02	−0.05
7	3-CH_2OH	6.28	−0.21	−0.22
8	4-NH_2	6.30	−0.03	−0.04
9	3,5-$(CH_2OH)_2$	6.31	−0.60	−0.60
10	4-F	6.35	−0.02	−0.02
11	3-$O(CH_2)_6CH_3$	6.39	−0.02	−0.05
12	4-$OCH_2CH_2OCH_3$	6.40	−0.27	−0.28
13	4-OH	6.45	0.11	0.10
14	4-Cl	6.45	−0.09	−0.09
15	3,4-$(OH)_2$	6.46	−0.22	−0.22
16	3-OH	6.47	0.04	0.04
17	4-CH_3	6.48	0.07	0.06
18	3-$OCH_2CH_2OCH_3$	6.53	−0.36	−0.37
19	3-$CH_2O(CH_2)_3CH_3$	6.55	−0.02	−0.03
20	3-OCH_2CONH_2	6.57	−0.17	−0.18
21	4-OCF_3	6.57	−0.08	−0.08
22	3-CH_2OCH_3	6.59	−0.16	−0.16
23	4-OSO_2CH_3	6.60	−0.40	−0.40
24	3-Cl	6.65	0.04	0.04
25	3-CH_3	6.70	0.22	0.21
26	4-$N(CH_3)_2$	6.78	−0.12	−0.13
27	3-$O(CH_2)_3CH_3$	6.82	0.06	0.05
28	4-OCH_3	6.82	0.16	0.16
29	4-Br	6.82	−0.03	−0.04
30	3-OH, 4-OCH_3	6.84	−0.17	−0.17
31	3-$O(CH_2)_5CH_3$	6.86	0.31	0.29
32	4-$NHCOCH_3$	6.89	0.25	0.25
33	4-$O(CH_2)_3CH_3$	6.89	0.35	0.33
34	4-$OCH_2C_6H_5$	6.89	0.04	0.03
35	3-OSO_2CH_3	6.92	−0.30	−0.30
36	3-OCH_3	6.93	0.13	0.13
37	4-C_6H_5	6.93	−0.18	−0.19
38	3-Br	6.96	0.03	0.03
39	3-NO_2,4-$NHCOCH_3$	6.97	0.23	0.24
40	3-$OCH_2C_6H_5$	6.99	−0.12	−0.12
41	3-CF_3	7.02	0.15	0.15
42	3,5-$(CH_3)_2$	7.04	0.18	0.18
43	3,4-OCH_2O	7.13	0.05	0.06
44	3-$O(CH_2)_7CH_3$,4-OCH_3	7.16	0.28	0.26
45	3,5-$(OCH_3)_2$,4-$O(CH_2)_7CH_3$	7.20	−0.16	−0.18
46	3,4-$(OCH_2CH_2OCH_3)_2$	7.22	−0.26	−0.27
47	3-I	7.23	0.13	0.12
48	3-OCH_2CH_3,4-$OCH_2C_6H_5$	7.35	−0.08	−0.08
49	3,5-$(OC_3H_7)_2$	7.41	−0.11	−0.12
50	3-OCH_3,4-$OCH_2C_6H_5$	7.53	0.13	0.12
51	3-OCH_3,4-OH	7.54	0.47	0.48
52	3,5-$(OCH_2CH_3)_2$,4-pyrryl	7.66	−0.17	−0.17
53	3-$OCH_2C_6H_5$,4-OCH_3	7.66	0.13	0.13
54	3,5-$(OCH_2CH_3)_2$	7.69	0.12	0.12
55	3-OC_2H_5,5-OC_3H_7	7.69	0.15	0.14
56	3-CF_3,4-OCH_3	7.69	0.27	0.27
57	3,5-$(OCH_3)_2$,4-$N(CH_3)_2$	7.71	−0.63	−0.62
58	3,5-$(OCH_3)_2$	7.71	0.23	0.23
59	3,4-$(OCH_3)_2$	7.72	0.32	0.32
60	3-OCH_3,4-$OCH_2CH_2OCH_3$	7.77	0.36	0.36
61	3-OSO_2CH_3,4-OCH_3	7.80	0.06	0.07
62	3,4,5-$(CH_2CH_3)_3$	7.82	0.24	0.23
63	3-OCH_3,4-OSO_2CH_3	7.94	0.28	0.29
64	3,5-$(OCH_3)_2$,4-SCH_3	8.07	−0.10	−0.09
65	3,4,5-$(OCH_3)_3$	8.08	0.03	0.04
66	3,5-$(OCH_3)_2$,4-$C(CH_3)=CH_2$	8.12	−0.07	−0.06
67	3,5-$(OCH_3)_2$,4-Br	8.18	−0.12	−0.11
68	3,5-$(OCH_3)_2$,4-$O(CH_2)_2OCH_3$	8.35	0.36	0.36

[a] $res_{MLR} = \log(1/K)_{exp} - \log(1/K)_{cal}$ Eq. (17).
[b] $res_{LOO} = \log(1/K)_{exp} - \log(1/K)_{pre}$ Eq. (17).

with $\log \beta = 0.175$, $MR_4^0 = 1.85$, and $\pi_3^0 = 0.73$. The same set of data has been studied by Richards [35] with the aid of an artificial neural network with four input neurons, six hidden neurons, and one output neuron. The neural network, with 38 adjustable coefficients, gave better results at the expense of a much larger number of optimizable parameters.

Weighting schemes. In the chemical graph theory, an organic compound containing heteroatoms and multiple bonds can be represented as a vertex- and edge-weighted molecular graph [18,19,24]. A vertex- and edge-weighted (VEW) molecular graph G consists of a vertex set $V = V(G)$ an edge set $E = E(G)$, a set of chemical symbols of the vertices $Sy = Sy(G)$, a set of topological bond orders of the

edges $Bo = Bo(G)$, a vertex weight set $Vw(w) = Vw(w, G)$ and an edge weight set $Ew(w) = Ew(G)$. The elements of the vertex and edge weight sets are computed with the weighting scheme w. Usually, hydrogen atoms are not considered in the molecular graph, and in a VEW graph the weight of a vertex corresponding to a carbon atom is 0, while the weight of an edge corresponding to a carbon–carbon single bond is 1. Also, the topological bond order Bo_{ij} of an edge e_{ij} takes the value 1 for single bonds, 2 for double bonds, 3 for triple bonds and 1.5 for aromatic bonds. Several procedures for computing vertex and edge weights were proposed in the literature [36–41].

In a weighting scheme w the vertex Vw and edge Ew parameters are computed from a property p_i associated with every vertex v_i from G, $v_i \in V(G)$, and the topological bond order Bo of all edges from the molecular graph. The vertex parameter $Vw_i(w)$ for the vertex v_i is:

$$Vw_i(w) = 1 - p_C/p_i \tag{2}$$

and the edge parameter $Ew_{ij}(w)$ for the edge between vertices v_i and v_j is:

$$Ew_{ij}(w) = p_C p_C/Bo_{ij} p_i p_j \tag{3}$$

where p_i is the atomic property of vertex v_i, p_j is the atomic property of vertex v_j, and p_C is the atomic property for carbon atom. Several weighting schemes for molecular graphs were defined by applying Eqs. (2) and (3) to different atomic properties [39]: A, when p is the atomic mass; P, when p is the atomic polarizability; E, when p is the atomic electronegativity; R, when p is the atomic radius. Similar equations were used to define the X, Y [38], and Z [36] weighting schemes.

The AH weighting scheme [39] uses the following equation to define the vertex parameter $Vw_i(AH)$ for the non-hydrogen atom i:

$$\begin{aligned} Vw_i(AH) &= 1 - A_C/(A_i + NoH_i A_H) \\ &= 1 - 12.011/(A_i + 1.0079NoH_i) \end{aligned} \tag{4}$$

The edge parameter $Ew_{ij}(AH)$ for the bond between atoms i and j is defined with the equation [39]:

$$\begin{aligned} Ew_{ij}(AH) &= A_C A_C/Bo_{ij}(A_i + NoH_i A_H)(A_j + NoH_j A_H) \\ &= 12.011 \times 12.011/Bo_{ij}(A_i + 1.0079NoH_i) \\ &\quad \times (A_j + 1.0079NoH_j) \end{aligned} \tag{5}$$

where $A_C = 10.011$ is the atomic mass for carbon, $A_H = 1.0079$ is the atomic mass for hydrogen, NoH_i is the number of hydrogen atoms bonded to the heavy atom i, and NoH_j is the number of hydrogen atoms bonded to the heavy atom j.

Molecular matrices. In this study the graph descriptors are computed from the following five molecular matrices: adjacency **A**, distance **D**, reciprocal distance **RD** [26–29], distance-path $\mathbf{D}_p$ [30,31], and reciprocal distance-path $\mathbf{RD}_p$ [30,31].

Structural descriptors. The scope of this paper is to investigate the ability of structural descriptors computed from molecular graph operators to generate good QSAR models for the inhibition of dihydrofolate reductase by diaminopyrimidines. Owing to the complexity of the molecular structure, it seems to be impossible to expect that a single set of descriptors would contain all the relevant information. This is the main reason why QSPR and QSAR models are developed by selecting descriptors from a large pool of structural descriptors. The list of the 271 structural descriptors used in the QSAR study is presented below:

1. the molecular weight, **MW**; this constitutional descriptor is a measure of molecular size, and experience shows that in many structure–property studies this is an important parameter.
2. Five Kier and Hall connectivity indices [14,15] ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi_p^v$, ${}^3\chi_c^v$; these topological indices are, by far, the most used descriptors in QSPR and QSAR models.
3. 75 Wiener indices computed with the Wiener operator $\mathbf{Wi}(\mathbf{M}^p,w)$ [19]; these descriptors represent an extension of the Wiener index that is defined for any molecular matrix. The Wiener operator $\mathbf{Wi}(\mathbf{M}^p,w,G)$ of a VEW molecular graph G with N vertices is computed from the pth power of the symmetric $N \times N$ molecular matrix $\mathbf{M}(w,G)$:

$$\mathbf{Wi}(\mathbf{M}^p, w, G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{M}_{ij}(w, G)]^p \tag{6}$$

where w is the weighting scheme used to compute the molecular matrix $\mathbf{M}(w)$, and p takes the values 1,2, and 3. Molecular graph descriptors computed with the Wiener operator were used with success to compute the boiling points of acyclic compounds containing oxygen or sulfur atoms [38].

4. 15 hyper-Wiener indices computed with the hyper-Wiener operator **HyWi**(**M**,w) [19,20]. The hyper-Wiener operator **HyWi**(**M**,w,G) of a vertex- and edge-weighted molecular graph G with N vertices is computed from the symmetric $N \times N$ molecular matrix **M**(w,G):

$$\mathbf{HyWi}(\mathbf{M}, w, G) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=i}^{N} [[\mathbf{M}_{ij}(w, G)]^2 + \mathbf{M}_{ij}(w, G)] \quad (7)$$

where w is the weighting scheme. Only the adjacency **A**, distance **D**, and reciprocal distance **RD** matrices were used to compute the hyper-Wiener indices.

5. 50 spectral descriptors computed with the spectral operators **MinSp**(**M**,w) and **MaxSp**(**M**,w) [19]. The matrix spectrum operator $\mathbf{Sp}(\mathbf{M}, w, G) = \{x_i, i = 1, 2, \ldots, N\}$ represents the eigenvalues of the matrix **M**(w) or the roots of the polynomial $\mathbf{Ch}(\mathbf{M}, G, x), \mathbf{Ch}(\mathbf{M}, w, G, x) = 0$. The spectral operators **MinSp**(**M**,w,G) and **MaxSp**(**M**,w,G) are equal to the minimum and maximum values of **Sp**(**M**,w,G), respectively:

$$\mathbf{MinSp}(\mathbf{M}, w, G) = \mathbf{min}\{\mathbf{Sp}(\mathbf{M}, w, G)\} \quad (8)$$

$$\mathbf{MaxSp}(\mathbf{M}, w, G) = \mathbf{max}\{\mathbf{Sp}(\mathbf{M}, w, G)\} \quad (9)$$

Structural descriptors computed with these two spectral operators were used with good results to develop QSPR models for the boiling points, heat of vaporization, molar refraction, molar volume, critical pressure, critical temperature, and surface tension of alkanes [29,42], to estimate the boiling points of acyclic compounds containing oxygen or sulfur atoms [38], and to define a chemical space for clustering molecules from chemical libraries [40,41].

6. 125 **Chi** indices computed with the operator **Chi**(**VSD**,**M**,w). Consider a vertex structural descriptor **VSD**(**M**,w,G) that assigns a numerical invariant $\mathbf{VSD}_i(\mathbf{M},w,G)$ to each vertex v_i from the VEW molecular graph G. The **Chi** operator **Chi**(**VSD**,**M**,w,G) of the graph G is:

$${}^{m}\mathbf{Chi}(\mathbf{VSD}, \mathbf{M}, w)_t = \sum_{i=1}^{s} \prod_{j=1}^{n} (\mathbf{VSD}(\mathbf{M}, w)_j)^{-1/2} \quad (10)$$

where s is the number of connected subgraphs of type t with m edges, n is the number of vertices of the subgraph, and w is the weighting scheme. The **Chi** operator represents a generalization of the Kier and Hall connectivity indices [14,15] by replacing the local invariant δ^v with any other vertex invariant. Five types of **Chi** indices were computed, namely ${}^{0}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)$, ${}^{1}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)$, ${}^{2}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)$, ${}^{3}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)_p$,and ${}^{3}\mathbf{Chi}(\mathbf{VSD},\mathbf{M},w)_c$. The vertex structural descriptors considered in this study are the valency **val**(w) and vertex sum **VS**(**M**,w) operators. The valency of the vertex v_i, $\mathbf{val}_i(w,G)$, is defined as the sum of the weights $\mathrm{Ew}_{ij}(w)$ of all edges e_{ij} incident with vertex v_i [42,43]:

$$\mathbf{val}_i(w, G) = \sum_{e_{ij} \in E(G)} \mathrm{Ew}_{ij}(w) \quad (11)$$

where w is the weighting scheme used to compute the Ew parameters.

Consider the vertex v_i from the weighted graph G with N vertices and the symmetric graph matrix $\mathbf{M}(w,G)$ computed with the weighting scheme w. The vertex sum of the vertex v_i, $\mathbf{VS}_i(\mathbf{M},w,G)$, is defined as the sum of the elements in the column i, or row i, of the molecular matrix **M**(w) [42,43]:

$$\mathbf{VS}_i(\mathbf{M}, w, G) = \sum_{j=1}^{N} \mathbf{M}_{ij}(w) = \sum_{j=1}^{N} \mathbf{M}_{ji}(w) \quad (12)$$

QSAR model. The QSAR equations were developed using multi-linear regression models. Because the exhaustive test of all MLR equations requires important computational resources, all studies that develop QSPR and QSAR models from a large set of computed descriptors use a wide range of algorithms for selecting significant descriptors, such as genetic and evolutionary algorithms [44–47]. In our investigation we generated the QSAR models using the genetic function approximation (GFA) [47] feature selection as implemented in Cerius^{2}.[2]

The GFA algorithm generates an initial population of QSAR equations by the random selection of molecular descriptors. In our study we have used QSAR models containing only linear terms; the experiments that employed non-linear terms did not show any significant improvement. The length of the equations is determined by the number of molecular descriptors selected and is allowed to increase or decrease with a certain probability. Each equation is fit to the experimental data using linear least-squares regression techniques, and the QSAR models are ranked according to their lack of fit (LOF), which is an adjusted least-squares error (LSE) statistical index:

$$\mathrm{LOF} = \frac{\mathrm{LSE}}{\left(1 - \frac{c + df}{n}\right)^2} \tag{13}$$

where c is the number of basis functions (number of terms in the QSAR equation), d the smoothing parameter, f the total number of structural descriptors contained in all basis functions, and n is the number of molecules in the calibration set. The addition of a new descriptor to a QSAR equation may reduce the LSE, but it also increases the values of c and f, and the LOF index may increase. In this way, the LOF index avoids overfitting of the data by limiting the tendency to add new descriptors and favoring simple, more compact models. The user-defined smoothing parameter d controls the growing of the QSAR equations. Because values larger than 1 for this parameter favor equations with fewer descriptors, we have used a value of 2 for d in all experiments.

The GFA algorithm uses a genetic algorithm to perform the genetic crossover of the best equations and the elimination of the poorer equations. Using pairs of QSAR equations with a low LOF, the genetic algorithm cuts and separates pieces of equations and then recombines the fragments to form new equations. Next, mutation operators are applied randomly in order to increase the diversity of the population. The following probability levels p were used for the equation mutation operators:

1. *Add new term*. $p = 0.5$; this mutation increases the diversity in the equation population by randomly adding a new term to a child equation.
2. *Reduce equation*. $p = 0.5$; this mutation operation generates smaller models by eliminating from a child equation the term with the lowest contribution to the model.
3. *Extend equation*. $p = 0.5$; this mutation operation generates larger models by adding to a child equation a new descriptor. Each descriptor not yet used in the QSAR model is tested and the selection algorithm retains the one that has the highest contribution to the model.

The genetic algorithm uses the LOF index to select equations for crossover and survival. With new generations, the equation population evolves to a set of higher quality QSAR models. The output of a GFA calculation consists of a population of QSAR equations that model the relationship between the structural descriptors and the biological activity.

3. Results and discussion

The equation population consists of 300 equations containing only linear terms. The evolution of the equation population was performed for 5000 generations, when the LOF index indicated that the population of equations stops significantly improving. From the final population of 300 equations we have selected the best ten QSAR models (according to LOF) with up to five structural descriptors. Selecting several QSAR models offers the possibility to study the distribution of structural descriptors, molecular matrices and weighting schemes. The statistical indices of Eqs. (14)–(23) are presented in Table 2

$$\begin{aligned}\log(1/K) = {} & -21.8289 - 0.001715\mathbf{Wi}\left(\mathbf{RD}_{\mathrm{p}}^{3}, E\right) \\ & + 0.487974\,{}^{0}\chi^{v} + 4.28779\mathbf{MaxSp}(\mathbf{RD}, E) \\ & - 1.09075\,{}^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_{\mathrm{p}}), E)\end{aligned} \tag{14}$$

$$\begin{aligned}\log(1/K) = {} & -18.6578 - 0.001509\mathbf{Wi}\left(\mathbf{RD}_{\mathrm{p}}^{3}, E\right) \\ & - 1.42469\,{}^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}), E) \\ & + 0.424785\,{}^{0}\chi^{v} + 4.05\mathbf{MaxSp}(\mathbf{RD}, E)\end{aligned} \tag{15}$$

$$\log(1/K) = -27.1388^{3}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}), A)_p - 0.005413\mathbf{MaxSp}(\mathbf{D}_\mathrm{p}, AH) - 0.000247\mathbf{Wi}(\mathbf{RD}^3, R) + 0.563187^{0}\chi^{v} + 41.9252^{3}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}), AH)_p \quad (16)$$

$$\log(1/K) = -23.7395 - 0.174763^{1}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), R) + 0.450132^{0}\chi^{v} + 4.56168\mathbf{MaxSp}(\mathbf{RD}, E) - 0.96052^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.001825\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) \quad (17)$$

$$\log(1/K) = -21.0551 + 0.460472^{0}\chi^{v} + 4.06631\mathbf{MaxSp}(\mathbf{RD}, E) - 0.878343^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.001599\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) - 0.007035\mathbf{MaxSp}(\mathbf{D}, R) \quad (18)$$

$$\log(1/K) = -21.5116 + 1.91467^{3}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}), AH)_p + 0.526038^{0}\chi^{v} + 4.17073\mathbf{MaxSp}(\mathbf{RD}, E) - 1.15385^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.001672\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) \quad (19)$$

$$\log(1/K) = -20.5146 - 0.000839\mathbf{MaxSp}(\mathbf{D_p}, R) + 3.99309\mathbf{MaxSp}(\mathbf{RD}, E) - 0.911014^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.001594\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) + 0.464312^{0}\chi^{v} \quad (20)$$

$$\log(1/K) = -22.161 + 1.46601^{3}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}), A)_p + 0.528748^{0}\chi^{v} + 4.28698\mathbf{MaxSp}(\mathbf{RD}, E) - 1.18821^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.001704\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) \quad (21)$$

$$\log(1/K) = -19.9524 + 0.016933\mathbf{Wi}(\mathbf{A}^3, A) + 0.532245^{0}\chi^{v} + 4.02885\mathbf{MaxSp}(\mathbf{RD}, E) - 1.25614^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.001616\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) \quad (22)$$

$$\log(1/K) = -20.9523 - 0.001606\mathbf{Wi}\left(\mathbf{RD}_\mathrm{p}^3, E\right) + 4.13974\mathbf{MaxSp}(\mathbf{RD}, E) - 1.13923^{0}\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_\mathrm{p}), E) - 0.000324\mathbf{Wi}(\mathbf{RD}^2, AH) + 0.550071^{0}\chi^{v} \quad (23)$$

From the set of 10 QSAR models, Eqs. (14) and (15) have four structural descriptors, while Eqs. (16)–(23) have five structural descriptors. In restricting our selection to models with up to five structural descriptors we have given preference to simple, more robust models, because with the increase of the number of descriptors it increases also the danger of overfitting. An inspection of the descriptors selected in Eq. (14) shows that this QSAR model is the basis for the development of Eqs. (17)–(23); all these equations contain the four topological indices from Eq. (14), and a fifth one that introduces a small variability of the statistical indices, as can be seen from Table 2. This indicates that the population of QSAR models converged to a set of related equations that represent a suboptimal solution for the investigated QSAR model. We have to mention that a genetic algorithm is not intended to find the optimal solution to a problem, but to search in a highly dimensional space of structural descriptors and to identify in a (relatively) short time a set of suboptimal solutions; this is the method of choice

Table 2
Statistical indices of the QSAR models from Eqs. (14)–(23)

Eq.	LOF	r^2	LSE	F	r^2_{LOO}	$PRESS_{LOO}$
(14)	0.074	0.859	0.050	96.18	0.861	3.368
(15)	0.076	0.855	0.052	92.80	0.857	3.473
(16)	0.080	0.859	0.050	97.14	0.861	3.375
(17)	0.080	0.863	0.049	78.44	0.865	3.267
(18)	0.081	0.862	0.049	77.72	0.864	3.294
(19)	0.081	0.862	0.049	77.70	0.864	3.295
(20)	0.081	0.862	0.049	77.58	0.864	3.299
(21)	0.081	0.862	0.049	77.54	0.864	3.300
(22)	0.081	0.862	0.049	77.44	0.864	3.293
(23)	0.081	0.862	0.049	77.40	0.864	3.292

whenever the exhaustive search of the solution space is not possible in a reasonable time.

An important problem in QSAR studies is the identification of structural descriptors that are to be used to model a given property. We have to point here that correlational analysis develops models by suggesting statistical (and not causal) relationships between structural descriptors and a physical, chemical, or biological property. Correlations can be observed not only because a causal relationship exists between a set of descriptors and a property, but also due to statistical bias resulting from errors in determining structural descriptors, experimental errors in measuring the property, or even due to chance alone. When a correlation appears due to errors or chance factors, the resulting QSAR model has no scientific content and offers misleading conclusions. As was shown by Topliss [48], such a situation can easily occur if large numbers of structural descriptors are screened for potential inclusion into the final correlation equation. Model validation techniques are used to distinguish between true and random correlations and to estimate the predictive power of the model [49]. For Eqs. (14)–(23) we have used the leave-one-out (LOO) cross-validation procedure; the obtained correlation coefficient (r_{LOO}) and $PRESS_{LOO}$ are reported in the last two columns of Table 2.

In Table 2 the QSAR models are arranged in the increasing order of LOF, as offered by the GFA algorithm. The statistical indices of the 10 QSAR models are very close, in both calibration and prediction, indicating that their statistical quality is very similar. In fact, these equations can be ordered in different ways by the statistical indices reported in Table 2 as already pointed, LOF ranks first Eq. (14); r^2, r^2_{LOO}, and $PRESS_{LOO}$ put Eq. (17) on the first place; Eq. (16) has the highest F value; Eqs. (17)–(23) all have the smallest LSE. However, the statistical differences of the QSAR models are not high, and this offers the possibility to study the distribution of structural descriptors, molecular matrices and weighting schemes.

The connectivity index ${}^0\chi^v$ appears in all QSAR models, and this is the single descriptor from its class selected by the GFA algorithm. The index ${}^0\chi^v$ represents the weighted contribution of subgraphs containing one non-hydrogen atom and it is a measure of molecular size. On the other hand, the molecular weight was not selected, indicating that this descriptor was successfully replaced by ${}^0\chi^v$.

The Wiener indices **Wi** were selected once in eight QSAR models and twice in two QSAR models, indicating the high significance of such descriptors in modeling the dihydrofolate reductase inhibition. The Wiener operator represents a global measure of molecular size, shape, and branching, and our previous studies showed that this class of graph descriptors have a high importance in modeling various molecular properties. The hyper-Wiener indices **HyWi** were not found important by the GFA algorithm, since none of them was selected in Eqs. (14)–(23); this observation is in line with our previous results, which indicated that this class of structural descriptors is not suitable for structure–property models.

The results provided by the GFA algorithm demonstrate the importance of the spectral indices computed with the **MaxSp** operator in this QSAR, because these indices appear twice in two QSAR models, and once in eight QSAR models. The **MaxSp** operator is a measure of molecular shape and branching, relatively independent of the molecular size. On the other hand, the **MinSp** operator is not significant for modeling the dihydrofolate reductase inhibition, because in Eqs. (14)–(23) this class of topological indices was not selected.

From the five types of descriptors derived from the **Chi** operator, only ^{0}**Chi**(**VSD**,**M**,w) and ^{3}**Chi**(**VSD**,**M**,w)$_p$ appear with a higher frequency: ^{0}**Chi**(**VSD**,**M**,w), nine times; ^{1}**Chi**(**VSD**,**M**,w), once; ^{3}**Chi**(**VSD**,**M**,w)$_p$, four times. The remaining two **Chi** operators, ^{2}**Chi**(**VSD**,**M**,w) and ^{3}**Chi**(**VSD**,**M**,w)$_c$, were not selected by the GFA algorithm. The index

^{0}**Chi**(**VSD**,**M**,*w*) represents the weighted contribution of subgraphs containing one non-hydrogen atom and is mainly a measure of molecular size; the shape component is introduced by the **VSD** atomic invariant. The descriptor ^{3}**Chi**(**VSD**,**M**,*w*)$_p$ is a weighted sum of the butane-like subgraphs from the molecular graph and represents the molecular shape and branching. From the two types of **VSD** atomic invariant investigated, only the **Chi** indices computed with the vertex sum operator **VS**(**M**,*w*,) were selected by the GFA algorithm. The structural information contained by the valency operator **val**(*w*) measures only the local (atomic) structure, while that from **VS**(**M**,*w*,) reflects at the atomic level the global (molecular) structure. The higher content of structural information contained into the **VS**(**M**,*w*,) indices can explain the absence of the **val**(*w*) indices.

An analysis of the presence of different molecular matrices in the structural descriptors selected in Eqs. (14)–(23) reveals several interesting tendencies: structural descriptors computed with the reciprocal matrices **RD** and **RD**$_p$ were selected in the majority of cases, while those derived from the adjacency **A**, distance **D**, and distance-path **D**$_p$ matrices were considered by the GFA algorithm to be less important. This observation is apparent from the frequency of selection of different molecular matrices: $\mathbf{A}^3$, once; **D**, once; **RD**, 14 times; $\mathbf{RD}^2$, once; $\mathbf{RD}^3$, once; **D**$_p$, twice; **RD**$_p$, 9 times; $\mathbf{RD}_p^3$, 9 times. From these results, it appears also that from the higher power matrices, only $\mathbf{RD}_p^3$ has a greater importance. Until recently, the molecular graph descriptors were mainly computed from the adjacency and distance matrices. Our finding indicates that the structural descriptors derived from the recently introduced molecular matrices are more suitable for developing relevant QSAR models.

Using various atomic properties, the weighting schemes used in this study offer the atom and bond parameters for the computation of the topological indices. The correlational ability of the topological indices depends heavily on the weighting scheme; those computed with atomic electronegativity parameters *E* were selected in the majority of cases, while those derived from the atomic polarizability parameters *P* were not selected in Eqs. (14)–(23): *E*, 27 times; *R*, 4 times; AH, 4 times; *A*, 3 times. While our results indicate that the structural descriptors computed with atomic electronegativity parameters give good results in modeling the dihydrofolate reductase inhibition, further experiments are required to determine if this is a particular behavior or represents a more general tendency.

As already mentioned, the statistical indices from Table 2 show that the statistical differences of the QSAR models in Eqs. (14)–(23) are not high; we will now examine in detail Eq. (17), because it gives the best prediction results. In Table 1 we give the calibration and prediction (LOO) residuals computed with Eq. (17). These results show that there are 18 compounds with an absolute residual greater than 0.25, both in calibration and prediction. Richards used this criterion to compare the MLR model from Eq. (1) and the neural network model [35]; Eq. (1) has 25 compounds and the neural network has 12 compounds with an absolute residual greater than 0.25 (residual outliers). From these data it appears that our QSAR model is between the previously proposed models. However, one has to consider that Eq. (17), with five structural descriptors, has six adjustable coefficients, Eq. (1) has eight adjustable coefficients, while the neural network has 38 adjustable coefficients. Artificial neural networks represent a simple and efficient way to generate non-linear QSAR models, but in many cases this is obtained at the expense of a large number of optimizable parameters; this is the case for the set of data investigated here, where the improvement in modeling the inhibition of dihydrofolate reductase is not justified by such an important increase in the number of optimizable parameters. A neural network that has too many optimizable parameters (connections between neurons) gives good results for the model calibration, but the prediction results are of poorer quality. For the data from Table 1, the neural network gives 12 residual outliers in calibration and 21 in LOO prediction [35]; this degradation of the results for prediction compared with calibration is a clear sign of overfitting, indicating that the model is not stable for predicting the activity for new compounds. The main objective in QSAR studies is to obtain a model with the highest predictive ability, and the neural network fails to attain this goal. As pointed above, Eq. (17) has 18 residual outliers both in calibration and prediction, demonstrating that this QSAR model is stable, is not overfitted, and it gives better predictions than the neural network model. Also, we have to point that

both Eq. (1) and neural network models represent nonlinear QSAR models, while in Eqs. (14)–(23) we have explored only linear models. Another advantage of Eqs. (14)–(23) is the use of molecular graph descriptors that can be calculated from the molecular graph for any organic compound.

The strong intercorrelation between structural descriptors from a MLR equation may lead to misinterpretation of the corresponding structure–activity model. The GFA algorithm does not test the intercorrelation of the structural descriptors selected by the genetic evolution, and can produce QSAR models that contain highly intercorrelated independent variables. In Table 3 we give the intercorrelation matrix of the five structural descriptors in Eq. (17), together with the individual correlation coefficient of the descriptors with the biological activity. From this matrix one can see that five pairs of descriptors are highly intercorrelated, with $|r_{ij}| > 0.80$ (their intercorrelation coefficients are highlighted in bold in Table 3): $\{{}^0\chi^v$,**MaxSp**(**RD**,E)$\}$, $\{{}^0\chi^v$,^{0}**Chi**(**VS**(**RD**$_p$),E)$\}$, $\{{}^0\chi^v$,^{1}**Chi**(**VS**(**RD**$_p$),R)$\}$, $\{$**MaxSp**(**RD**,E), **Wi**(**RD**$_p^3$,E)$\}$, $\{{}^0$**Chi**(**VS**(**RD**$_p$),E),^{1}**Chi**(**VS**(**RD**$_p$),R)$\}$. Several techniques can be applied to overcome the problem of highly intercorrelated descriptors: PCA, PLS, or sequential orthogonalization. We have selected the recently introduced sequential orthogonalization [50] that was applied with success for modeling the chromatographic retention indices [22,51].

In the sequential orthogonalization algorithm, a descriptor from the set of intercorrelated structural descriptors can be made orthogonal, by removing the part of its information content that it shares with the other descriptors in the set. The order in which descriptors are orthogonalized is important, because it strongly affects the information content of so obtained orthogonal descriptors. We present briefly the algorithm implementation used in the present investigation. Consider the MLR equation that models the property Y with the set of three structural descriptors X_1, X_2, and X_3:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 \tag{24}$$

The scope is to orthogonalize the set of structural descriptors to obtain the orthogonalized set of descriptors $\Omega(X_1)$, $\Omega(X_2)$, and $\Omega(X_3)$. For the set of three intercorrelated structural descriptors orthogonalized in the order X_1, X_2, and X_3, the construction of the orthogonal descriptors consists of the following steps:

1. The first orthogonal descriptor $\Omega(X_1)$ is identical with the original descriptor X_1:

$$\Omega(X_1) = X_1 \tag{25}$$

2. The linear regression equation between descriptor X_2 and orthogonal descriptor $\Omega(X_1)$ is computed:

$$X_2 = A_{2,1} + B_{2,1}\Omega(X_1) \tag{26}$$

The second orthogonal descriptor $\Omega(X_2)$ is the residual of the above equation, i.e. the difference between the real value of X_2 and that computed with Eq. (26):

$$\Omega(X_2) = X_2 - A_{2,1} - B_{2,1}\Omega(X_1) \tag{27}$$

3. The orthogonalization of the third descriptor begins with the computation of the linear regression equation between descriptor X_3 and orthogonal descriptor $\Omega(X_1)$:

$$X_3 = A_{3,1} + B_{3,1}\Omega(X_1) \tag{28}$$

The residual of Eq. (28) gives $\Omega(X_{3,1})$, the part of X_3 that is orthogonal to $\Omega(X_1)$:

$$\Omega(X_{3,1}) = X_3 - A_{3,1} - B_{3,1}\Omega(X_1) \tag{29}$$

The vector $\Omega(X_{3,1})$ is then orthogonalized against $\Omega(X_2)$ by computing the linear regression equation between these two descriptors:

$$\Omega(X_{3,1}) = A_{3,2} + B_{3,2}\Omega(X_2) \tag{30}$$

Finally, the third orthogonal descriptor $\Omega(X_3)$ is the residual of Eq. (30):

$$\Omega(X_3) = \Omega(X_{3,1}) - A_{3,2} - B_{3,2}\Omega(X_2) \tag{31}$$

The structure–property model from Eq. (24) is computed with the orthogonal descriptors $\Omega(X_1)$, $\Omega(X_2)$, and $\Omega(X_3)$:

$$Y = b_0 + b_1\Omega(X_1) + b_2\Omega(X_2) + b_3\Omega(X_3) \tag{32}$$

Regardless of the sequence of orthogonalization, Eqs. (24) and (32) have identical statistical indices (correlation coefficient r, standard deviation s, and Fisher test F). In practice, the orthogonalization of descriptors can be used to simplify QSAR models that

Table 3
Intercorrelation matrix of the structural descriptors from Eq. (17) and correlation coefficient between each descriptor and the experimental activity. The intercorrelation coefficients larger than 0.80 are highlighted in bold

		1	2	3	4	5	6
${}^0\chi^v$	**1**	1.000	**0.811**	**0.969**	0.736	**0.904**	0.481
MaxSp(RD,*E*)	**2**	**0.811**	1.000	0.772	**0.986**	0.777	0.646
^{0}Chi(VS(RD$_p$),*E*)	**3**	**0.969**	0.772	1.000	0.695	**0.948**	0.324
Wi(RD$_p^3$,*E*)	**4**	0.736	**0.986**	0.695	1.000	0.699	0.602
^{1}Chi(VS(RD$_p$),*R*)	**5**	**0.904**	0.777	**0.948**	0.699	1.000	0.330
log(1/*K*)	**6**	0.481	0.646	0.324	0.602	0.330	1.000

contain many intercorrelated descriptors, by removing the variables with a small contribution.

The QSAR model from Eq. (17) was orthogonalized considering the descriptors in the order from Table 3:

$$\begin{aligned}\log(1/K) = {} & 5.377785 + 0.142529\Omega({}^0\chi^v) \\ & + 0.773778\Omega(\mathbf{MaxSp}(\mathbf{RD}, E)) \\ & - 1.021522\Omega({}^0\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_p), E)) \\ & - 0.001715\Omega\left(\mathbf{Wi}\left(\mathbf{RD}_p^3, E\right)\right) \\ & - 0.174682\Omega({}^1\mathbf{Chi}(\mathbf{VS}(\mathbf{RD}_p), R)) \qquad (33)\end{aligned}$$

The above MLR equation obtained after the orthogonalization has the same statistical indices as Eq. (17), but all descriptors are orthogonal; in this way, the statistical interpretation of the QSAR model is simplified.

4. Concluding remarks

Molecular graph descriptors represent valuable structural descriptors extensively used to develop QSPR and QSAR models. The interest of developing new topological indices was stimulated in recent years by their use in virtual screening of combinatorial libraries, database mining, similarity and diversity assessment. Recently we have defined several molecular graph operators as a convenient and efficient method to compute from a unique mathematical equation a family of related molecular graph descriptors; such an operator can be applied to all molecular matrices and sets of parameters for atoms and bonds. In the present investigation we have developed QSAR models for the dihydrofolate reductase inhibition with diaminopyrimidines, using as structural descriptors a set of 271 topological indices.

All QSAR models were generated with the genetic function approximation method; the genetic algorithm proved to be an efficient method for feature selection in a highly dimensional space of structural descriptors. Another advantage of the genetic algorithm is the generation of several good QSAR models; this set of equations can be studied for information on the use of structural descriptors, graph operators, molecular matrices, or weighting schemes. As we found in the present study, the QSAR equation that has the best calibration statistical indices may not have the best prediction results. The investigation of the top equations offers the possibility to identify the QSAR model that gives the best predictions. Using a group of five molecular graph descriptors, we have obtained a QSAR model that gives better calibration and prediction results than other models developed with substituent constants or neural networks.

The topological indices derived from the new graph operators proved to be efficient measures of the molecular structure, with the important advantage that they can be easily computed for any organic compound. As we have already mentioned, topological indices complement (and do not substitute) the information encoded into other classes of structural descriptors, such as constitutional, geometrical, electrostatic, quantum, and grid (field) descriptors. A number of successful 3D QSAR approaches have been developed in recent years [52,53], many of them using grid (field) descriptors together with a pharmacophore-dependent alignment of the molecules. However, several studies [53–55] pointed out that

QSAR models obtained with structural descriptors derived from the molecular graph provide statistical indices as good as those generated with 3D QSAR models. The successful development of alternative QSAR models (i.e. with different classes of structural descriptors and with different statistical equations) confirms the existence of a structure–activity relationship in a data set. The approach proposed in this paper is an effective and simple complement to the sophisticated 3D QSAR models.

Acknowledgements

Ovidiu Ivanciuc thanks the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche of France for a PAST grant. Teodora and Ovidiu Ivanciuc acknowledge the kind hospitality of the LARTIC group during their stay in Nice. We acknowledge the partial financial support of this research by the Romanian Ministry of National Education under Grant 7001 T34.

References

[1] J.M. Sutter, S.L. Dixon, P.C. Jurs, J. Chem. Inf. Comput. Sci. 35 (1995) 77.
[2] B.E. Mitchell, P.C. Jurs, J. Chem. Inf. Comput. Sci. 38 (1998) 489.
[3] B.E. Mitchell, P.C. Jurs, J. Chem. Inf. Comput. Sci. 36 (1996) 58.
[4] D.L. Clouser, P.C. Jurs, J. Chem. Inf. Comput. Sci. 36 (1996) 168.
[5] O. Mekenyan, S. Karabunarliev, D. Bonchev, Comput. Chem. 14 (1990) 193.
[6] O.G. Mekenyan, S.H. Karabunarliev, J.M. Ivanov, D.N. Dimitrov, Comput. Chem. 18 (1994) 173.
[7] R. Murugan, M.P. Grendze, J.E. Toomey, A.R. Katritzky, M. Karelson, V.S. Lobanov, P. Rachwal, CHEMTECH 24 (1994) 17.
[8] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. (1995) 279.
[9] M. Karelson, V.S. Lobanov, A.R. Katritzky, Chem. Rev. 96 (1996) 1027.
[10] A.R. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Karelson, J. Chem. Inf. Comput. Sci. 38 (1998) 720.
[11] A.R. Katritzky, V.S. Lobanov, M. Karelson, J. Chem. Inf. Comput. Sci. 38 (1998) 28.
[12] O. Ivanciuc, T. Ivanciuc, A.T. Balaban, Tetrahedron 54 (1998) 9129.
[13] O. Ivanciuc, T. Ivanciuc, P.A. Filip, D. Cabrol-Bass, J. Chem. Inf. Comput. Sci. 39 (1999) 515–524.
[14] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.
[15] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Analysis, Research Studies Press, Letchworth, 1986.
[16] O. Ivanciuc, A.T. Balaban, in: P.v.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III, P.R. Schreiner (Eds.), The Encyclopedia of Computational Chemistry, Wiley, Chichester, 1998, p. 1169.
[17] O. Ivanciuc, A.T. Balaban, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, Amsterdam, 1999, p. 59.
[18] O. Ivanciuc, T. Ivanciuc, A.T. Balaban, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, Amsterdam, 1999, p. 169.
[19] O. Ivanciuc, T. Ivanciuc, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, Amsterdam, 1999, p. 221.
[20] O. Ivanciuc, T. Ivanciuc, M.V. Diudea, SAR QSAR Environ. Res. 7 (1997) 63.
[21] M.V. Diudea, Croat. Chem. Acta 72 (1999) 835.
[22] O. Ivanciuc, T. Ivanciuc, D.J. Klein, W.A. Seitz, A.T. Balaban, SAR QSAR Environ. Res. 11 (2001) 419.
[23] O. Ivanciuc, T. Ivanciuc, D.J. Klein, W.A. Seitz, A.T. Balaban, J. Chem. Inf. Comput. Sci. 41 (2001) 536.
[24] O. Ivanciuc, J. Chem. Inf. Comput. Sci. 40 (2000) 1412.
[25] O. Ivanciuc, ACH—Model. Chem. 137 (2000) 607.
[26] T.S. Balaban, P.A. Filip, O. Ivanciuc, J. Math. Chem. 11 (1992) 79.
[27] D. Plavšić, S. Nikolić, N. Trinajstić, Z. Mihalić, J. Math. Chem. 12 (1993) 235.
[28] O. Ivanciuc, T.S. Balaban, A.T. Balaban, J. Math. Chem. 12 (1993) 309.
[29] M.V. Diudea, O. Ivanciuc, S. Nikolić, N. Trinajstić, MATCH (Commun. Math. Comput. Chem.) 35 (1997) 41.
[30] M.V. Diudea, J. Chem. Inf. Comput. Sci. 36 (1996) 535.
[31] M.V. Diudea, J. Chem. Inf. Comput. Sci. 36 (1996) 833.
[32] C. Hansch, R.-L. Li, J.M. Blaney, R. Langridge, J. Med. Chem. 25 (1982) 777.
[33] C.D. Selassie, R.-L. Li, M. Poe, C. Hansch, J. Med. Chem. 34 (1991) 46.
[34] C. Hansch, T. Fujita, J. Am. Chem. Soc. 86 (1964) 1616.
[35] S.-S. So, W.G. Richards, J. Med. Chem. 35 (1992) 3201.
[36] M. Barysz, G. Jashari, R.S. Lall, V.K. Srivastava, N. Trinajstić, in: R.B. King (Ed.), Chemical Applications of Topology and Graph Theory, Elsevier, Amsterdam, 1983, p. 222.
[37] A.T. Balaban, MATCH (Commun. Math. Chem.) 21 (1986) 115.
[38] O. Ivanciuc, T. Ivanciuc, A.T. Balaban, J. Chem. Inf. Comput. Sci. 38 (1998) 395.
[39] O. Ivanciuc, Rev. Roum. Chim. 45 (2000) 289.
[40] O. Ivanciuc, S.L. Taraviras, D. Cabrol-Bass, J. Chem. Inf. Comput. Sci. 40 (2000) 126.
[41] S. Taraviras, O. Ivanciuc, D. Cabrol-Bass, J. Chem. Inf. Comput. Sci. 40 (2000) 1128.

[42] O. Ivanciuc, Rev. Roum. Chim. 44 (1999) 519.
[43] O. Ivanciuc, Rev. Roum. Chim. 45 (2000) 587.
[44] R. Leardi, R. Boggia, M. Terrile, J. Chemom. 6 (1992) 267.
[45] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci. 37 (1997) 306.
[46] H. Kubinyi, Quant. Struct.-Act. Relat. 13 (1994) 285.
[47] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854.
[48] J.G. Topliss, R.J. Costello, J. Med. Chem. 15 (1972) 1066.
[49] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), QSAR: Chemometric Methods in Molecular Design, Methods and Principles in Medicinal Chemistry, vol. 2, Verlag Chemie, Weinheim, Germany, 1995, p. 309.
[50] M. Randić, New J. Chem. 15 (1991) 517.
[51] O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, A.T. Balaban, J. Chem. Inf. Comput. Sci. 40 (2000) 732.
[52] O. Ivanciuc, in: M.V. Diudea (Ed.), QSPR/QSAR Studies by Molecular Descriptors, Nova Science, Huntington, NY, 2001, p. 233.
[53] H. Kubinyi, Drug Discovery Today 2 (1997) 538.
[54] B. Hoffman, S.J. Cho, W. Zheng, S. Wyrick, D.E. Nichols, R.B. Mailman, A. Tropsha, J. Med. Chem. 42 (1999) 3217.
[55] B.T. Hoffman, T. Kopajtic, J.L. Katz, A.H. Newman, J. Med. Chem. 43 (2000) 4151.