

## Application to QSAR studies of 2-furylethylene derivatives

Cristina D. Moldovan · Adina Costescu ·  
Gabriel Katona · Mircea V. Diudea

Published online: 6 August 2008  
© Springer Science+Business Media, LLC 2008

**Abstract** A quantitative structure–activity relationship (QSAR) is a mathematical model that relates a molecular structure to a physicochemical property or a biological activity. The log P of a set of 38 of 2-furylethylenes, biologically active substances exhibiting a broad spectrum of antimicrobial, antiparasitic, cytotoxic, carcinogenic and mutagenic activities, was modeled by using topological indices provided by TOPOCLUJ and DRAGON software packages. The models derived showed good stability and predictability (as given by the leave-one-out LOO cross-validation data). The results are compared with those reported in literature, obtained by different methodology.

**Keywords** QSAR · 2-furylethylene · log P · Similarity · TOPOCLUJ · DRAGON

### 1 Introduction

Quantitative structure–activity relationship (QSAR) techniques became indispensable in all aspects of research regarding the molecular interpretation of biological properties [1]. It is obvious that physical, chemical, or biological properties of a compound depend on the three-dimensional (3D) arrangement of atoms in the molecule. The ability to produce quantitative correlation between 3D structure of molecules and their biological activity is important in deciding upon the synthetic ways of bioactive chemicals [2].

A QSAR is a mathematical model that relates, in a quantitative manner, the chemical structure and a physico-chemical property or a biological effect. Under the future REACH (Registration, Evaluation and Authorization of CHEMicals) system, proposed by the Commission's White Paper [3] on a Future Chemicals Policy, it is anti-

---

C. D. Moldovan · A. Costescu · G. Katona · M. V. Diudea (✉)  
Faculty of Chemistry and Chemical Engineering, Babes-Bolyai University, Cluj-Napoca, Romania  
e-mail: diudea@chem.ubbcluj.ro

pated that QSARs will be used more extensively, in the interests of time- and cost-effectiveness and animal welfare. In particular, QSARs are likely to play an important role in the assessment of chemicals produced or imported in quantities between 1 and 10 tons, for which minimal animal testing is foreseen by the White Paper. In principle, QSARs could be used for a number of purposes in the implementation of legislation on chemical substances and products. It was considered necessary to develop a framework for the independent development, validation and dissemination of QSARs.

## 2 Materials and methods

### 2.1 Data sets

We studied a set of 38 2-furylethylenes, with the activity taken from the publications of Miguel Angel Cabrera Pérez [4] and Yovani Marrero Ponce [5]. 2-Furylethylenes are biologically active substances exhibiting a broad spectrum of antimicrobial, antiparasitic, cytotoxic, but in some case also carcinogenic and mutagenic activities [6]. The interest in the research of 2-furylethylenes has increased in recent years as a consequence of the discoveries of potent microcidal compounds having this chemical structure [7, 8].

2-Furylethylenes are derivatives of the ethene where a furan ring is attached to one of its carbon atoms. The exocyclic double bond is frequently substituted at position  $\beta$  by different atomic groups, which modify in some extent the physicochemical and antimicrobial properties of such compounds.

Several physicochemical properties of drug molecules such as lipophilicity have been used to predict passive absorption in vivo. One of them, the partition coefficient, is used to characterize the lipophilicity of drugs [4].

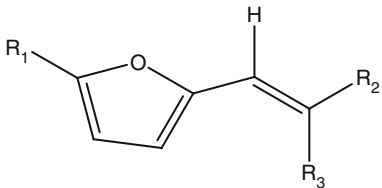
In the present work 38 compounds are taken in the modeling study of partition coefficients ( $\log P$ ). The molecular structures of all these compounds are illustrated in Table 1. This set was many times utilized for the evaluation of the performances of new QSAR methods [9]. The values  $\log P$  (partition coefficient in water) of these compounds were determined experimentally and reported in literature [4].

The values of *n*-octanol/water partition coefficient for four derivatives of 2-furylethylene 35, 36, 37 and 38 were taken from the article by Miguel Angel Cabrera Pérez [4]. The steps followed in deriving the model are given in Fig. 1.

### 2.2 Computation of molecular descriptors

Topological molecular descriptors are used in QSAR studies because of their accessibility, being easily computed by available software programs. The set of molecular descriptors used in this study is calculated by **TOPOCLUJ** [10] and **DRAGON** [11] software packages. The structures were optimized by using the semi-empirical PM3 Hamiltonian, available in **HyperChem** [12].

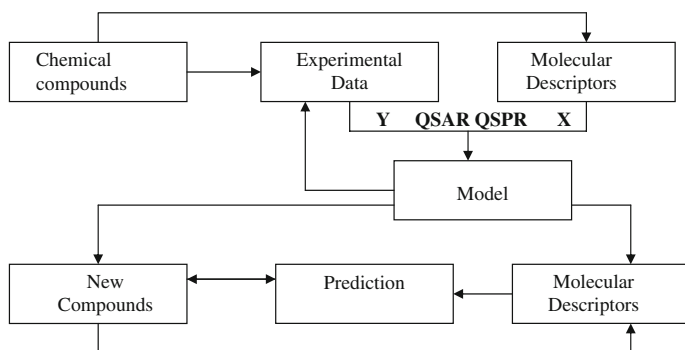
The DRAGON software provided 1600 molecular descriptors (denoted here as *DI*). The most relevant descriptors proved to be *MW* (molecular weight), *X4v* (valence

**Table 1** Data set of 2-furylethylenes derivatives


Structure	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	log P
1	H	NO <sub>2</sub>	COOCH <sub>3</sub>	1.879
2	CH <sub>3</sub>	NO <sub>2</sub>	COOCH <sub>3</sub>	2.439
3	Br	NO <sub>2</sub>	COOCH <sub>3</sub>	2.739
4	COOCH <sub>3</sub>	NO <sub>2</sub>	COOCH <sub>3</sub>	1.869
5	NO <sub>2</sub>	NO <sub>2</sub>	COOCH <sub>3</sub>	1.599
6	NO <sub>2</sub>	COOC <sub>2</sub> H <sub>5</sub>	COOC <sub>2</sub> H <sub>5</sub>	2.504
7	NO <sub>2</sub>	H	NO <sub>2</sub>	1.303
8	H	H	NO <sub>2</sub>	1.583
9	NO <sub>2</sub>	H	CONHC <sub>2</sub> H <sub>5</sub>	1.386
10	NO <sub>2</sub>	H	CONH(CH <sub>2</sub> ) <sub>2</sub> CH <sub>3</sub>	1.86
11	NO <sub>2</sub>	H	CONHCH(CH <sub>3</sub> ) <sub>2</sub>	1.803
12	NO <sub>2</sub>	H	CONH(CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>	2.356
13	NO <sub>2</sub>	H	CONHCH <sub>2</sub> CH(CH <sub>3</sub> ) <sub>2</sub>	2.225
14	NO <sub>2</sub>	H	CONHCH(CH <sub>3</sub> )C <sub>2</sub> H <sub>5</sub>	2.284
15	NO <sub>2</sub>	H	CONHC(CH <sub>3</sub> ) <sub>3</sub>	2.333
16	NO <sub>2</sub>	H	CONHCH <sub>2</sub> C(CH <sub>3</sub> ) <sub>3</sub>	2.605
17	NO <sub>2</sub>	H	COOCH <sub>3</sub>	1.652
18	NO <sub>2</sub>	H	COOC <sub>2</sub> H <sub>5</sub>	2.098
19	NO <sub>2</sub>	H	COO(CH <sub>2</sub> ) <sub>2</sub> CH <sub>3</sub>	2.673
20	NO <sub>2</sub>	H	COOCH(CH <sub>3</sub> ) <sub>2</sub>	2.641
21	NO <sub>2</sub>	H	COO(CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>	2.827
22	NO <sub>2</sub>	H	COOCH <sub>2</sub> CH(CH <sub>3</sub> ) <sub>2</sub>	3.135
23	NO <sub>2</sub>	H	COOCH(CH <sub>3</sub> )C <sub>2</sub> H <sub>5</sub>	3.091
24	NO <sub>2</sub>	H	COOC(CH <sub>3</sub> ) <sub>3</sub>	3.06
25	NO <sub>2</sub>	H	COO(CH <sub>2</sub> ) <sub>4</sub> CH <sub>3</sub>	3.404
26	NO <sub>2</sub>	H	Br	2.447
27	NO <sub>2</sub>	H	CN	1.05
28	NO <sub>2</sub>	H	OCH <sub>3</sub>	1.591
29	NO <sub>2</sub>	H	H	1.611
30	NO <sub>2</sub>	CN	COOCH <sub>3</sub>	1.488
31	I	NO <sub>2</sub>	COOCH <sub>3</sub>	2.999
32	NO <sub>2</sub>	H	CONH <sub>2</sub>	0.649
33	NO <sub>2</sub>	H	CONHCH <sub>3</sub>	0.984
34	NO <sub>2</sub>	H	CON(CH <sub>3</sub> ) <sub>2</sub>	0.819
35	Br	NO <sub>2</sub>	Br	2.820
36	Br	NO <sub>2</sub>	CH <sub>3</sub>	2.730
37	H	NO <sub>2</sub>	H	1.290
38	H	NO <sub>2</sub>	CH <sub>3</sub>	1.940

connectivity index  $\chi$ -4), and  $G(N \dots O)$  (the geometric distance between  $N \dots O$ ). Geometric descriptors indicate the size of the molecule and are derived from the three-dimensional coordinates.

Topological indices were calculated with TOPOCLUJ software package, developed in our laboratory. A single number, representing a chemical structure, in



**Fig. 1** Schematic representation of the steps followed in building the model

graph-theoretical terms, is called a topological descriptor. Being a structural invariant, it does not depend on the labeling or the pictorial representation of the graph [13–15].

The most relevant descriptors computed with TOPOCLUJ are:  $VAA1$ ,  $VAD1$ ,  $CS[LM[Electronegativity]]$ ,  $IE[C_f \text{ Max}[Density]]$ ,  $CS[Sh[W4[Charge\_Adjacency]]]$ .

### 2.3 Methods

Principal components analysis PCA is a powerful statistical technique useful in data reduction. The regression equations were derived by STATISTICA 6.0 software package [16].

Once the desired set of descriptors had been calculated and stored, the process of descriptor analysis is started. It is important to examine the pool of descriptors in an objective manner and to remove from further consideration those descriptors which are redundant or do not contain enough discriminatory information to be of any significant value. All descriptors containing identical values for 90% or more of the compounds in a given data set, including both zero and non-zero values, were removed.

All possible pairs of remaining descriptors were examined to identify those pairs which are highly correlated. As a rule of thumb, a critical value of 0.950 for the correlation coefficient ( $r$ ) was used. If two descriptors were correlated at or above the critical value, one descriptor was discarded. The decision of which one to retain was based on the possible physical interpretation of the descriptor, ease of calculation, or usefulness in the past studies. The result of this analysis is a reduced pool of information-rich descriptors which can then be screened by using multiple linear regression analysis. After all of these procedures we reduced the searching space from 1600 to 548 descriptors.

Linear regression models were developed by multiple regressions with stepwise addition of descriptors, where the inclusion of a given term is based on the F statistic values. A deletion process is then used, where each independent variable is held out in turn, and a model is developed by using the remaining pool of descriptors. Then all pairs and triplets are held out, and the process is repeated. This series of steps has the effect of finding the best regression equations.

The best found descriptors and models were also examined for robustness and predictive ability through both internal and external validation methods. These evaluations are included in the discussion below.

Molecular similarity has been considered here in two terms: (i) topological similarity defined by connectivity descriptors and (ii) geometrical similarity, when geometrical aspects of the molecular structure are taken into account.

Similarity searching in databases of 3D chemical structures is widely used for virtual screening and lead discovery. A similarity measure, that quantifies the degree of structural resemblance between the target structure and each of the database structure, is based on molecular descriptors encoding the molecular structure, with similarity between pairs of such representations being computed using the *Tanimoto* coefficient.

Another topological similarity measure of increasing interest (although more computationally demanding) is detection of 3D maximum common subgraphs (MCS) [17, 18]. The values of the similarity coefficient  $sim^{cv}(G_1, G_2)$  of two compared molecular graphs range between 0 and 1, according to the equality:

$$sim^{cv}(G_1, G_2) = \frac{(V(G_{12}) + E(G_{12}))^2}{(|V(G_1)| + |E(G_1)|) \cdot (|V(G_2)| + |E(G_2)|)} \quad (1)$$

where  $G_{12}$  is the maximum common subgraph among of two graphs  $G_1$  and  $G_2$ , with the vertex set  $V(G)$  and edge set  $E(G)$ .

The binding affinity of the ligand to the receptor site, which usually expresses the biological activity, is related to a single geometrical configuration of the ligand. This method takes a full account of the conformational flexibility, by the imposed upper bond conditions.

The training set was tested for similarity against the molecules 35 to 38, belonging to the prediction set. Those structures with similarity coefficient higher than 0.70, at least for one of these four structures, have been included in the training set (Table 2).

The most significant molecular descriptors computed with DRAGON software and the partition coefficient *n*-octanol/water (log P), are given in Tables 3 and 4.

To find the best correlations, the four subsets were submitted to STATISTICA software and the results are shown in Eqs. 2 and 3. The best QSAR equation, obtained in MLR was used for prediction in the testing set. The quality of the model is expressed by the squared regression coefficient ( $R^2$ ), Fisher-ratio, standard error of estimate ( $s$ ) and the *leave-one-out* (LOO) ( $Q^2$ ) in the cross-validation procedure.

### 3 Results and discussion

The regression equations are:

*Bivariate regression*

$$\log P = 0.086403 + 3.442395 \cdot DI_1 + 2.163165 \cdot DI_3 \quad (2)$$

$$n = 12, \quad R^2 = 0.984, \quad s = 0.092, \quad F = 282.346$$

*Multivariate regression*

**Table 2** Similarity data for 2-furylethylene derivatives against the selected leaders and their log P (partition coefficient *n*-octanol/water)

Structure	#35	#36	#37	#38	log P
1	0.595	0.720	0.714	0.786	1.879
2	0.556	0.672	0.667	0.733	2.439
3	0.672	0.8	0.667	0.733	2.739
4	0.463	0.560	0.556	0.611	1.869
5	0.490	0.593	0.588	0.647	1.599
6	0.459	0.482	0.451	0.501	2.504
7	0.641	0.641	0.769	0.699	1.303
8	0.833	0.833	1	0.909	1.583
9	0.501	0.613	0.602	0.668	1.386
10	0.470	0.574	0.564	0.626	1.86
11	0.470	0.574	0.564	0.626	1.803
12	0.490	0.540	0.531	0.590	2.356
13	0.490	0.540	0.531	0.590	2.225
14	0.490	0.540	0.531	0.590	2.284
15	0.490	0.540	0.531	0.590	2.333
16	0.463	0.510	0.502	0.557	2.605
17	0.537	0.656	0.645	0.716	1.652
18	0.501	0.613	0.602	0.668	2.098
19	0.470	0.574	0.564	0.626	2.673
20	0.470	0.574	0.564	0.626	2.641
21	0.490	0.540	0.531	0.589	2.827
22	0.490	0.540	0.531	0.589	3.135
23	0.490	0.540	0.531	0.589	3.091
24	0.490	0.540	0.531	0.589	3.06
25	0.463	0.510	0.501	0.557	3.404
26	0.835	0.684	0.820	0.746	2.447
27	0.627	0.766	0.752	0.835	1.05
28	0.627	0.627	0.752	0.684	1.591
29	0.752	0.752	0.903	0.820	1.611
30	0.470	0.574	0.564	0.626	1.488
31	0.556	0.672	0.667	0.733	2.999
32	0.578	0.707	0.694	0.771	0.649
33	0.537	0.656	0.645	0.716	0.984
34	0.501	0.613	0.602	0.668	0.819
35	1	0.840	0.833	0.758	2.820
36	0.840	1	0.833	0.917	2.730
37	0.833	0.833	1	0.909	1.290
38	0.758	0.917	0.909	1	1.940

$$\log P = 0.391261 - 0.003500 \cdot DI_1 + 3.287415 \cdot DI_2 - 0.043587 \cdot DI_3 \quad (3)$$

$$n = 12, \quad R^2 = 0.987, \quad s = 0.079, \quad F = 195.31$$

The equations show a good coefficient of correlation:  $R^2 = 0.984$  in bivariate regression (Eq. 2, Fig. 2a), and  $R^2 = 0.987$  (Eq. 3, Fig. 2b) in multivariate regression with 3 descriptors. This last correlation is slightly better than those previously reported (the best model, with 7 descriptors, by Y. M. Ponce et al. shows  $R^2 = 0.968$ ;  $Q^2 = 0.938$ ) [5].

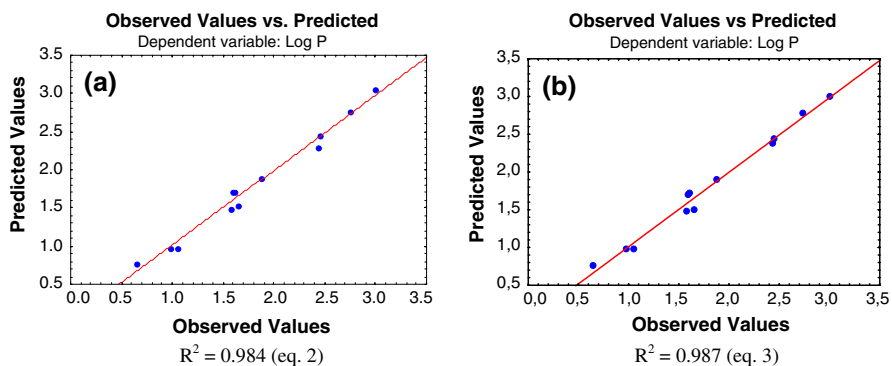
**Table 3** Topological descriptors<sup>a</sup> by DRAGON software and log P (partition coefficient *n*-octanol/water) for the training set

Structure	$DI_1$	$DI_2$	$DI_3$	log P
1	197.16	0.799	9.619	1.879
2	211.19	0.956	9.613	2.439
3	276.05	1.15	9.626	2.739
8	139.12	0.52	3.096	1.583
17	197.16	0.748	14.97	1.652
26	218.01	0.853	0	2.447
27	164.13	0.649	22.162	1.05
28	169.15	0.665	6.598	1.591
29	139.12	0.55	0	1.611
31	323.05	1.265	9.641	2.999
32	182.15	0.694	29.088	0.649
33	196.18	0.771	29.063	0.984

<sup>a</sup> $DI_1$ ,  $DI_2$ ,  $DI_3$  are molecular descriptors computed by DRAGON software:  $MW$  is the molecular weight,  $X4v$  is the valence connectivity index chi-4,  $G(N...O)$  counts the geometrical distances between  $N...O$  atoms

**Table 4** Topological descriptors by DRAGON and log P for the prediction/testing set

Structure	$DI_1$	$DI_2$	$DI_3$	log P
35	296.9	1.121	4.405	2.49
36	232.04	0.984	4.516	2.37
37	139.12	0.52	4.386	1.56
38	153.15	0.632	4.517	1.92

**Fig. 2** Plots of observed vs. predicted property for Eqs. 2 and 3 in the training set.**Table 5** Experimental and predicted log P values for the testing set

	log P (exp)	log P(calc) (Eq. 2)	log P(calc) (Eq. 3)
35	2.820	2.924846	2.845186
36	2.730	2.580922	2.617008
37	1.290	1.438395	1.422570
38	1.940	1.709787	1.735940
$Q^2$		0.933	0.959
CV%		10,395	8,196

**Table 6** Topological descriptors<sup>a</sup> by TOPOCLUJ and log P (partition coefficient *n*-octanol/water) for the training set

Structure	$TI_1$	$TI_2$	$TI_3$	$TI_4$	$TI_5$	log P
1	570	34	220	8.5	46	1.879
2	650	36	290	9	52	2.439
3	650	37	300	9	52	2.739
8	290	21	62	6.1	27	1.583
17	570	33	220	8.5	51	1.652
26	350	26	130	6.6	31	2.447
27	410	29	120	7.3	38	1.05
28	420	28	140	7.3	38	1.591
29	290	23	62	6.1	25	1.611
31	650	27	350	9	52	2.999
32	490	23	120	7.7	44	0.649
33	560	25	220	8.5	51	0.984

<sup>a</sup> $TI_1, TI_2, TI_3, TI_4, TI_5$  are molecular descriptors  $CS[LM[Electronegativity]]$ ,  $CS[Sh [W_4[Charge\_Adjacency]]]$ ,  $IE[CfMax[Density]]$ ,  $VAA1$  and  $VAD1$  computed with TOPOCLUJ software

**Table 7** Topological descriptors by TOPOCLUJ and log P for the prediction set

Nr.	$TI_1$	$TI_2$	$TI_3$	$TI_4$	$TI_5$	log P
35	410	22	190	7.1	36	2.82
36	410	19	170	7.1	36	2.73
37	290	11	62	6.1	27	1.29
38	340	17	120	6.6	31	1.94

Every QSAR model must be validated on an external prediction set. In this case the prediction set consists of #35 to #38 2-furylethylene derivatives, for which the predicted log P values are listed in Table 5.

Predictability and stability of the models obtained using the above molecular descriptors are determined here by means of LOO cross-validation. The model showing the best cross-validation regression coefficient of  $Q^2 = 0.959$  (Table 5).

Following the same steps we derived the models with the descriptors provided by TOPOCLUJ software. The results are as follows.

The most significant computed descriptors by TOPOCLUJ are listed in Table 6 along with the corresponding log P values, for the training set while in Table 7 the same data are given for the testing set.

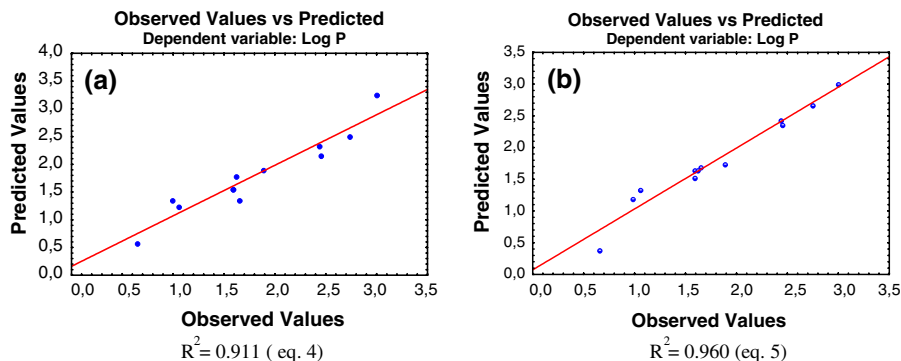
The best equations of the model, used for prediction are:

*Bivariate regression*

$$\log P = 3.562116 + 0.015455 \cdot TI_3 - 0.109763 \cdot TI_5 \quad (4)$$

$$n = 12, \quad R^2 = 0.911, \quad s = 0.520, \quad F = 46.196$$





**Fig. 3** The plot of observed vs. predicted log P, by Eqs. 4 and 5.

**Table 8** Experimental and predicted log P values for the prediction set

Nr.	log P(exp)	log P(calc) Eq. 4	log P(calc) Eq. 5
35	2.820	2.547	2.502
36	2.730	2.238	1.951
37	1.290	1.556	0.925
38	1.940	2.014	1.632
Q <sup>2</sup>		0.924	0.904
CV%		11.094	12.492

### Multivariate regression

$$\log P = 7.99004 + 0.05992 \cdot TI_2 + 0.01853 \cdot TI_3 - 1.45456 \cdot TI_4 \quad (5)$$

$$n = 12, \quad R^2 = 0.960, \quad s = 0.233, \quad F = 64.311$$

The above equations show the best correlation coefficient  $R^2 = 0.960$  (Eq. 5) in three variables. Figure 3 illustrates the plot of predicted vs. observed values of log P, by Eqs. 4 and 5.

The QSAR models were next validated on the external prediction set. This set consists of #35 to 38 of 2-furylethylenes (Table 8).

## 4 Conclusions

The partition coefficient *n*-octanol/water (log P) plays an important role in the understanding of the biological behavior of chemicals, particularly 2-furylethylene derivatives. For the modeling and prediction of this molecular property a clustering method based on similarity was proposed. It is used to reduce the number of descriptors in the regression equation. The models obtained here are compared with those reported in the literature and have a good ability of prediction.

## References

1. A.R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson, Perspective: structurally diverse quantitative structure–property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **40**, 1–18 (2000)
2. D.D. Robinson, P.J. Winn, P.D. Lyne, W.G. Richards, Self-organizing molecular field analysis: a tool for structure–activity studies. *J. Med. Chem.* **42**, 573–583 (1999)
3. <http://europa.eu.int>
4. M.A. Cabrea Pérez et al., Experimental and theoretical determination of physicochemical properties in a novel family of microcidal compounds. *Eur. Bull. Drug Res.* **9**, 1 (2001)
5. Y. Marrero Ponce et al., Atom, atom-type, and total linear indices of the molecular pseudograph's atom adjacency matrix: application to QSPR/QSAR studies of organic compounds. *Molecules* **9**, 1100–1123 (2004)
6. J.Ch. Dore, C. Viel, Antitumoral chemotherapy. X. Cytotoxic and antitumoral activity of  $\beta$ -nitrostyrenes and nitrovinyl derivatives. *Farmaco* **30**, 81–109 (1975)
7. N. Castañedo, R. Goizueta, J. Perez, J. Gonzalez, E. Silveira, M. Cuesta, A. Martinez, E. Lugo, E. Estrada, A. Carta, O. Navia, M. Delgado, Cuban Pat. 22446, 1994; Can. Pat. 2,147,594, 1999
8. J.M. Blondeau, N. Castañedo, O. Gonzalez, R. Medina, E. Silveira, In vitro evaluation of G-1: a novel antimicrobial compound. *Antimicrob. Agents Chemother.* **11**, 1663–1669 (1999)
9. S. Balaz, E. Sturdik, M. Rosenberg, J. Augustin, B. Skara, Kinetics of drug activities as influenced by their physicochemical properties: antibacterial effects of alkylating 2-furylethylenes. *J. Theor. Biol.* **131**, 115–134 (1988)
10. M.V. Diudea, O. Ursu, TOPOCLUJ (Copyright Babes-Bolyai Univ. Cluj, 2002)
11. *Dragon 5* software, <http://www.disat.unimib.it/chm/Dragon.htm>
12. HyperChem [TM], release 4.5 for SGI, © 1991–1995, Hypercube, Inc.
13. A.T. Balaban, I. Motoc, D. Bonchev, O. Mekenyan, Topological indices for structure–activity correlations. *Top. Curr. Chem.* **114**, 21–55 (1993)
14. D.H. Rouvray, The challenge of characterizing branching in molecular species. *Discr. Appl. Math.* **19**, 317–338 (1988)
15. M. Randić, Design of molecules with desired properties. A molecular similarity approach to property optimization, in *Concepts and Applications of Molecular Similarity*, Chap. 5, ed. by M.A. Johnson, G.M. Maggiora (John Wiley & Sons, Inc., 1990), pp. 77–145
16. StatSoft, Inc., STATISTICA (data analysis software system), version 6 (2001), [www.statsoft.com](http://www.statsoft.com)
17. J. Raymond, E. Gardiner, P. Willett, RASCAL: calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* **45**, 631–644 (2002)
18. J. Raymond, E. Gardiner, P. Willett, Heuristics for rapid similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.* **42**, 305–316 (2002)