

QSAR studies of imidazopyridine derivatives as Et-PKG inhibitors using the PSO-SVM approach

Zhengjun Cheng · Yuntao Zhang · Wenjun Zhang

Received: 20 May 2009 / Accepted: 17 September 2009 / Published online: 29 October 2009
© Birkhäuser Boston 2009

Abstract In this work, the multiple linear regression (MLR) and support vector machine (SVM) methods were applied for modeling and predicting the inhibitory activity of imidazopyridine derivatives with two-dimensional (2D) autocorrelation descriptors calculated from the molecular structure alone for the first time. We define the new objective function as fitness, and it can guide the particle swarm optimization (PSO) method to select important descriptors which are responsible for the inhibitory activity of these compounds. The square of the correlation coefficient ($R^2 = 0.897$), the square of correlation coefficients of the test set ($R_{\text{ext}}^2 = 0.660$), and the obtained statistical parameter of the calibrating set in the PSO-SVM model was 0.743, which demonstrated the reliability of the model. The PSO-SVM model is superior over the PSO-MLR method in the dataset with imidazopyridine derivatives as Et-PKG inhibitors. Our best quantitative structure–activity relationship model illustrates the importance of an adequate distribution of atomic properties represented in topological frames and reveals atomic masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities to be the most influential atomic properties in the structures of the imidazopyridine derivatives.

Keywords Quantitative structure–activity relationship · Imidazopyridine derivatives · Support vector machine · Two-dimensional (2D) autocorrelation descriptor · Particle swarm optimization

Introduction

Protozoan parasites of the subphylum Apicomplexa are significant threats to both human and animal health worldwide (Diaz *et al.*, 2006). Coccidiosis is caused by

Z. Cheng · Y. Zhang (✉) · W. Zhang
Institute of Applied Chemistry, China West Normal University, Nanchong Sichuan 637002, China
e-mail: nczyt@yahoo.com.cn

related apicomplexans; the *Eimeria* species of parasites, which invade intestinal epithelial cells, are the causative agents (Liang *et al.*, 2007). Some of the most significant *Eimeria* species in poultry are *E. tenella*, *E. acervulina*, *E. necatrix*, *E. brunetti*, and *E. maxima* (Scribner *et al.*, 2007). During acute infections, these parasites cause significant morbidity and mortality in broiler breeds of chicken. Hence, more than 35 billion chickens are raised annually worldwide and all major poultry operations use anticoccidial agents prophylactically (Scribner *et al.*, 2007). Nearly 30 years ago, polyether ionophore anticoccidials were discovered, which have been successfully used as prophylactic agents to combat coccidiosis (Biftu *et al.*, 2006). Not surprisingly, reports of development of resistance due to the extended and constant chemotherapeutic pressure exerted by this class of compounds are not uncommon (Gurnett *et al.*, 2002). Thus, the need to identify and develop new drugs for the control of coccidiosis is critically important. Recently, some groups (Scribner *et al.*, 2008) reported on some novel anticoccidial agents with potent *in vitro* and *in vivo* activity against *Eimeria* parasites. It was determined that inhibition of parasite growth by these compounds was due to inhibition of the parasite-specific cyclic GMP (cGMP)-dependent protein kinase (PKG). Cyclic GMP (cGMP)-dependent protein kinase (PKG) was biochemically purified from *E. tenella* (Donald *et al.*, 2002), and it has recently been identified as the likely molecular target of a newly disclosed broad-spectrum coccidiostat (Salowe *et al.*, 2002).

In previous studies, many groups were more interested in synthesis of novel compounds such as *E. tenella* cGMP-dependent protein kinase (Et-PKG) inhibitors. However, time and cost considerations make it unfeasible to carry out binding bioassays on every molecule. Alternatively, an untested molecule might be evaluated using the information from already obtained bioassays and the ability to build a quantitative structure–activity relationship (QSAR) model. The QSAR model seeks to discover and use mathematical relationships between chemical structure and biological activity. The approach does not depend on any experimental properties; it requires the molecular descriptors that can be calculated from the molecular structure alone. Once the structure of a compound is known, any molecular descriptor can be calculated, whether or not the compound is synthesized. When a reliable model is established, it can predict the inhibitory activity of compounds and evaluate the structural factors responsible for the inhibitory activity. So, in the last decade, computational-based rational design of drugs has increased rapidly. Most of these approaches are focused on using different kinds of molecular descriptors to encode chemical information (Doležal *et al.*, 2009; Li *et al.*, 2009; Mercader *et al.*, 2008). Topological index, geometrical descriptors, and other descriptors are included in QSAR studies. The two-dimensional (2D) autocorrelation descriptor has been successfully used to construct same kinds of QSAR model for modeling biological activities (Saíz-Urra *et al.*, 2007; Sharma *et al.*, 2008; Caballero *et al.*, 2008).

In the past several years, Ertepinar *et al.*, (1995) applied the QSAR model to study a set of benzimidazole and imidazopyridine derivatives that have previously been tested for their antibacterial activities against *Bacillus subtilis*. The results revealed that the activity contributions of benzimidazoles and imidazopyridines

against *B. subtilis* depend almost entirely on their relative lipophilic character as defined by their octanol/water partition coefficients, $\log P$. Later, Curtin *et al.*, (1998) discovered and evaluated a series of 3-acylindole imidazopyridine platelet-activating factor (PAF) antagonists. The PAF antagonist, an endogenous phospholipid inflammatory mediator, is a D-glycerol derivative bearing a phosphorylcholine at C3, an acetyl group at C2, and a long-chain alkyl ether moiety at C1. They applied structure–activity relationships (SARs) to study these compounds. The results indicated that modification of the indole and benzoyl spacer of the lead compound 6-(4-fluoro-phenyl)-3-(4-(2-methyl-imidazo[4,5-*c*]pyridin-3-ylmethyl)-benzoyl)-indole-1-carboxylic acid dimethylamide gave analogs that were more potent, longer-lived, and bioavailable. However, until recently, no authors have investigated a QSAR model for predicting inhibitory activity of imidazopyridine derivatives against Et-PKG.

In this context, we calculated 723 descriptors by E-Dragon 1.0 software and present a modeling approach based on SVM to study imidazopyridine derivatives as Et-PKG inhibitors. The particle swarm optimization (PSO) method was used to preselect the proper descriptors from whole descriptor sets. The main purpose of the work is to develop a QSAR model for predicting inhibitory activity of the 107 imidazopyridine derivatives and to better understand the structural features of these compounds and their relation with the inhibitory activity using the 2D autocorrelation descriptor. This study may help us to design new analogs with a better biological profile.

Materials and computational methods

Materials

Experimental data

The studied compounds are 107 imidazopyridine derivatives, which were taken from the literature (Liang *et al.*, 2007; Scribner *et al.*, 2007, 2008; Biftu *et al.*, 2006); their structures are shown in Table 1. These compounds were tested for in vitro efficacy using the Ten_K (*Tenella* kinase) assay. The Ten_K assay measures inhibition of Et-PKG activity, which is reported as the amount of compound required to inhibit activity by 50% (IC₅₀; nM). And the 50% inhibitory concentration (IC₅₀) values are listed in Table 1. If a compound was tested more than once and the corresponding two IC₅₀ values approximated equality, the average IC₅₀ value is listed in Table 1.

Descriptor calculation

To develop a QSAR model, molecular structures are often represented using molecular descriptors, which encode structural information. The calculation process of the descriptors involved the following steps: the structures of the compounds were drawn using Molinspiration WebME Editor (<http://www.molinspiration.com:9080/mi/webme.html>) and saved as C.SMI files. Then the C.SMI files were transferred

into the software E-Dragon 1.0 (<http://www.vcclab.org/lab/edragon/>) to calculate different dimension structural descriptors. The software E-Dragon 1.0 can calculate Edge adjacency indexes, GETAWAY, WHIM, 2D autocorrelation indexes, Burden eigenvalues, and 3D MoRSE descriptors. And these descriptors have been successfully used in various QSAR/QSPR studies (Khan *et al.*, 2009; Panek *et al.*, 2005; Davood *et al.*, 2009). In the prereduction step, the calculated descriptors were

Table 1 Structures of imidazopyridine derivatives (1–107) and experimental (Exp.) and predicted (Pred.) values of IC₅₀ by the PSO-SVM model related to the 2D autocorrelation descriptor

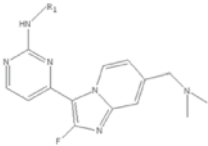
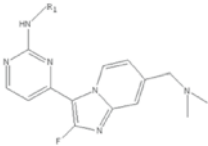
No.	Exp.	Pred.	No.	Exp.	Pred.		
Training set			Calibrating and test set				
							
R ₁			R ₁				
1	CH ₂ CH ₃	0.26	0.915	1	<i>c</i> -Bu	0.26	0.458
2	<i>n</i> -Propyl	0.13	0.408	2	CH ₂ CH ₂ CH ₂ CH ₂ CH ₃	0.19	0.195
3	CH ₂ CH ₂ CH ₂ OH	0.33	0.348	3	CH ₂ -4-Cl-Ph	0.27	0.339
4	CH ₂ CH ₂ CH ₂ CH ₃	2.14	0.363	4	C(=O)CH ₂ CH ₂ CH ₃	0.63	0.228
5	CH ₂ CH ₂ CH ₂ OCH ₃	0.39	0.667	5	C(=O)- <i>c</i> -Pentyl	0.26	0.535
6	<i>c</i> -Hexyl	0.21	0.135	6	C(=O)H	0.56	0.379
7	CH ₂ -2-F-Ph	0.087	0.182	7*	CH ₂ CH ₂ OCH ₃	0.36	0.996
8	CH ₂ -2-F-4-F-Ph	0.21	0.488	8*	CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₃	0.22	0.047
9	CH ₂ -3-F-4-F-Ph	0.06	-0.218	9*	CH ₂ -2-F-3-F-Ph	0.21	-0.317
10	CH ₂ -3-Cl-Ph	0.24	-0.037	10*	CH ₂ -3-F-5-F-Ph	0.21	1.541
11	Pyridin-2-yl	0.43	0.528	11*	C(=O)CH ₂ CH ₃	0.52	0.197
12	C(=O)- <i>c</i> -Pr	0.73	0.452	12*	C(=O)- <i>c</i> -Bu	0.35	0.497

Table 1 continued

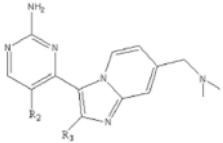
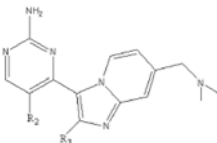
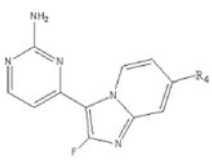
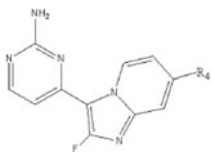
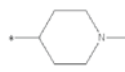
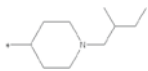
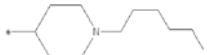
13	$C(=O)CH_2CH_2CH_2CH_3$	0.34	0.063						
14	$C(=O)NC(CH_3)_3$	4.39	4.112	R_2	R_3				
15	CO_2CH_3	0.63	0.353	13	H	2-F-4-F-Ph	0.1	0.768	
				14*	F	4-F-Ph	0.26	0.310	
	R_2	R_3							
16	H	4-F-Ph	0.11	0.622	R_4				
17	CN	4-F-Ph	1.6	1.322	15	$CH_2N(CH_3)CH_2CH_2OCH_3$	0.69	0.425	
18	Br	4-F-Ph	5.8	5.522	16	$CH_2N(CH_3)CH_2CN$	1.33	0.334	
19	H	3-Br-4-F-Ph	0.16	0.437	17	$N(CH_3)_2$	3	2.911	
20	H	3-CH ₃ -4-F-Ph	0.1	0.477	18	$CH_2CH_2N(CH_3)_2$	0.54	0.532	
21	H	4-Cl-Ph	0.41	0.688	19	$C(CH_3)_2NH_2$	0.20	-0.043	
22	H	2-F-4-F-6-F-Ph	0.52	0.798	20	$C(CH_3)_2CH_2N(CH_3)_2$	0.20	0.322	
				21		0.044	0.011		
	R_4			22		0.13	0.104		
23	$CH_2N(CH_3)CH_2C\equiv CH$	0.65	0.372	23		0.13	0.235		

Table 1 continued

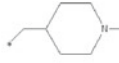
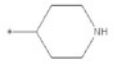

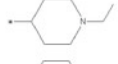
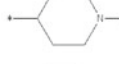
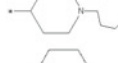
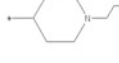


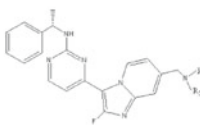
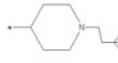

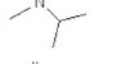
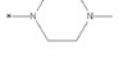
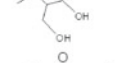
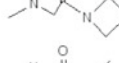
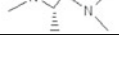
24	$\text{CH}_2\text{N}(\text{CH}_2\text{CH}_3)_2$	0.30	0.578	24	CH_2NHMe	0.15	0.380
25	$\text{C}(\text{=NH})\text{N}(\text{CH}_3)_2$	2.5	2.222	25*	$\text{CH}_2\text{NHC}(\text{CH}_3)_3$	0.52	-0.008
26	$\text{C}(\text{=O})\text{NHCH}_3$	3.1	2.823	26*	$\text{CH}_2\text{N}(\text{CH}_3)\text{CH}_2\text{CH}_2\text{OH}$	0.5	-0.020
27	$\text{CH}_2\text{CH}_2\text{CH}_2\text{N}(\text{CH}_3)_2$	0.27	-0.007	27*		2.2	0.902
28	$\text{CH}(\text{CH}_2\text{CH}_3)\text{N}(\text{CH}_3)_2$	0.22	0.091	28*	$\text{C}(\text{=O})\text{NH}_2$	1.8	2.773
29	$\text{CH}(\text{CH}_2\text{CH}_2\text{CH}_3)\text{N}(\text{CH}_3)_2$	0.065	0.342	29*	$\text{CH}(\text{CH}_3)\text{N}(\text{CH}_3)_2$	0.09	0.227
30	$\text{C}(\text{CH}_3)_2\text{N}(\text{CH}_3)_2$	0.11	-0.167	30*	$\text{CH}(\text{CH}_2\text{CH}_3)$	0.4	0.274
					$\text{CH}_2\text{N}(\text{CH}_3)_2$		
31	$\text{C}(\text{CH}_3)_2\text{CH}_2\text{NH}_2$	0.26	0.537	31*		0.046	-0.107
32	$\text{CHFCH}_2\text{N}(\text{CH}_3)_2$	0.59	0.312	32*		0.044	-0.482
33		0.12	-0.006	33*		0.12	0.294
34		0.065	-0.109	34*		0.07	0.244
35		0.12	-0.103	35*	CH_2OH	2.7	1.149
36		0.095	-0.103				
37		0.11	0.388		$\text{R}_5\text{-N-R}_6$		
38		0.11	0.383	36		0.05	0.251
39		0.23	0.507	37		0.15	0.408
40	Cl	2.7	2.422	38		0.57	0.464
41	CH_2NH_2	0.18	0.458	39		0.26	0.755

Table 1 continued

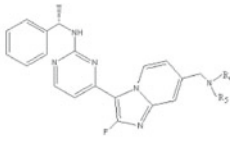

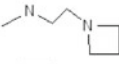

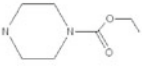

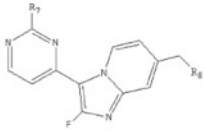
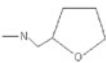




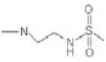

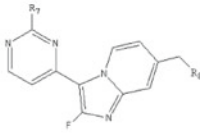







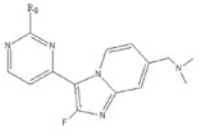






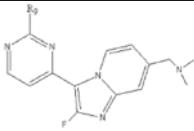
				40*		0.11	0.296	
	R ₅ -N-R ₆			41*		0.10	0.335	
42		0.13	0.226	42*		5.6	4.084	
43		0.11	0.003					
44		0.27	0.546	R ₇	R ₈			
45		0.07	0.348	43		0.06	0.135	
46		0.39	0.231	44		2.8	1.634	
47		0.15	0.428	45		1.3	1.137	
				46*		0.081	0.070	
	R ₇	R ₈		47*		0.60	0.601	
48			0.28	0.428	48*		0.20	0.552
49			0.29	0.497				
50			0.45	0.479	R ₉			
51			0.31	0.587	49	NH- <i>n</i> Pr	0.13	0.408
52			0.86	0.583	50*	NHMe	0.17	1.292

Table 1 continued


	R ₉		
53	H	1.9	1.623
54	N(CH ₃) ₂	2.5	2.223
55	NH-neopentyl	0.15	0.428
56	NH-Bn	0.081	0.070
57	OCH ₃	2.6	1.830

* Test-set compound

searched for constant values for all molecules; those detected descriptors were removed, and the other calculated descriptors were as the original variable set.

Data splitting

Rational division of the experimental data set into training, calibration, and test sets is a crucial step in the development and validation of reliable QSAR models. Ideally, this division is performed such that points representing the training, calibration, and test sets are distributed within the whole descriptor space occupied by the entire data set, and each point in the calibration and test set is close to at least one point in the training set. The k-MCA (k-means cluster analysis) may be used in training, calibration, and test series design (Kowalski and Wold, 1982). The idea consists of carrying out partition of the set of compounds under study into several statistically representative classes of chemicals. Thence, one may select from the members of all these classes the training, calibration, and test series. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in the compound series (training, calibration, and test). It permits the design of training, calibration, and test series which are representative of the entire experimental universe. As a result, k-MCA splits imidazopyridine derivatives into five clusters, with 13, 29, 36, and 45 training-set members and 24 calibration- and test-set members. Selection of the training, calibration, and test sets was carried out by randomly selecting compounds belonging to each cluster. Twenty-five compounds (see Table 1) were selected as the test set, another 25 compounds (see Table 1) were selected as the calibration set, and the remaining 57 compounds formed the training set. The training and calibration set were used to adjust the parameters of the models. And the test set was used to evaluate the predicted ability of the models once they were built.

Computational strategies

Variable selection strategies

The PSO algorithm was proposed by Kennedy and Eberhart, (1995). Similarly to other population-based algorithms, PSO can solve a variety of complicated optimization problems and shows a fast convergence rate in certain problems (Kennedy and Eberhart, 2001). In addition, it is very easy to configure and can be rapidly implemented.

For a population of M particles, particle i can be described as $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, where $i = 1, 2, \dots, M$. Each particle has a velocity along each dimension, represented as $V_i = (v_{i1}, v_{i2}, \dots, v_{iD})$. It also memorizes its previous best position $P_i = (p_{i1}, p_{i2}, \dots, p_{iD})$ (also known as pbest) and the best position P_g of all the particles (also known as gbest) (Pavlidis *et al.*, 2005).

$$P_i(t+1) = \begin{cases} X_i(t+1) & \text{if } f(X_i(t+1)) \geq f(P_i(t)) \\ P_i(t) & \text{if } f(X_i(t+1)) < f(P_i(t)) \end{cases} \quad (1)$$

$$f(P_g(t)) = \max\{f(P_1(t)), \dots, f(P_i(t)), \dots, f(P_M(t))\} \quad (2)$$

After finding the two best values, P_i and P_g , the particle updates its velocity according to the following formula:

$$V_i(t+1) = wV_i(t) + c_1r_1(t)(P_i(t) - X_i(t)) + c_2r_2(t)(P_g(t) - X_i(t)) \quad (3)$$

where c_1 and c_2 are the learning factors that determine how far a particle will move in each iteration, r_1 and r_2 are the elements from two uniform random sequences in the range of [0, 1], and w is the inertia weight. After updating its velocity, the particle's position is updated:

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (4)$$

The velocity $V_i(t+1)$ and position $X_i(t+1)$ are generally confined to an area to prevent the particles from flying out of the solution space.

The iterations are represented by the parameter t in the above formulae. The iterations are terminated when the maximum generation or the designated fitness value is reached.

To select the most significant variables, we define a new objective function, fitness, which can be calculated as follows:

$$\text{fitness} = a \times R^2 + b \times R_v^2 \quad (5)$$

where a and b are clamped to the range (0,1); moreover, $a+b = 1$, R^2 denotes the square of correlation coefficients of the training set, and R_v^2 denotes the square of correlation coefficients of the calibrating set. Figure 1 shows the flow diagram of the proposed PSO selection variable approach.

The new objective function, fitness, combined the result of the training set and calibration set to guide the PSO-selected important variables, and the values of a and b in the objective function are evaluated on the basis of the square of correlation coefficients (R_{ext}^2) of the test set. In the present work, to improve the predictive

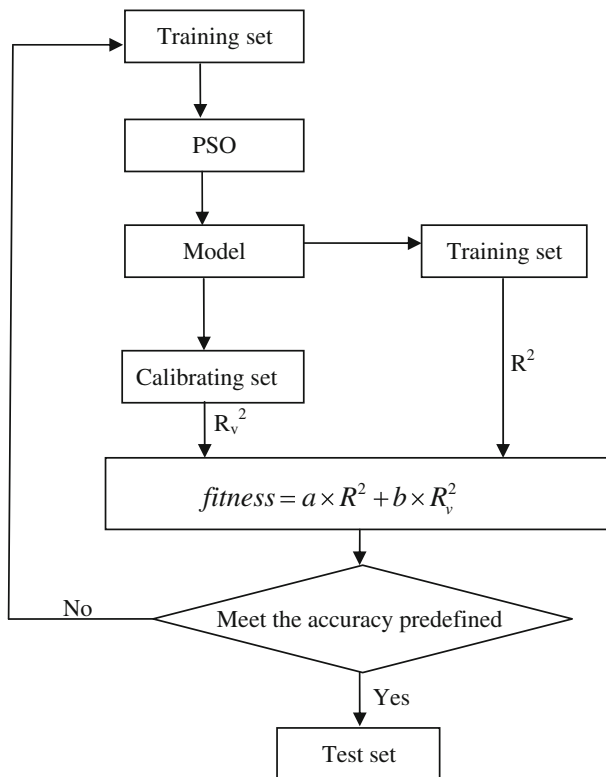


Fig. 1 Flow diagram of the proposed PSO selective variables approach

ability of the model, we apply the objective function fitness ($\text{fitness} = a \times R^2 + b \times R_v^2$) to guide the PSO-selected most relevant descriptors from the pool of 96 2D autocorrelation descriptors. Optimized parameters of the objective function fitness are $a = 0.4$ and $b = 0.6$ on the basis of maximization of the square of correlation coefficients (R_{ext}^2) of the test set. Optimized parameters of PSO are as follows: a population of 20 particles, $c_1 = c_2 = 1.8$, $V_{\text{max}} = 0.2$, $X_{\text{max}} = 1.0$, $X_{\text{min}} = 0.0$, and evolution for 100 generations. And the selected variables are reported in Table 2.

Methods of model construction

As mentioned in the previous section, the PSO method was used to select the most significant descriptors. The selected descriptors were used to construct some models by using multiple linear regression (MLR) and support vector machine (SVM) (Chihchung and Chihjen, 2009) techniques; these models are called PSO-MLR, and PSO-SVM. In each SVM model, the Gaussian radial basis function kernel was used because of its effectiveness and speed in the training process. The following SVM parameters were chosen to be optimized: c , cost (set the parameter C); ν (set the parameter ν of ν -SVC); and g , γ (set γ in kernel function [default $1/k$]). The values of optimized SVM parameters are listed in Table 3.

Table 2 Statistical parameters of different constructed QSAR models based on the 2D autocorrelation descriptor

Model	Variable	Training set						Test set	Calibration set
		R^2	R^2_{adj}	F	S	AIC	FIT	R^2_{ext}	R^2_v
PSO-MLR	ATS2m, ATS5m, ATS6m, ATS1v, ATSSe, ATS8e, MATS7v, MATS7e, GATS1v, GATS5e, GATS8p	0.550	0.427	4.480	0.872	124.228	0.268	0.166	0.304
PSO-SVM	ATS3m, ATS7m, ATS8m, ATS5v, ATS7v, ATS8e, MATS1v, MATS1e, MATS3e, MATS1p, GATS5p	0.897	0.872	35.727	0.412	84.154	2.208	0.660	0.743

Table 3 Parameters of SVM in different QSAR models

Descriptor	Variables	C	n	g
2D autocorrelations	ATS3m, ATS7m, ATS8m, ATS5v, ATS7v, ATS8e, MATS1v, MATS1e, MATS3e, MATS1p, GATS5p	2040	0.267	0.785
GETAWAY	H4m, HTe, HATS6e, HATSe, H8p, RTu, R7u+, R5 m+	1000	0.1	0.02
3D-MoRSE	Mor14u, Mor28u, Mor15m, Mor25m, Mor25v, Mor21e, Mor30e, Mor12p, Mor30p	900	0.3	0.034
WHIM	L2m, L3m, L2p, G2p, G3s, E2s	1550	0.4	0.065
Burden eigenvalues	BEHm1, BELm6, BEHv4, BELv6, BELv7, BELv8, BEHp3, BELp3	1500	0.45	0.03
Edge adjacency indexes	EPS1, EEig06x, EEig09x, EEig12x, EEig13x, EEig03d, EEig08d, EEig02r, EEig07r, ESpm13r	1500	0.1	0.1

All algorithms were written in MATLAB and run on a personal computer [Intel(R) Pentium(R), 4/3.20 GHz, 1.00-GB RAM].

Results and discussion

The studied data set includes selective compounds which are potent inhibitors of *E. tenella* cGMP-dependent protein kinase (Et-PKG). Et-PKG is essential for survival and represents a desirable therapeutic target (Scribner *et al.*, 2007). High-throughput screening of known kinase inhibitors resulted in the discovery of imidazopyridine derivatives as PKG inhibitors and broad-spectrum anticoccidial agents (Scribner *et al.*, 2008). The imidazopyridine analogs were tested for in vitro efficacy via the Ten_K (*Tenella* kinase) assay, which measures inhibition of Et-PKG enzyme activity.

QSAR model

Three 2D autocorrelation MLR and SVM models are reported in this work, respectively. In total, 21 descriptors were employed from the whole 2D autocorrelation pool. Statistical parameters of the two QSAR models are listed in Table 2. As reported in the table, the PSO-SVM model has a higher value of the square of the correlation coefficient ($R^2 = 0.897$), the adjusted square regression coefficient ($R_{\text{adj}}^2 = 0.872$), the Fisher ratio ($F = 35.727$), and the Kubinyi, (1994a, b) function ($\text{FIT} = 2.208$) and a lower value of the standard deviation ($\text{SD} = 0.412$) and Akaike's information criterion ($\text{AIC} = 84.154$) (Kuzmic *et al.*, 2006) than the PSO-MLR model does. So the statistical parameters of the PSO-SVM model are more robust than those of the PSO-MLR model. Comparison of the square of correlation coefficients (R_{ext}^2) of the test set, the value of PSO-SVM model was higher than another. From the comparison of the calibrating set results, the data indicate that the model is reliable and accurate.

Using the PSO-SVM model, satisfactory results were obtained. The results are shown in Fig. 2. With this model, R^2 for the training set was increased to 0.897 and S was reduced to 0.412. For the calibration set, R_V^2 was increased to 0.743, while S was reduced to 0.540. For the test set, R^2 was increased to 0.660, while S was reduced to 0.975, showing the good generalization ability of the SVM model. The smaller scatter of data points in Fig. 2 demonstrates that the SVM model is clearly superior both in fitness and in prediction performance.

The PSO-MLR model gave a squared standard error (s) of 0.872 for the training set, 0.925 for the calibration set, and 1.59 for the test set, and the corresponding correlation coefficients (R^2) were 0.550, 0.304, and 0.166, respectively.

Comparing the PSO-SVM and PSO-MLR models developed above, the PSO-SVM model is expected to be a better predictor for Et-PKG inhibitors than the PSO-MLR

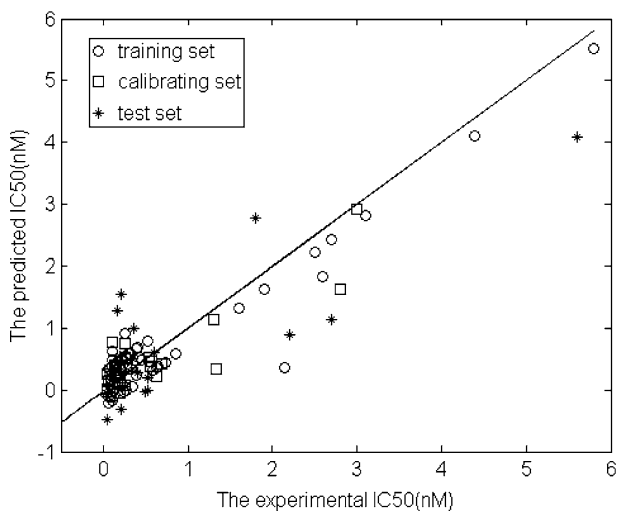


Fig. 2 Predicted vs. experimental IC₅₀ (nM) by SVM

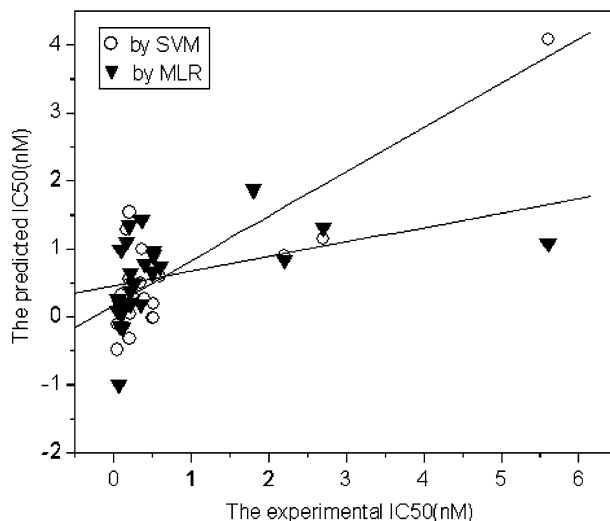


Fig. 3 Comparison of the predicted capability for the test by SVM and MLR

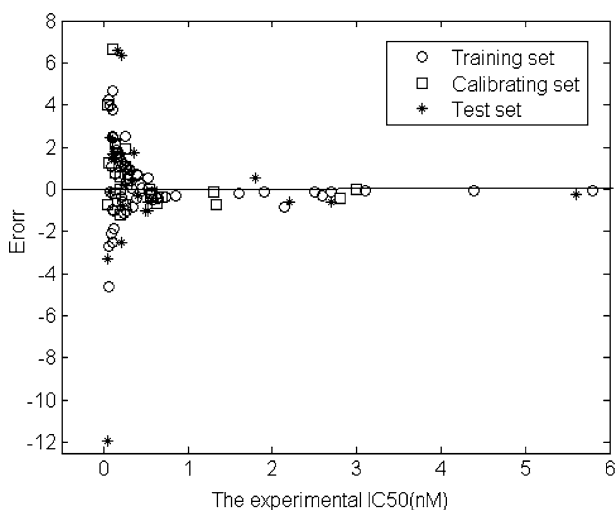


Fig. 4 The relative deviation vs. the corresponding experimental IC₅₀ by SVM

model. This indicates the good generalization capability of the PSO-SVM model. This is also demonstrated in Fig. 3, where we compared the predicted results for the test set alone by the PSO-SVM and PSO-MLR models. Clearly, the results using the nonlinear model show a relatively smaller bias than those using the PSO-MLR model.

Figures 4 and 5 show further comparison of the results using the PSO-SVM and PSO-MLR models, which plots the experimental IC₅₀ versus the relative deviation of the corresponding predicted and experimental IC₅₀ values. The plot shows that the MLR model can correctly predict rate constants for 57.89, 60.00, and 56.00%

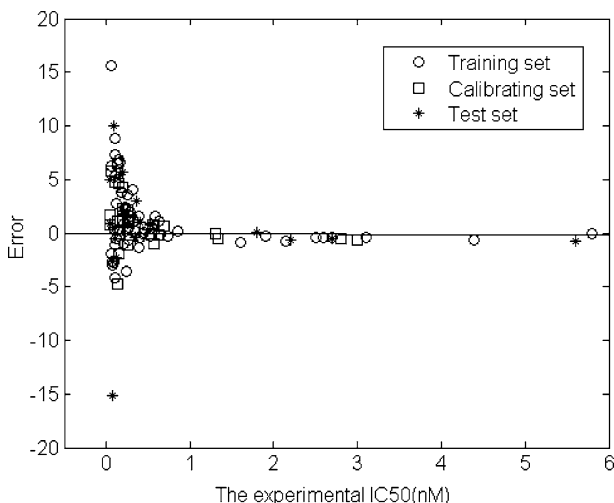


Fig. 5 The relative deviation vs. the corresponding experimental IC₅₀ by MLR

Table 4 Statistical parameters of the PSO-SVM model obtained for the six kinds of descriptors involved in the comparison

Models	Descriptor	Training set						Calibration set	Test set
		R^2_{adj}	R^2	S	F	AIC	FIT	R^2_v	R^2_{ext}
PSO-SVM	2D autocorrelations	0.872	0.897	0.412	35.727	84.154	2.208	0.743	0.660
	GETAWAY	0.349	0.442	0.930	4.749	113.660	0.314	0.143	0.054
	3D-MoRSE	0.595	0.660	0.733	10.148	107.093	0.662	0.514	0.161
	WHIM	0.515	0.567	0.802	10.929	104.643	0.705	0.247	0.207
	Burden eigenvalues	0.063	0.196	1.116	1.467	125.895	0.097	0.161	0.017
	Edge adjacency indexes	0.257	0.390	0.993	2.940	123.454	0.187	0.108	0.012

compounds in the training, calibration, and test sets, respectively, within a relative error of between -1.5 and 1.5 ; for the SVM model, the corresponding proportions were 70.18, 72.00, and 64.00%, respectively. Within a relative error of between -2.0 and 2.0 , the SVM model gave a corresponding accuracy of 80.70, 88.00, and 76.00%, respectively, whereas the MLR model can only predict 64.91, 68.00, and 64.00% of the compounds in the respective data set.

As we explained previously, one of the objectives of the current work was to compare the reliability and applicability of the 2D autocorrelation descriptors to describe the property under study compared with those of other different descriptors. Consequently, we developed another six models using the same data set that was included in the 2D autocorrelation QSAR model. The results obtained with Edge adjacency indexes, WHIM, 2D autocorrelation indexes, Burden eigenvalues, 3D MoRSE, and GETAWAY descriptors (Todeschini and Consonni, 2000) are listed in Table 4.

As reported in Table 4, the value of R^2 is <0.670 for all approaches except the 2D autocorrelation, which has an $R^2 = 0.897$. This approach also yields the best value for other statistical parameters like the standard deviation (S) of the training set and the AIC, which has the lowest value in comparison with the rest of the approaches. In similarly, the Fisher ratio (F) and the Kubinyi function (FIT) are the highest. Additional, the 2D autocorrelation descriptor presents the best R_v^2 and R_{ext}^2 values, which implies that the PSO-SVM model with the 2D autocorrelation descriptor has better predictive ability than the other models. For all these reasons, we consider that the 2D autocorrelation descriptor can be a useful tool for prediction of inhibitory activities taking into account the imidazopyridine derivatives.

Variables' interpretation of the best model

In the text, the 11 most significant descriptors have been selected by PSO as independent variables of the best model: ATS3m, ATS7m, ATS8m, ATS5v, ATS7v, ATS8e, MATS1v, MATS1e, MATS3e, MATS1p, and GATS5p.

The 2D autocorrelation descriptors represent the topological structure of the compounds but are more complex in nature than the classical topological descriptors. Computation of these descriptors involves the summations of different autocorrelation functions corresponding to different structural lags and leads to different autocorrelation vectors corresponding to the lengths of substructural fragments. Hence, it can distinguish the details of important substructural differences, however, the “traditional” descriptors (for example, $\log P$ or pK_a) cannot solve these questions. In the last decades, the 2D autocorrelation descriptor has been proven advantageous for establishing a QSAR model (Saíz-Urra *et al.*, 2007; Caballero *et al.*, 2006; Bauknecht *et al.* 1994; Moreau and Broto, 1980). Three spatial autocorrelation vectors are employed for modeling inhibitory activities: Broto-Moreau's autocorrelation coefficients [ATS; Eq. 6] (Moreau and Broto, 1980), Moran's, (1950) indexes (MATS; Eq. 7)], and Geary's, (1954) coefficients (GATS; Eq. 8].

$$ATS(p_k, l) = \sum_i \delta_{ij} p_{ki} p_{kj} \quad (6)$$

$$MATS(p_k, l) = \frac{N \sum_{ij} \delta_{ij} (p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{2L \sum_i (p_{ki} - \bar{p}_k)} \quad (7)$$

$$GATS(p_k, l) = \frac{N - 1 \sum_{ij} \delta_{ij} (p_{ki} - \bar{p}_k)}{4L \sum_i (p_{ki} - \bar{p}_k)} \quad (8)$$

where $ATS(p_k, l)$, $MATS(p_k, l)$, and $GATS(p_k, l)$ are Broto-Moreau's autocorrelation coefficients, Moran's indexes, and Geary's coefficients at spatial lag l , respectively; p_{ki} and p_{kj} are the values of physicochemical properties (i.e., atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities, and

atomic polarizabilities) k of atoms i and j , respectively; \bar{p}_k is the average value of property k ; and $\delta(l, d_{ij})$ is a Dirac-delta function defined as

$$\delta(l, d_{ij}) = \begin{cases} 1 & \text{if } d_{ij}=1 \\ 0 & \text{if } d_{ij} \neq 1 \end{cases} \quad (9)$$

where d_{ij} is the topological distance or spatial lag between atom i and atom j .

The 2D autocorrelation descriptors can be obtained by summing up the products of certain properties of the two atoms located at a given topological distance or spatial lag. There are slight differences among the descriptors of type ATSD, GATSD, and MATSD (Saíz-Urra *et al.*, 2007), but in general, they describe how the considered property is distributed along the topological structure. The most important factor in interpreting them in the model is the topological distance, once weighted equally.

In point of fact, the best model selected an optimum descriptor combination which includes atomic masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities as the most relevant key features. This result illustrates that a certain distribution of these properties is necessarily required to typify the imidazopyridine derivatives.

Aiming to figure out the impact of each input in the model, we performed a sensibility analysis in two ways (Cherqaoui *et al.*, 1998). In the first, the descriptor under study is removed from the model and the statistical coefficient R^2_{train} for the training set is analyzed. Comparison between these R^2 values and that calculated when no descriptor is removed gives an idea of the importance of the descriptor removed. In the other method, the mean of the absolute deviation values Δ_{mi} between the observed and the estimated value for all compounds is calculated. Finally, the contribution factor C_i (Cherqaoui *et al.*, 1998) of descriptor i ($i = 1-11$) is given by

$$C_i = (100 \times \Delta_{mi}) / \sum \Delta_{mi} \quad (10)$$

The results of the sensitivity analysis are reported in Table 5. As can be seen, greater decreases in R^2 correspond to higher values of C_i . Among the 11 descriptors in the best model, the descriptors with the highest impacts ($C_i > 9.5\%$) are GATS5p, ATS8m, ATS5v, and ATS8e. GATS5p indicates that the presence of polarizable atoms at a topological distance of 5 contributes positively to inhibition of Et-PKG enzyme activity. A possible polarizable atom pair at a topological distance of 5 is presented for compound **32**, the top inhibitor of Et-PKG enzyme activity in the calibration and test sets. Likewise, ATS8m, ATS5v, and ATS8e indicate that the

Table 5 Results of the sensitivity analysis for evaluating the impact of each descriptor in the best model by PSO-SVM

Removed descriptor	C% _{train}	R ² _{train}	Removed descriptor	C% _{train}	R ² _{train}
ATS3m	8.61	0.886	MATS1v	8.55	0.890
ATS7m	8.86	0.879	MATS1e	8.63	0.888
ATS8m	9.56	0.865	MATS3e	8.95	0.891
ATS5v	9.56	0.865	MATS1p	8.35	0.892
ATS7v	9.12	0.872	GATS5p	10.30	0.856
ATS8e	9.51	0.868	None	100	0.897

interaction between each pair of atoms at topological distances of 8, 5, and 8 contributes to the inhibition of Et-PKG enzyme activity, weighted by atomic masses, van der Waals volumes, and Sanderson electronegativities, respectively.

This reflects the fact that the effect of van der Waals volumes on inhibitory activities (IC₅₀) is greater (27.23%) than that of atomic masses (27.03%), Sanderson electronegativities (27.09%), and polarizabilities (18.65%). The vector of ATS7v is weighted by van der Waals volumes representing topological substructures of size 7. The impact of IC₅₀ is similar to that of ATS8m. The other descriptors, ATS3m, ATS7m, MATS1v, MATS1e, MATS3e, and MATS1p, influence inhibitory activities to about the same extent that ATS8m does.

This fact may be viewed in terms of association of activity information content with structural fragments of such size and could be related to the great importance of an adequate molecular size and/or shape for proper matching with the receptor. However, further interpretation of the information content of these descriptors is very complex, as their computations involve integration of the structural fragments, and thus it is not possible to traverse backward from a higher state to a lower one (Fernández *et al.*, 2005). These facts make our model interesting mainly as a predictive tool rather than for designing trends that are encoded within the model. Regarding this, our PSO-SVM model related to the 2D autocorrelation descriptor predictor should be useful to predict inhibitory activities of the new synthetic imidazopyridine derivatives against Et-PKG using the Ten_K assay *in vitro*.

Conclusion

We applied QSAR methodology to model the inhibitory activities of 107 imidazopyridine derivatives using 2D autocorrelation descriptors. To select the pertinent variables, we carried out PSO searches. Satisfactory quantitative models were obtained using the SVM method, achieving good results in calibration-set validations and external predictions. The present work demonstrates that the PSO-SVM approach is a robust and sensitive method to apply to QSAR research of imidazopyridine derivative inhibitory activities. In addition, comparisons with other descriptors such as the 3D MoRSE, Edge adjacency indexes, WHIM, 2D autocorrelation indexes, Burden eigenvalues, and GETAWAY descriptors were also carried out. The statistical parameters of the model and the validation results produced by the methodology that we propose are superior to those produced by the other descriptors. In this sense, the combination of SVM with PSO led to a useful method for the scientific community interested in QSAR development for imidazopyridine derivative inhibitory activities.

References

- Bauknecht H, Zell A, Bayer H, Levi P, Wagener M, Sadowski J, Gasteiger J (1994) Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J Chem Inf Comput Sci* 36:1205–1213
- Biftu T, Feng D, Fische M, Liang GB, Qian XX, Scribne A, Dennis R, Lee S, Liberato PA, Brown C, Gurnett A, Leavitt PS, Thompson D, Mathew J, Misura A, Samaras S, Tamas T, Sina JF, McNulty

- KA, McKnight CG, Schmatz DM, Wyvratt M (2006) Synthesis and SAR studies of very potent imidazopyridine antiprotozoal agents. *Bioorg Med Chem Lett* 16:2479–2483
- Caballero J, Garriga M, Fernández M (2006) 2D Autocorrelation modeling of the negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks. *Bioorg Med Chem* 14:3330–3340
- Caballero J, Fernández M, González-Nil FD (2008) Structural requirements of pyrido[2,3-d]pyrimidin-7-one as CDK4/D inhibitors: 2D autocorrelation, CoMFA and CoMSIA analyses. *Bioorg Med Chem* 16:6103–6115
- Cherqaoui D, Esseffar M, Villemin D, Cense JM, Chastrette M, Zakarya D (1998) Structure-musk odour relationship studies of tetralin and indan compounds using neural networks. *New J Chem* 22:839–843
- Chihchung C, Chihjen L (2009) LIBSVM—a library for support vector machines. Available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Curtin ML, Davidsen SK, Heyman HR, Garland RB, Sheppard GS, Florjancic AS, Xu LH, Carrera GM, Steinman DH, Trautmann JA, Albert DH, Magoc TJ, Tapang P, Rhein DA, Conway RG, Luo GJ, Denissen JF, Marsh KC, Morgan DW, Summers JB (1998) Discovery and evaluation of a series of 3-acylindole imidazopyridine platelet-activating factor antagonists. *J Med Chem* 41:74–95
- Davoud A, Nematollahi A, Iman M, Shafiee A (2009) Computational studies of new 1,4-dihydropyridines containing 4-(5)-chloro-2-ethyl-5-(4)-imidazolyl substituent: QSAR and docking. *Med Chem Res*. doi:10.1007/s00044-009-9171-2
- Diaz CA, Allocco J, Powles MA, Yeung L, Donald RGK, Anderson JW, Liberator PA (2006) Characterization of *Plasmodium falciparum* cGMP-dependent protein kinase (PfPKG): antiparasitic activity of a PKG inhibitor. *Mol Biochem Parasitol* 146:78–88
- Doležal R, Damme SV, Bultinck P, Waissner K (2009) QSAR analysis of salicylamide isosteres with the use of quantum chemical molecular descriptors. *Eur J Med Chem* 44:869–876
- Donald RGK, Allocco J, Singh SB, Nare B, Salowe SP, Wiltsie J, Liberator PA (2002) *Toxoplasma gondii* cyclic GMP-dependent kinase: chemotherapeutic targeting of an essential parasite protein kinase. *Eukaryot Cell* 1:317–328
- Ertepinar H, Gök Y, Geban Ö, Zden SÖ (1995) A QSAR study of the biological activities of some benzimidazoles and imidazopyridines against *Bacillus subtilis*. *Eur J Med Chem* 30:171–175
- Fernández M, Caballero J, Morales AH, Castro EA, González MP (2005) Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. *Bioorg Med Chem* 13:3269–3277
- Gearty RF (1954) The contiguity ratio and statistical mapping. *Incorp Stat* 5:115–145
- Gurnett AM, Liberator PA, Dulski PM, Salowe SP, Donald RGK, Anderson JW, Wiltsie J, Diaz CA, Harris G, Chang B, Darkin-Ratray SJ, Nare B, Crumley T, Blum PS, Misura AS, Tamas T, Sardana MK, Yuan J, Biftu T, Schmatz DM (2002) Purification and molecular characterization of cGMP-dependent protein kinase from *Apicomplexan paras.* *J Biol Chem* 277:15913–15922
- Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of the IEEE International joint conference on neural networks. Vol 4, pp 1942–1948
- Kennedy J, Eberhart RC (2001) Swarm intelligence. Morgan Kaufmann Publishers, San Francisco
- Khan AKR, Sahu VK, Singh RK, Khan SA (2009) Comparative QSTR study of saturated alcohols based on topological, constitutional, geometrical, and getaway descriptors. *Med Chem Res*. doi:10.1007/s00044-009-9166-z
- Kowalski RB, Wold S (1982) Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN (eds) Handbook of statistics. North-Holland, Amsterdam, pp 673–677
- Kubinyi H (1994a) Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant Struct Act Relat* 13:285–294
- Kubinyi H (1994b) Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. *Quant Struct Act Relat* 13:393–401
- Kuzmic P, Cregar L, Millis SZ, Goldman M (2006) Mixed-type noncompetitive inhibition of anthrax lethal factor protease by aminoglycosides. *FEBS J* 273:3054–3062
- Li F, Chen JW, Wang ZJ, Li J, Qiao XL (2009) Determination and prediction of xenoestrogens by recombinant yeast-based assay and QSAR. *Chemosphere* 74:1152–1157
- Liang GB, Qian XX, Feng D, Fische M, Brown CM, Gurnett A, Leavitt PS, Liberator PA, Misura AS, Tamas T, Schmatz DM, Wyvratt M, Biftu T (2007) Synthesis and SAR studies of potent imidazopyridine anticoccidial agents. *Bioorg Med Chem Lett* 17:3558–3561

- Mercader AG, Duchowicz PR, Fernández FM, Castro EA (2008) Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemometr Intell Lab* 92:138–144
- Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* 37:17–23
- Moreau G, Broto P (1980) The autocorrelation of a topological structure: a new molecular descriptor. *Nouv J Chim* 4:359–360
- Panek JJ, Jezierska A, Vračko M (2005) Kohonen network study of aromatic compounds based on electronic and nonelectronic structure descriptors. *J Chem Inf Model* 45:264–272
- Pavlidis NG, Parsopoulos KE, Vrahatis MN (2005) Computing Nash equilibria through computational intelligence methods. *J Comput Appl Math* 175:113–136
- Saíz-Urra L, González MP, Teijeira M (2007) 2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma. *Bioorg Med Chem* 15:3565–3571
- Salowe SP, Wiltsie J, Liberator PA, Donald RGK (2002) The role of a parasite-specific allosteric site in the distinctive activation behavior of *Eimeria tenella* cGMP-dependent protein kinase. *Biochemistry* 41:4385–4391
- Scribner A, Dennis R, Hong J, Lee S, McIntyre D, Perrey D, Feng D, Fisher M, Wyvratt M, Leavitt P, Liberator P, Gurnett A, Brown C, Mathew J, Thompson D, Schmatz D, Biftu T (2007) Synthesis and biological activity of imidazopyridine anticoccidial agents: part I. *Eur J Med Chem* 42:1334–1357
- Scribner A, Dennis R, Lee S, Ouvry G, Perrey D, Fisher M, Wyvratt M, Leavitt P, Liberator P, Gurnett A, Brown C, Mathew J, Thompson D, Schmatz D, Biftu T (2008) Synthesis and biological activity of imidazopyridine anticoccidial agents: part II. *Eur J Med Chem* 43:1123–1151
- Sharma S, Prabhakar YS, Singh P, Sharma BK (2008) QSAR study about ATP-sensitive potassium channel activation of cromakalim analogues using CP-MLR approach. *Eur J Med Chem* 43:2354–2360
- Todeschini R, Consonni V (2000) *Handbook of molecular descriptors*. Wiley, Weinheim