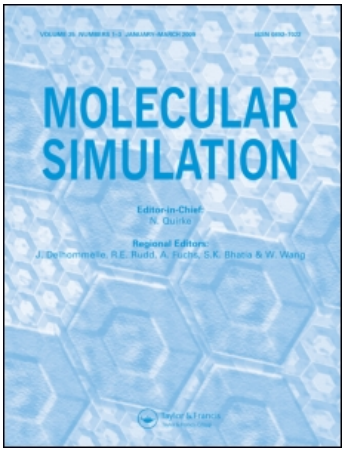


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

MIA-QSAR, PCA-ranking and least-squares support-vector machines in the accurate prediction of the activities of phosphodiesterase type 5 (PDE-5) inhibitors

Mohammad Goodarzi[ab]; Matheus P. Freitas[c]
[a] Department of Chemistry, Faculty of Sciences, Islamic Azad University, Arak, Markazi, Iran [b] Young Researchers Club, Islamic Azad University, Arak, Markazi, Iran [c] Departamento de Química, Universidade Federal de Lavras - UFLA, Lavras, MG, Brazil

Online publication date: 15 October 2010

To cite this Article Goodarzi, Mohammad and Freitas, Matheus P.(2010) 'MIA-QSAR, PCA-ranking and least-squares support-vector machines in the accurate prediction of the activities of phosphodiesterase type 5 (PDE-5) inhibitors', Molecular Simulation, 36: 11, 871 — 877
To link to this Article: DOI: 10.1080/08927022.2010.490261
URL: http://dx.doi.org/10.1080/08927022.2010.490261

Molecular Simulation
Vol. 36, No. 11, September 2010, 871–877

Taylor & Francis
Taylor & Francis Group

MIA-QSAR, PCA-ranking and least-squares support-vector machines in the accurate prediction of the activities of phosphodiesterase type 5 (PDE-5) inhibitors

Mohammad Goodarzi[ab] and Matheus P. Freitas[c]*

[a]*Department of Chemistry, Faculty of Sciences, Islamic Azad University, Arak Branch, Arak, Markazi, Iran;* [b]*Young Researchers Club, Islamic Azad University, Arak Branch, Arak, Markazi, Iran;* [c]*Departamento de Química, Universidade Federal de Lavras – UFLA, Caixa Postal 3037, 37200-000 Lavras, MG, Brazil*

(*Received 13 January 2010; final version received 28 April 2010*)

Phosphodiesterase type-5 (PDE-5) is a key enzyme involved in the erection process. PDE-5 inhibitors, such as Sildenafil (Viagra™), Vardenafil (Levitra™) and Tadalafil (Cialis™), are used for the treatment of erectile dysfunction. Computer-assisted modelling of biological activities of PDE-5 inhibitors may make quantitative structure–activity relationship (QSAR) models useful for the development of safer (low side effects) and more potent drugs. The multivariate image analysis applied to QSAR (MIA-QSAR) method, coupled to partial least-squares (PLS) regression, has provided highly predictive QSAR models. Nevertheless, regression methods which take into account nonlinearity, such as least-squares support-vector machines (LS-SVMs), are supposed to predict biological activities more accurately than the usual linear methods. Thus, together with prior variable selection using principal component analysis ranking, MIA-QSAR and LS-SVM regression were applied to model the bioactivities of a series of cyclic guanine derivatives (PDE-5 inhibitors), and the results were compared with those based on linear methodologies. MIA-QSAR/LS-SVM was found to improve greatly the prediction performance when compared with MIA-QSAR/PLS, MIA-QSAR/N-PLS, CoMFA/PLS and CoMSIA/PLS models.

Keywords: MIA-QSAR; PCA ranking; LS-SVM; PDE-5

1. Introduction

The term 'erectile dysfunction' can mean the inability to achieve erection, an inconsistent ability to do so or the ability to achieve only brief erections. These various definitions make estimating the incidence of erectile dysfunction difficult. The National Institute of Health estimates that erectile dysfunction affects as many as 30 million men in the USA [1]. Physical and psychological factors, in addition to lifestyle choices (such as smoking), may cause erectile dysfunction; biochemically, inhibition of phosphodiesterase type-5 (PDE-5), a cyclic guanosine 3′,5′-monophosphate (cGMP)-specific isozyme in the *corpus cavernosum* of the penis, increases the effective concentration of cGMP in the *corpus cavernosum*, leading to vasodilatation and increasing flow and erection. In order to produce cGMP, nitric oxide (NO) must be released from non-adrenergic, non-cholinergic neurons in the penis upon sexual stimulation. NO activates guanylyl cyclase, which produces cGMP [2–4]. Sildenafil (Viagra™) [5], Vardenafil (Levitra™) [6] and Tadalafil (Cialis™) [7] are currently the most commonly used drugs for the treatment of erectile dysfunction.

Recently, some derivatives of these milestone compounds were computationally designed in order to improve potency and side effects [8]. Also, a quantitative structure–activity relationship analysis based on multivariate image analysis (MIA-QSAR) was performed to model the bioactivities of some cyclic guanine derivatives (PDE-5 inhibitors) [9], demonstrating predictive ability comparable with the ones obtained through comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) [2]. QSAR/QSPR (quantitative structure–property relationship) studies are important for designing new compounds in lieu of testing experimentally which molecular system would be more appropriate, reducing time and cost. Consequently, researchers have paid attention to introduce new methodologies with high quality and minimum cost to achieve a robust QSAR model. Partial least squares (PLS) is a famous and popular regression technique, which has been used in a variety of 3D-QSARs, such as the one previously performed on this class of compounds, through CoMFA and CoMSIA [2]. However, external validation gave poor correlation between experimental and predicted activity values in all previous works, which was attributed to external samples not calibrated in the prior modelling step. The badly predicted results may also be due to nonlinearity, which is not accounted for in linear methods, such as the PLS method, during regression. In turn, support-vector machines (SVMs), originally proposed and developed by Vapnik [10], are based on linear or nonlinear radial basis function (RBF) kernels, and thus can be applied to improve correlation of data with nonlinear

*Corresponding author. Email: matheus@dqi.ufla.br

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927022.2010.490261
http://www.informaworld.com

nature [11]. This, together with principal component (PC) ranking as a variable selection method, has provided an improved QSAR model for a series of checkpoint kinase WEE1 inhibitors [12]. Thus, selection of MIA descriptors by PC ranking followed by least-squares SVM (LS-SVM) regression was carried out here to derive a QSAR model for a series of PDE-5 inhibitors, in order to circumvent the problem of nonlinearity and to enhance the estimation performance for an external test set, making the prediction of activities of novel congeners to be more reliable.

2. Computational methods

The building of the **X** matrix (descriptor block) to proceed with the MIA-QSAR analysis was performed as previously described [9], and is detailed as follows. The structures of compounds **1**–**49** (Table 1) were systematically built using appropriate software, ACD/ChemSketch [13], and then converted to bitmaps in 250 × 250 pixel windows, with a resolution of 102 × 102 points per inch. All the molecular structures were fixed by a common point among them in a given coordinate, since the shapes should be superimposed afterwards, as a 2D alignment to allow maximum similarity. In our data-set, the pixel located at the 50 × 100 coordinate (on the carbonyl carbon, present in the whole series) was used as the reference in the alignment step. Each 2D image was read and converted into binaries (double array in Matlab [14]), and a three-way array, the predictor block, was built by grouping the 49 treated images, giving a 49 × 250 × 250 array. The 3D array was unfolded to an **X** matrix (49 × 62,500), in order to be correlated with the **Y** block (the column vector of activities) using LS-SVM regression.

In LS-SVM, a linear estimation is done in kernel-induced feature space ($y = w^{\mathrm{T}}\phi(x) + b$, where w corresponds to the weights and ϕ denotes a feature map). However, the optimisation problem is described as

$$\min J(w, \xi) = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{N}\xi_i^2,$$

$$\text{subjected to} \quad y_i = w^{\mathrm{T}}\phi(xi) + b + \xi_i, \quad i = 1, 2, \ldots, n.$$

It is worth mentioning that C is the punishment factor, which determines the trade-off between the complexities of the LS-SVM model and had to be optimised by the user.

The LS-SVM's loss function is different from the standard SVM. For the optimisation problem, the Lagrange function is introduced as follows:

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i\{w^{\mathrm{T}}\phi(x_i) + b + \xi_i - y_i\}.$$

In this equation, α_i is the Lagrange multiplier. By eliminating w and ξ, the Karush–Kuhn–Tucker (KKT) system is obtained as

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_1, x_1) + 1/C & \cdots & K(x_1, x_n) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & K(x_1, x_n) & \cdots & K(x_n, x_n) + 1/C \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix},$$

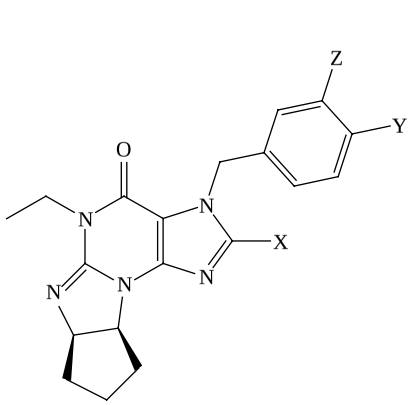
where $\alpha = [\alpha_1, \ldots, \alpha_n]^{\mathrm{T}}$ and Mercer's condition [15], $K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j)$ is a kernel function and determines both a nonlinear mapping, $x \rightarrow \phi(x)$, and the corresponding inner product, $\phi(x_i)^{\mathrm{T}}\phi(x_j)$. This leads to the following nonlinear regression function:

$$y(x) = \sum_{i}^{N}\alpha_i^* K(x_i, x_j) + b^*,$$

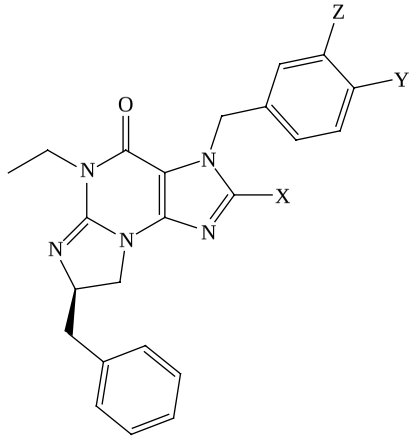
where α_i^* and b^* are optimal solutions to a linear system. In this paper, the RBF kernel was used as the kernel function, $\exp(-(\|x_i - x_j\|2)/2\sigma^2)$, a simple Gaussian function, where σ^2 is the width of the Gaussian, so C and σ, which are the relative weights of the regression error, and the kernel parameter of the RBF kernel should be optimised by the user to obtain the support vector. It should be noted that a five-fold cross-validation was used to tune the optimised C and σ^2, which were 87.53 and 9.45, respectively.

However, in order to select the more relevant MIA descriptors, principal component analysis (PCA) ranking variable selection was performed prior to regression. The data-set composed of 49 samples was divided into training set (37 compounds, 75% of the whole data-set) and test set (12 compounds), exactly as previously reported in CoMFA, CoMSIA and MIA-QSAR/PLS studies [2,9]. Also, the folded (three-way) array was treated using multilinear PLS (N-PLS). The model validation was achieved through leave-one-out cross-validation (LOO-CV) and leave-20%-out cross-validation (L20%O-CV), as well as by an external validation (test set), and the quality of the results was evaluated by analysing r^2 and q^2, the squared correlation coefficients of experimental vs. fitted/predicted activities of calibration and validation, respectively. In addition, the root-mean-square errors of calibration and external validation/prediction (RMSEP), relative standard error of prediction (RSEP), mean absolute error (MAE), Fischer (F)-test and t-test were used as statistical parameters to account for the predictive ability of the model.

Table 1. Forty-nine PDE-5 inhibitors used in the QSAR analysis, and experimental, fitted and predicted pIC_{50} values obtained using different QSAR methods.



1–50



51–59

No.	X	Y	Z	Exp.	MIA-QSAR/LS-SVM	MIA-QSAR/N-PLS	MIA-QSAR/PLS [9]	CoMFA [2]	CoMSIA [2]
1	CH_2-Ph	H	H	6.89	6.83	6.92	6.81	7.06	6.97
2	CH_2-Ph	Me	H	7.40	7.37	7.45	7.22	7.14	7.08
3	CH_2-Ph	Cl	H	6.80	7.14	6.81	6.87	7.01	7.05
4	CH_2-Ph	OMe	Cl	7.24	7.22	7.24	7.19	7.11	7.32
5	CH_2-Ph	3,4-OCH_2O		7.34	7.24	7.02	7.34	7.35	7.31
6	CCPh	Cl	H	8.30	8.47	8.44	8.92	8.59	8.49
7	CCPh	OMe	H	8.51	8.53	8.36	8.53	8.47	8.60
8	CCPh	OH	H	9.52	9.34	9.14	9.03	9.06	9.18
9	OEt	OMe	H	7.74	7.78	7.93	8.03	7.90	7.84
10	SEt	OMe	H	7.89	7.76	7.94	8.03	7.92	8.00
11	COOMe	OMe	H	8.17	8.30	8.15	8.23	8.00	8.02
12	CN	OMe	H	8.20	7.85	8.19	8.21	8.10	8.14
13	$CONH_2$	OMe	H	8.40	8.29	8.58	8.51	8.50	8.41
14	CF_3	OH	H	8.15	8.49	8.49	8.43	8.32	8.18
15	$CONH_2$	OH	H	8.68	8.98	8.54	8.52	8.83	8.70
16	SEt	OH	H	8.74	8.80	8.46	8.48	8.59	8.56
17	H	OMe	Br	8.19	7.99	8.05	8.48	8.11	8.30
18	H	OMe	Cl	8.08	7.80	7.98	8.46	8.11	7.97
19	$CONH_2$	OMe	Br	8.82	8.70	8.73	8.56	8.64	8.71
20	$CONH_2$	OH	Br	9.05	9.00	8.96	9.02	9.04	9.08
21	$CONH_2$	OH	Cl	8.96	8.98	8.90	9.04	8.99	8.91
22	CONHMe	OMe	Cl	8.64	8.67	8.63	8.35	8.62	8.55
23	CONHMe	OH	Cl	8.89	8.92	8.87	8.88	8.98	9.01
24	OEt	OMe	Br	8.27	8.23	8.09	8.08	8.31	8.51
25	OEt	OMe	Cl	8.19	8.05	8.01	8.06	8.08	8.12

Table 1 – *continued*

No.	X	Y	Z	Exp.	MIA-QSAR/LS-SVM	MIA-QSAR/N-PLS	MIA-QSAR/PLS [9]	CoMFA [2]	CoMSIA [2]
26	OEt	OH	Cl	8.68	8.54	8.74	8.82	8.63	8.67
27	OMe	OMe	Br	8.55	8.44	8.52	8.41	8.63	8.58
28	OMe	OMe	Cl	8.30	8.29	8.44	8.39	8.33	8.16
29	OMe	OH	Br	8.80	8.75	8.87	8.89	8.86	8.85
30	OMe	OH	Cl	8.70	8.70	8.81	8.91	8.88	8.77
31	$CONH_2$	OMe	Cl	8.55	8.66	8.66	8.53	8.61	8.50
32	$CONH_2$	OH	Cl	8.52	8.50	8.89	8.78	8.79	8.87
33	OBn	OH	Cl	8.68	8.68	8.67	8.34	8.62	8.77
34	OMe	OMe	CN	7.96	7.97	8.20	7.93	8.02	7.78
35	OMe	OMe	Br	8.66	8.55	8.51	8.47	8.42	8.56
36	OMe	OH	Br	8.89	8.75	8.86	8.74	8.74	8.78
37	OMe	OH	Cl	8.72	8.76	8.80	8.76	8.76	8.70
38[a]	CH_2-Ph	OMe	H	7.70	7.69	7.16	7.16	6.89	7.17
39[a]	CH_2-Ph	OH	H	8.41	8.14	7.51	7.16	7.35	7.66
40[a]	CH_2-Ph	NH_2	H	6.26	6.81	7.51	7.20	7.04	6.87
41[a]	CCPh	H	H	8.10	8.11	8.54	8.94	8.58	8.40
42[a]	H	OMe	H	7.26	7.60	7.90	8.43	7.95	7.83
43[a]	CF_3	OMe	H	7.32	7.81	8.09	8.14	7.89	7.74
44[a]	NH_2	OH	H	7.66	7.54	8.51	8.72	8.63	8.21
45[a]	OEt	OH	H	9.05	8.80	8.46	8.49	8.55	8.50
46[a]	N_3	OH	H	9.21	8.67	8.48	8.42	8.68	8.13
47[a]	H	OH	Br	8.20	8.38	8.75	9.12	7.95	8.54
48[a]	OMe	OMe	Cl	8.80	8.53	8.44	8.45	8.17	8.10
49[a]	OMe	OH	CN	7.37	7.60	8.55	8.18	8.38	8.36

[a]Compounds pertain to the test set.

3. Results and discussion

In the previous work using MIA-QSAR to this series of PDE-5 inhibitors [9], PLS was used as a regression method and the correlation results were $r^2 = 0.864$ and $q^2 = 0.605$, using five latent variables. In order to investigate the effect of using a three-way array during the modelling, we have used N-PLS as a regression method, since it is supposed to be advantageous over unfolded PLS [16]. The results of Tables 1 and 2 indicate that a slight improvement in estimation and prediction was achieved through MIA-QSAR/N-PLS when compared with the PLS-based model. Although very similar to previous results using CoMFA and CoMSIA, the MIA-QSAR/N-PLS model was not reliable for the prediction of activities of external compounds. In order to improve the existing results for the series of PDE-5 inhibitors, PC ranking as a feature selection and a nonlinear regression using LS-SVM were applied.

PCA creates p latent variables (Y) as linear combinations of the original p variables (X), in such a way that new orthogonal axes are built to explain the maximum variance possible in just a few dimensions,

$$Y_i = e_i^{\mathrm{T}} X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p,$$

where the unknown vector e_i establishes the ith linear combination, for $i = 1, \ldots, p$.

The PC ranking technique was chosen as the feature selection method. This method is an extremely useful explorative tool that maps samples through scores and individual variables through scores in a new vector space defined by high correlative PCs; in this work, the order of the four selected PCs based on their decreasing correlation coefficients was PC3 > PC5 > PC7 > PC1 (Table 3). It can be seen that among the PCs, the four components PC3, PC5, PC7 and PC1 show, respectively, the largest correlation coefficients with pIC_{50}. Although PC1 represents 89.2% of the variance in the space of scores, it shows a lower correlation with pIC_{50} when compared with PC3, PC6 and PC7. Therefore, in the PCA ranking method, the criterion for the selection of the space of PCs containing the important response matrix is the correlation of PCs with pIC_{50}. Obviously, there is no significant correlation (>0.9) among descriptors selected by PC ranking. On the other hand, to demonstrate the absence of chance correlation on the LS-SVM model, a **Y** scrambling test was performed, in which the **Y** block (pIC_{50}) was randomised while the **X** one was not, to determine the correlation and predictability of the resulting model. The correlation obtained using this procedure was $r^2 = 0.119 \pm 0.041$ (average of 40 repetitions). This demonstrates that the good correlation obtained in the real calibration was not casual. After feature selection, variables were used as input for nonlinear regression, namely LS-SVMs.

The use of LS-SVM as the regression procedure gave a high correlation both in calibration and validation/prediction, as illustrated in Figure 1. The propagation of residuals shown at both sides of the zero line in Figure 2 indicates that no systematic error exists in the development

Table 2. Statistics for the MIA-QSAR models and for those in the literature.

Parameters	MIA-QSAR/PC ranking/LS-SVM	MIA-QSAR/N-PLS	MIA-QSAR/PLS [9]	CoMFA [2]	CoMSIA [2]
RMSEP					
Training set	0.15	0.15	0.22	0.15	0.14
Test set	0.33	0.78	0.87	0.73	0.66
RSEP (%)					
Training set	1.76	1.84	2.67	1.85	1.72
Test set	4.07	9.76	10.95	9.12	8.21
MAE (%)					
Training set	5.35	5.62	6.88	5.73	5.52
Test set	15.06	24.73	26.42	23.98	22.65
r^2					
Training set	0.940	0.933	0.864	0.932	0.941
Test set	0.930	0.190	0.137	0.260	0.358
q^2_{LOO}					
Training set	0.876	0.724	0.605	0.755	0.788
$q^2_{\mathrm{L20\%O}}$					
Training set	0.853	0.580			
F					
Training set	544.41	485.07	221.48	479.68	559.47
Test set	132.07	2.34	1.58	3.51	5.56
t					
Training set	23.33	22.02	14.88	21.90	23.65
Test set	11.49	1.53	1.26	1.87	2.36

Table 3. Correlation matrix of selected descriptors.

	pIC_{50}	PC3	PC5	PC7	PC1
pIC_{50}	1				
PC3	0.2485	1			
PC5	0.1814	0.0001	1		
PC7	0.1077	0.003	0.0123	1	
PC1	0.0386	0.0011	0.0015	0.1656	1

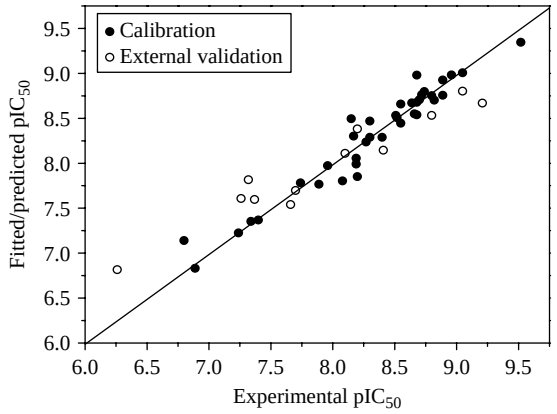


Figure 1. Plot of experimental vs. fitted/predicted pIC_{50}, using the MIA-QSAR/PC ranking/LS-SVM method.

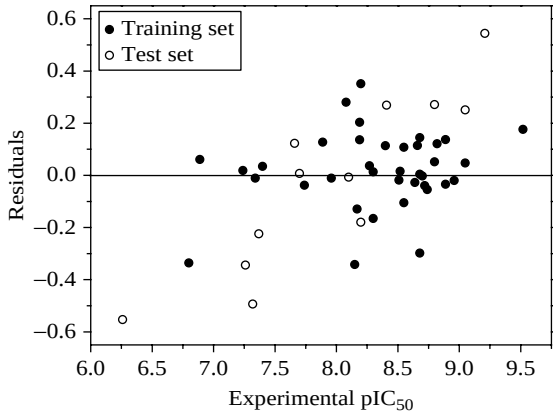


Figure 2. Plot of residuals vs. experimental pIC_{50}.

of the LS-SVM model. Tables 1 and 2 show that the results using this approach improved significantly, i.e. higher correlation (r^2 and q^2), lower errors (RMSEP, RSEP and MAE) and better statistics (F and t) were achieved through the MIA-QSAR/LS-SVM method when compared with MIA-QSAR/PLS, MIA-QSAR/N-PLS, CoMFA and CoMSIA for a series of 49 cyclic guanine derivatives. In addition, the model showed to be stable, since independent of the number of samples left out, q^2 did not vary significantly, as confirmed by the LOO-CV and L20%O-CV results of Table 2. Nevertheless, there are substituents not calibrated in the test set, e.g. NH_2 and N_3 as X, and this extrapolation is not accurately predicted by any model (e.g. compound **46** showed the highest deviation from the experimental value even in the MIA-QSAR/LS-SVM model).

Further tests on validation are required to achieve a reliable QSAR model; according to Golbraikh and Tropsha [17], in addition to r being close to 1, at least one of the correlation coefficients for regressions through the origin (observed vs. predicted activities or predicted vs. observed activities), i.e. r_0^2 or $r_0'^2$, should be close to r^2. Furthermore, at least one slope of regression lines (k or k') through the origin should be close to 1. Models would be considered acceptable, if they satisfy all of the following conditions: (i) $q^2 > 0.5$, (ii) $r^2 > 0.6$ and (iii) r_0^2 or $r_0'^2$ is close to r^2, such that $[(r^2 - r_0^2)/r^2]$ or $[(r^2 - r_0'^2)/r^2] < 0.1$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$. Moreover, other parameters have been implemented to guarantee the validity of QSAR models [18–20]: an additional statistic for external validation r_m^2 is calculated as $r_m^2 = r^2(1 - (r^2 - r_0^2)^{1/2})$. The parameters r^2 and r_0^2 are squared correlation coefficient values between observed and predicted values of the test-set compounds with and without the intercept, respectively. For a model with good external predictability, the r_m^2 value should be greater than 0.5. A similar procedure can be performed to obtain $r_{m(\mathrm{LOO})}^2$, based on the LOO predicted values. We have found that the MIA-QSAR/LS-SVM model obeys all these requirements, as depicted in Table 4, while others do not positively fulfil the whole set of the above validation tests.

Table 4. Validation tests performed for the various QSAR models.

Models	Sets	$r_{m(\mathrm{LOO})}^2$	r_m^2	r_0^2	$r_0'^2$	k	k'	$(r^2 - r_0^2)/r^2$	$(r^2 - r_0'^2)/r^2$	$\lvert r_0^2 - r_0'^2 \rvert$
LS-SVM	Training set	0.658	0.905	0.938	0.938	0.997	1.002	0.004	0.001	0.001
	Test set		0.439	0.652	0.844	1.000	0.999	0.299	0.094	0.191
N-PLS	Training set	0.438	0.880	0.923	0.925	0.999	1.001	0.003	0.001	0.002
	Test set		−0.050	−1.374	0.166	1.019	0.973	0.825	0.127	1.539
PLS [9]	Training set	0.348	0.785	0.858	0.865	1.000	0.999	0.010	0.002	0.007
	Test set		0.010	−0.715	−0.023	1.025	0.966	6.230	1.168	0.692
CoMFA [2]	Training set	0.506	0.864	0.927	0.932	1.000	1.000	0.006	0.000	0.005
	Test set		0.034	−0.494	0.215	1.001	0.991	−2.901	0.174	0.709
CoMSIA [2]	Training set	0.539	0.887	0.938	0.941	1.000	1.000	0.003	0.000	0.003
	Test set		−6.990	−0.682	0.357	0.995	0.998	2.907	0.000	1.039

Overall, we found that nonlinearity must be accounted for to achieve accurate estimations and predictions. This might be easily achieved using LS-SVM regression, which increased r^2 for an external test set, in comparison with linear-based methods, from 0.14–0.36 to 0.93. The high correlation between experimental and predicted pIC_{50} obtained for a series of 49 PDE-5 inhibitors makes the MIA-QSAR/LS-SVM model useful for the prediction of activities of new related derivatives. These compounds may be designed by combining significant substructures or substituents previously calibrated, e.g. those possessing high influence on the bioactivity values, forming a new chemical structure. For example, compounds **8** and **20**, which exhibit pIC_{50} values above 9, can be mixed to give a new, potentially useful drug with X, Y and Z substituents as CCPh, OH and Br, respectively. A similar procedure, assisted by docking studies, has been performed previously [8,21].

4. Conclusion

Previous QSAR analyses, including those based on MIA, were found to be very estimative, but not equally predictive, as shown by the poor results of external validation. Despite some substituents not calibrated being used in the test set, it should not be supposed that bad predictions are only due to the choice of compound sets, but also selection of suitable descriptors and, especially, the use of nonlinear regression between dependent and independent variables must be considered. This made the model for a series of PDE-5 inhibitors to be approximately 10% more estimative than the MIA-QSAR/PLS-based model and really useful for the prediction of unknown, potential derivatives.

Acknowledgements

The authors are grateful to the Young Researcher Club of Islamic Azad University and FAPEMIG for the financial support of this research, as well as to CNPq for the fellowship (to M.P.F.).

References

[1] National Institute of Health (NIH) Consensus Conference, *NIH consensus development panel on impotence*, J. Am. Med. Assoc. 270 (1993), pp. 83–90.

[2] G.-F. Yang, H.-T. Lu, Y. Xiong, and C.-G. Zhan, *Understanding the structure–activity and structure–selectivity correlation of cyclic guanine derivatives as phosphodiesterase-5 inhibitors by molecular docking, CoMFA and CoMSIA analyses*, Bioorg. Med. Chem. 14 (2006), pp. 1462–1473.

[3] D.A. Pissarnitski, T. Asberom, C.D. Boyle, S. Chackalamannil, M. Chintala, J.W. Clader, W.J. Greenlee, Y. Hu, S. Kurowski, J. Myers, J. Palamanda, A.W. Stamford, S. Vemulapalli, Y. Wang, P. Wang, P. Wu, and R. Xu, *SAR development of polycyclic guanine derivatives targeted to the discovery of a selective PDE5 inhibitor for treatment of erectile dysfunction*, Bioorg. Med. Chem. Lett. 14 (2004), pp. 1291–1294.

[4] C.D. Boyle, R. Xu, T. Asberom, S. Chackalamannil, J.W. Clader, W.J. Greenlee, H. Guzik, Y. Hu, Z. Hu, C.M. Lankin, D.A. Pissarnitski, A.W. Stamford, Y. Wang, J. Skell, S. Kurowski, S. Vemulapalli, J. Palamanda, M. Chintala, P. Wu, J. Myers, and P. Wang, *Optimization of purine based PDE1/PDE5 inhibitors to a potent and selective PDE5 inhibitor for the treatment of male ED*, Bioorg. Med. Chem. Lett. 15 (2005), pp. 2365–2369.

[5] N.K. Terrett, A.S. Bell, D. Brown, and P. Ellis, *Sildenafil (Viagra), a potent and selective inhibitor of Type 5cGMP phosphodiesterase with utility for the treatment of male erectile dysfunction,* Bioorg. Med. Chem. Lett. 6 (1996), pp. 1819–1824.

[6] H. Haning, U. Niewöhner, T. Schenke, M. Es-Sayed, G. Schmidt, T. Lampe, and E. Bischoff, *Imidazo[5,1-f][1,2,4]triazin-4(3H)-ones, a new class of potent PDE 5 inhibitors*, Bioorg. Med. Chem. Lett. 12 (2002), pp. 865–868.

[7] A. Daugan, P. Grondin, C. Ruault, A.C.M. de Gouville, H. Coste, J.M. Linget, J. Kirilovsky, F. Hyafil, and R. Labaudinière, *The discovery of tadalafil: A novel and highly selective PDE5 inhibitor. 2: 2,3,6,7,12,12a-hexahydropyrazino[1′,2′:1,6]pyrido [3,4-b]indole-1,4-dione analogues*, J. Med. Chem. 46 (2003), pp. 4533–4542.

[8] J.E. Antunes, M.P. Freitas, E.F.F. da Cunha, T.C. Ramalho, and R. Rittner, In silico *prediction of novel phosphodiesterase type-5 inhibitors derived from Sildenafil, Vardenafil and Tadalafil*, Bioorg. Med. Chem. 16 (2008), pp. 7599–7606.

[9] J.E. Antunes, M.P. Freitas, and R. Rittner, *Bioactivities of a series of phosphodiesterase type 5 (PDE-5) inhibitors as modelled by MIA-QSAR*, Eur. J. Med. Chem. 43 (2008), pp. 1632–1638.

[10] V.N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998, 736p.

[11] H. Li, Y. Liang, and Q. Xu, *Support vector machines and its applications in chemistry*, Chemom. Intell. Lab. Syst. 95 (2008), pp. 188–198.

[12] R. Cormanich, M. Goodarzi, and M.P. Freitas, *Improvement of multivariate image analysis applied to quantitative structure–activity relationship (QSAR) analysis by using wavelet-principal component analysis ranking variable selection and least-squares support vector machine regression: QSAR study of checkpoint kinase WEE1 inhibitors*, Chem. Biol. Drug Des. 73 (2009), pp. 244–252.

[13] ACD/ChemSketch, *Version 8.17*, Advanced Chemistry Development, Inc., Toronto, 2005.

[14] Matlab, *Version 7.5*, MathWorks, Inc., Natick, 2007.

[15] J. Mercer, *Functions of positive and negative type and their connection with the theory of integral equations*, Philos. Trans. R. Soc. Lond. A 209 (1909), pp. 415–446.

[16] M.M.C. Ferreira, *Multivariate QSAR*, J. Braz. Chem. Soc. 13 (2002), pp. 742–753.

[17] A. Golbraikh and A. Tropsha, *Beware of q2!*, J. Mol. Graph. Modell. 20 (2002), pp. 269–276.

[18] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.

[19] P.P. Roy, S. Paul, I. Mitra, and K. Roy, *On two novel parameters for validation of predictive QSAR models*, Molecules 14 (2009), pp. 1660–1701.

[20] K. Roy and P.P. Roy, *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques*, Eur. J. Med. Chem. 44 (2009), pp. 2913–2922.

[21] J.R. Pinheiro, M. Bitencourt, E.F.F. da Cunha, T.C. Ramalho, and M.P. Freitas, *Novel anti-HIV cyclotriazadisulfonamide derivatives as modeled by ligand- and receptor-based approaches*, Bioorg. Med. Chem. 16 (2008), pp. 1683–1690.