Support vector machines: Development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives

ARTICLE *in* EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY · APRIL 2010

Impact Factor: 3.45 · DOI: 10.1016/j.ejmech.2010.01.002 · Source: PubMed

Andreea R Schmitzer
Université de Montréal

Didier Villemin
Université de Caen Normandie

Abdellah Jarid
Cadi Ayyad University

Driss Cherqaoui
Cadi Ayyad University

Vol. 45 • No 4 • April 2010 • ISSN 0223-5234
ELSEVIER
EUROPEAN JOURNAL OF
MEDICINAL
CHEMISTRY
Available online at www.sciencedirect.com
ScienceDirect

European Journal of Medicinal Chemistry 45 (2010) 1590–1597

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Support vector machines: Development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives

Rachid Darnag [a], E.L. Mostapha Mazouz [a], Andreea Schmitzer [b], Didier Villemin [c], Abdellah Jarid [a], Driss Cherqaoui [a,*]

[a] Département de Chimie, Faculté des Sciences Semlalia, BP 2390, Université Cadi Ayyad, Marrakech, Morocco
[b] Université de Montréal- Faculté des Arts et Sciences – Département de Chimie 2900 Edouard Montpetit CP 6128 Succursale Centre Ville, H3C 3J7 Montréal, Québec, Canada
[c] Ecole Nationale Supérieure d'Ingénieurs (E.N.S.I.) I. S. M. R. A., LCMT, UMR CNRS n 6507, 6 boulevard Maréchal Juin, 14050 Caen, France

ARTICLE INFO

Article history:
Received 12 May 2009
Received in revised form
4 January 2010
Accepted 6 January 2010
Available online 14 January 2010

Keywords:
QSAR
Anti-HIV
TIBO
Support vector machines
Neural networks

ABSTRACT

The tetrahydroimidazo[4,5,1-jk][1,4]benzodiazepinone (TIBO) derivatives, as non-nucleoside reverse transcriptase inhibitors, acquire a significant place in the treatment of the infections by the HIV. In the present paper, the support vector machines (SVM) are used to develop quantitative relationships between the anti-HIV activity and four molecular descriptors of 82 TIBO derivatives. The results obtained by SVM give good statistical results compared to those given by multiple linear regressions and artificial neural networks. The contribution of each descriptor to structure-activity relationships was evaluated. It indicates the importance of the hydrophobic parameter. The proposed method can be successfully used to predict the anti-HIV of TIBO derivatives with only four molecular descriptors which can be calculated directly from molecular structure alone.

1. Introduction

Among infectious diseases, Acquired ImmunoDeficiency Syndrome (AIDS) is the most fatal disorder for which no curative chemotherapy has been developed so far [1]. This infection targets cells of the immune system expressing the CD4 receptors and leads to defects in cell-mediated immunity [2]. After severe depletion of immuno-competent cells, the ultimate phase is the appearance of opportunistic infections, neurologic and neoplastic diseases, and ultimately the death.

Theoretically, an anti-HIV agent may exert its activity by inhibiting a variety of steps in the life cycle of the virus. However, medicinal chemists have focused their attention predominantly on the following stages: Viral binding to target cells, Virus cell fusion, Virus uncoating, Reverse transcription of genomic RNA, Viral integration, Gene expression, Cleavage event and finally Virion maturation, by hitting any of these stages, the viral replication can be terminated [3].

In order to replicate, HIV-1, the causative agent of AIDS, converts its RNA into pro-viral DNA. The viral reverse transcriptase (RT) enzyme catalyzes this reaction. Therefore, the inhibition of this key biochemical event in the viral life cycle provides the most attractive target for anti-HIV drug development. One class of RT inhibitors is the Nucleoside analogues (NRTIs) like AZT, DDI, DDC and D4T. These dideoxy compounds are incorporated during reverse transcription and result in the termination of viral DNA synthesis. A second class of RT inhibitors is the Non-Nucleoside Inhibitors (NNRTIs). They function by binding directly to the enzyme, inhibiting catalysis without blocking substrate binding. These NNRTIs, as demonstrated with some representatives [4], are able to completely suppress virus replication in cell cultures for at least 3 months (and probably longer) [5]. Coming from 1-[2-hydroxyethoxy-methyl]-6-(phenylthio)thymine] (HEPT) and TIBO derivatives, more than 30 different classes of molecules have been identified as NNRTIs. These molecules specifically inhibit HIV type 1 because of a specific interaction with the reverse transcriptase of the virus [6].

The quantitative structure-activity relationship (QSAR) approach became very useful and largely widespread for the prediction of biological activities, particularly in drug design. This approach is based on the assumption that the variations in the properties of the compounds can be correlated with changes in their molecular features, characterized by the so-called "molecular descriptors". Many different techniques for QSAR modeling have been found useful for the establishment of the relationships between molecular structures and anti-HIV activity [7–14]. A certain number of computational techniques have been found

* Corresponding author.
E-mail address: cherqaoui@ucam.ac.ma (D. Cherqaoui).

0223-5234/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2010.01.002

useful for the establishment of the relationships between molecular structures and anti-HIV activity such as Multiple Linear Regression (MLR), partial least square regression and different types of Artificial Neural Networks (ANN) [15]. For these methods, linear model is much limited for a complex biological system. The flexibility of ANN enables them to discover more complex nonlinear relationships in experimental data. However, these neural systems have some problems inherent to its architecture such as over training, overfitting and network optimization. Other problems with the use of ANN concern the reproducibility of results, due largely to random initialization of the networks and variation of stopping criteria. Owing to the reasons mentioned above, there is a growing interest in the application of SVM in the field of QSAR.

An SVM is a supervised learning technique from the field of machine learning applicable to both classification and regression. The SVM is a relatively recent approach introduced by Vapnik [16] and Burges [17] in order to solve supervised classification and regression problems, or more colloquially learning from examples.

The application of SVM has been appeared in several areas of chemistry and biology. Recently they have been successfully used to establish models of structure–molecular properties, such as cancer diagnosis [18–21], identification of HIV protease cleavages sites [22], protein class prediction [23], etc. They have also been applied to the prediction of retention index of protein [24] and the investigation of QSAR studies [25–27].

The aim of this work is to provide an application of SVM to the structure-anti-HIV-1 activity relationship of TIBO derivatives. The results obtained will be compared to those given by MLR and ANN. Thereafter, we sought to measure the contribution of each descriptor to the structure-anti-HIV-1 activity relationship.

2. Methodology

2.1. Support vector machines

SVM is gaining popularity due to many attractive features and promising empirical performance. It originated from early concepts developed by Cortes and Vapnik [28]. This method has proven to be very effective for addressing general purpose classification and regression problems.

The main advantage of SVM is that it adopts the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle [17], employed by conventional neural networks. SRM minimizes an upper bound of the generalization error on Vapnik–Chernoverkis dimension ("generalization error"), as opposed to ERM that minimizes the training error. So SVM is usually less vulnerable to the overfitting problem. Since various introductions into SVM were already stated before [29], only the main ideas about SVM are given in this paper.

2.2. Theory of SVM for regression

SVM can be applied to regression problems by the introduction of an alternative loss function that is modified to include a distance measure. Considering the problem of approximating the set of data $G = \{(x_i, d_i)\}_{i=1}^{n}$ (x_i is the input vector, d_i is the desired value, and n is the total number of data patterns). In SVM method, the regression function is approximated, in a feature space F, by the following function:

$$f(x) = w^T \Phi(x_i) + b \tag{1}$$

Where w is a vector in F and $\Phi(x_i)$ maps the input x to a vector in F. The coefficients w and b are estimated by minimizing the regularized risk function, as shown in Eq. (2):

$$R(C) = \frac{1}{2} w^T w + C \frac{1}{n} \sum_{i=1}^{n} L_\varepsilon(d_i, y_i) \tag{2}$$

Where

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & d - y \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

ε is a prescribed parameter.

In Eq. (2), the first term $(1/2)w^T w = (1/2)\|w\|^2$ is called regularized term. Minimizing $(1/2)\|w\|^2$ will make a function as flat as possible, thus playing role of controlling the function capacity. The second term is the empirical error measured by the ε-insensitive loss function, which is defined by Eq. (3). This defines a ε tube so that if predicted value is within the tube, the loss is zero, while if predicted point is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius ε of the tube. C is penalty parameter, which is a regularized constant to determine the trade-off between training error and model flatness. To get the estimations of w and b, Eq. (2) is transformed to the primal objective Eq. (4) by introducing ξ_i and ξ_i^* (slack variables representing upper and lower constraints on the outputs of the system).

$$R\left(w, \xi_i, \xi_i^*\right) = \frac{1}{2}\|w\|^2 + C \frac{1}{n} \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right) \tag{4}$$

Subject to:

$$\begin{cases} w^T \Phi(x_i) + b - d_i \leq \varepsilon + \xi_i^* \\ d_i - w^T \Phi(x_i) - b \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0, \quad i = 1 \ldots n \end{cases} \tag{5}$$

Thus, decision function (1) becomes the following form:

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + b \tag{6}$$

In Eq. (6), α_i and α_i^* are the introduced Lagrange multipliers. They satisfy the equality $\alpha_i \alpha_i^* = 0$, $\alpha_i \geq 0$, $\alpha_i^* \geq 0$ ($i = 1 \ldots$n) and are obtained by maximizing the dual form of Eq. (4) which has the following form:

$$\begin{aligned} \Phi\left(\alpha_i, \alpha_i^*\right) = & \sum_{i=1}^{n} d_i \left(\alpha_i, \alpha_i^*\right) - \varepsilon \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) \\ & - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right) K(\alpha_i, \alpha_j) \end{aligned} \tag{7}$$

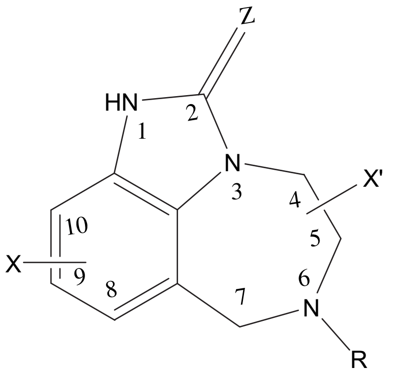

Fig. 1. General structure of TIBO derivatives studied.

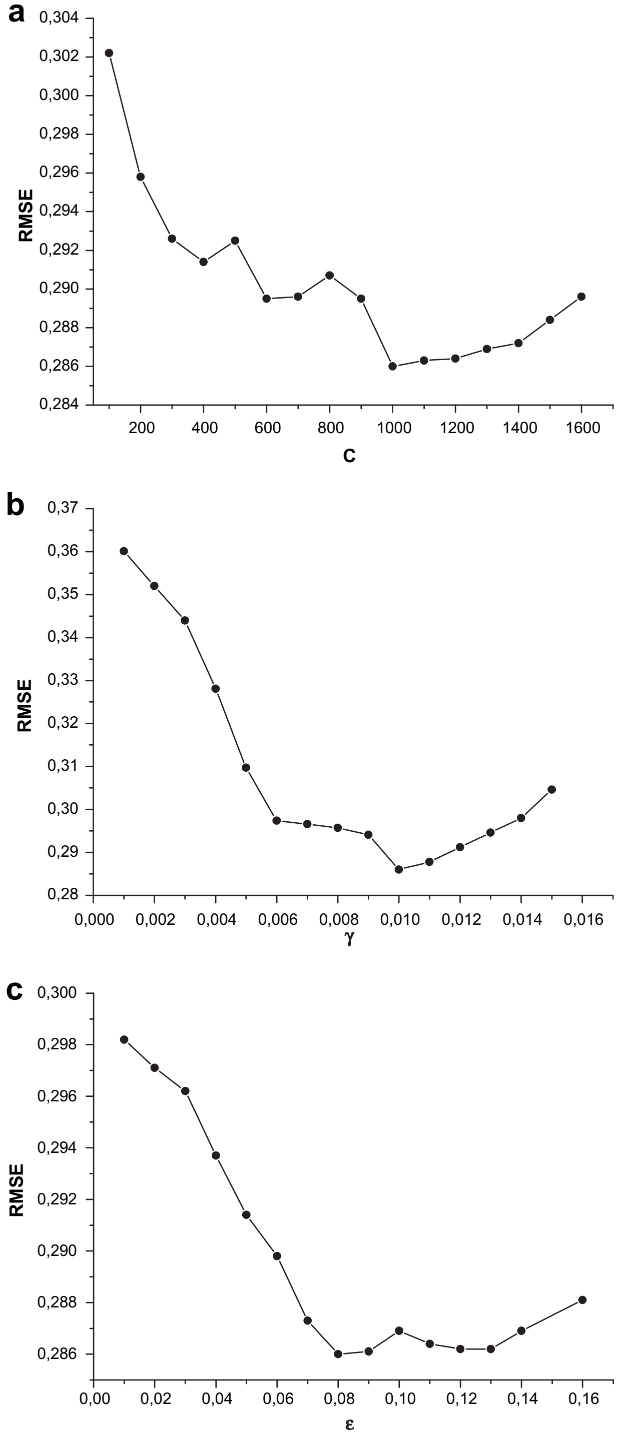

Fig. 2. a. RMSE versus C parameter ($\gamma = 0.01$, $\varepsilon = 0.08$), b. RMSE versus γ parameter ($C = 1000$, $\varepsilon = 0.08$), c. RMSE versus εparameter ($C = 1000$, $\gamma = 0.01$).

Subject to:

$$\begin{cases} \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \quad i = 1 \ldots n \end{cases}$$

Through selecting the appropriate kernel function, the non-linear relation between the building cooling load and its correlative influence parameters based on SVM is established.

Table 1
RMSE and q^2 of SVM, MLR and ANN using LOOCV procedure.

Method	q^2	RMSE
SVM	0.90	0.407
MLR	0.81	0.624
4-5-1	0.83	0.586
4-6-1	0.84	0.567
4-7-1	0.85	0.565
4-8-1	0.84	0.571
4-9-1	0.84	0.576
4-10-1	0.84	0.573
4-11-1	0.84	0.575
4-12-1	0.84	0.570
4-13-1	0.84	0.572

Any function satisfying Mercer's condition [30] can be used as the kernel function, and the typical kernel functions include linear, polynomial, Gaussian and sigmoid functions. Among these functions, the Gaussian function can map the sample set from the input space into a high dimensional feature space effectively, which is good for representing the complex non-linear relationship between the output and input samples. Moreover, there is only one variable (the width parameter) in it needed to be determined, which ensures the high calculation efficiency. Because of the above advantages, the Gaussian function is used widely. In this paper, Gaussian function is also selected as the kernel function, whose expression is shown as follows:

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\gamma^2}\right) \tag{8}$$

where γ is the width parameter.

The overall performance of SVM was evaluated in terms of root-mean-square error (RMSE) which was calculated from the following equation.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n_s} \frac{\left(y_i - \hat{y}_i\right)^2}{n_s}} \tag{9}$$

In this equation y_i is the desired output, $\hat{y}_i$ is the predicted value by model, and n_s is the number of the molecules in data set.

The predictive power of the SVM models, developed on the selected training sets, is estimated on the predictions of an external test set and examined by the statistical parameter q^2:

$$q^2 = 1 - \frac{\sum_{i=1}^{n_s}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_s}\left(y_i - y_m\right)^2} \tag{10}$$

y_m is the mean of dependent variable.

2.3. Software

All calculations of ANN and MLR were carried out, on computer workstation equipped by dual Xeon Processor, using our program written in C language.

All SVM models, in our present study, were implemented using the software LIBSVM for classification and regression developed by Chin-Chang and Chih-Jen Lin [31].

3. Compounds studied and molecular descriptors used

3.1. Compounds studied

A series of 82 TIBO compounds [7] were taken under consideration in this study. All the molecules studied had the same parent

Table 2
Chemical structures of the compounds studied and their anti-HIV-1 activity.

	Substituents				log(1/IC_{50})						
N	X	Z	R	X′	Exp[a]	SVM[b]	Diff	MLR[b]	Diff	ANN[b]	Diff
1	H	S	DMA	5-Me(S)	7.36	7.27	0.09	6.94	0.42	7.27	0.09
2	9-Cl	S	DMA	5-Me(S)	7.47	7.61	−0.14	7.22	0.25	7.20	0.27
3[t]	8-Cl	S	DMA	5-Me(S)	8.37	7.39	0.98	7.29	1.08	7.49	0.88
4	8-F	S	DMA	5-Me(S)	8.24	7.48	0.76	7.03	1.21	7.18	1.06
5	8-SMe	S	DMA	5-Me(S)	8.30	8.38	−0.08	7.50	0.80	7.99	0.31
6[t]	8-OMe	S	DMA	5-Me(S)	7.47	7.21	0.26	7.01	0.46	7.50	−0.03
7	8-OC_2H_5	S	DMA	5-Me(S)	7.02	6.94	0.08	7.16	−0.14	7.01	0.01
8	8-CN	O	DMA	5-Me(S)	5.94	5.86	0.08	5.64	0.30	5.90	0.04
9	8-CN	S	DMA	5-Me(S)	7.25	7.33	−0.08	7.04	0.21	7.20	0.05
10	8-CHO	S	DMA	5-Me(S)	6.73	6.65	0.08	6.95	−0.22	6.72	0.01
11	8-$CONH_2$	O	DMA	5-Me(S)	5.20	5.12	0.08	5.28	−0.08	4.91	0.29
12	8-Br	O	DMA	5-Me(S)	7.33	7.25	0.08	6.12	1.21	7.32	0.01
13	8-Br	S	DMA	5-Me(S)	8.52	8.44	0.08	7.52	1.00	8.46	0.06
14[t]	8-I	O	DMA	5-Me(S)	7.06	7.10	−0.04	6.40	0.66	7.75	−0.69
15	8-I	S	DMA	5-Me(S)	7.32	7.40	−0.08	7.81	−0.49	7.38	−0.06
16	8-C=-CH	O	DMA	5-Me(S)	6.36	6.31	0.05	5.70	0.66	5.83	0.53
17	8-C=-CH	S	DMA	5-Me(S)	7.53	7.50	0.03	7.10	0.43	7.20	0.33
18	8-Me	O	DMA	5-Me(S)	6.00	6.70	−0.70	5.84	0.16	6.09	−0.09
19[t]	8-Me	S	DMA	5-Me(S)	7.87	7.37	0.50	7.25	0.62	7.65	0.22
20	9-NO_2	O	CPM	5-Me(S)	4.48	4.40	0.08	4.02	0.46	4.47	0.01
21	8-NH_2	O	CPM	5-Me(S)	3.07	4.14	−1.07	3.96	−0.89	4.04	−0.97
22[t]	8-NMe_2	O	CPM	5-Me(S)	5.18	4.61	0.57	4.49	0.69	4.33	0.85
23	9-NH_2	O	CPM	5-Me(S)	4.22	4.14	0.08	3.92	0.30	3.95	0.27
24	9-NMe_2	O	CPM	5-Me(S)	5.18	5.26	−0.08	4.48	0.70	4.62	0.56
25	9-NHCOMe	O	CPM	5-Me(S)	3.80	3.88	−0.08	3.92	−0.12	4.04	−0.24
26	9-NO_2	S	CPM	5-Me(S)	5.61	5.53	0.08	5.43	0.18	5.88	−0.27
27	9-F	S	DMA	5-Me(S)	7.60	7.40	0.20	7.00	0.60	7.57	0.03
28	9-CF_3	O	DMA	5-Me(S)	5.23	5.31	−0.08	6.00	−0.77	6.25	−1.02
29	9-CF_3	S	DMA	5-Me(S)	6.31	6.23	0.08	7.41	−1.10	7.21	−0.90
30	10-OMe	O	DMA	5-Me(S)	5.18	5.26	−0.08	5.59	−0.41	5.65	−0.47
31[t]	9,10-di-Cl	S	DMA	5-Me(S)	7.60	7.44	0.16	7.66	−0.06	7.52	0.08
32	10-Br	S	DMA	5-Me(S)	5.97	6.68	−0.71	7.49	−1.52	6.25	−0.28
33	H	O	$CH_2CH{=}CH_2$	5-Me(S)	4.15	4.10	0.05	3.98	0.17	3.98	0.17
34[t]	H	O	2-MA	5-Me(S)	4.33	4.36	−0.03	4.24	0.09	4.51	−0.18
35	H	O	CH_2CO_2Me	5-Me(S)	3.07	3.15	−0.08	3.63	−0.56	3.32	−0.25
36	H	O	$CH_2C{\equiv}CH$	5-Me(S)	3.24	3.16	0.08	3.79	−0.55	3.50	−0.26
37	H	O	CH_2-2-furanyl	5-Me(S)	3.97	4.05	−0.08	4.06	−0.09	4.05	−0.08
38	H	O	$CH_2CH{=}CH_2$[S(+)]	5-Me(S)	4.18	4.10	0.08	3.98	0.20	3.88	0.30
39	H	O	$CH_2CH_2CH{=}CH_2$	5-Me(S)	4.30	4.38	−0.08	4.17	0.13	4.24	0.06
40	H	O	$CH_2CH_2CH_3$	5-Me(S)	4.05	4.14	−0.09	4.08	−0.03	4.09	−0.04
41	H	O	2-MA[S(+)]	5-Me(S)	4.72	4.51	0.21	4.24	0.48	4.35	0.37
42	H	O	CPM	5-Me(S)	4.36	4.29	0.07	4.18	0.18	4.32	0.04
43[t]	H	O	$CH_2CH{=}CHMe$(*E*)	5-Me(S)	4.24	4.29	−0.05	4.19	0.05	4.34	−0.10
44	H	O	$CH_2CH{=}CHMe$(*Z*)	5-Me(S)	4.46	4.38	0.08	4.19	0.27	4.25	0.21
45	H	O	$CH_2CH_2CH_2Me$	5-Me(S)	4.00	4.44	−0.44	4.33	−0.33	4.43	−0.43
46	H	O	DMA	5-Me(S)	4.90	4.82	0.08	5.53	−0.63	5.02	−0.12
47	H	O	$CH_2C(Br){=}CH_2$	5-Me(S)	4.21	4.26	−0.05	4.12	0.09	4.20	0.01
48	H	O	$CH_2C(Me){=}CHMe$(*E*)	5-Me(S)	4.54	4.54	0.00	4.38	0.16	4.49	0.05
49	H	O	DMA[*R*(+)]	5-Me(S)	4.66	4.82	−0.16	5.53	−0.87	5.02	−0.36
50	H	O	DMA[*S*(+)]	5-Me(S)	5.4	4.82	0.58	5.53	−0.13	5.22	0.18
51	H	O	$CH_2C(C_2H_5){=}CH_2$	5-Me(S)	4.43	4.51	−0.08	4.39	0.04	4.49	−0.06
52	H	O	$CH_2CH{=}CHC_6H_5$(*Z*)	5-Me(S)	3.91	3.99	−0.08	4.92	−1.01	4.70	−0.79
53[t]	H	O	$CH_2C(CH{=}CH_2){=}CH_2$	5-Me(S)	4.15	4.32	−0.17	4.21	−0.06	4.42	−0.27
54	8-Cl	S	DMA	H	7.34	7.42	−0.08	7.05	0.29	7.18	0.16
55	9-Cl	S	DMA	H	6.8	6.88	−0.08	6.98	−0.18	7.15	−0.35
56	H	O	2-MA	5,5-di-Me	4.64	4.69	−0.05	4.42	0.22	4.52	0.12
57[t]	H	O	2-MA	4-Me	4.5	4.34	0.16	4.22	0.28	4.34	0.16
58	9-Cl	S	2-MA	4-Me(S)	6.17	6.09	0.08	5.91	0.26	6.00	0.17
59[t]	9-Cl	S	CPM	4-Me(R)	5.66	5.84	−0.18	5.83	−0.17	5.69	−0.03
60	H	O	C_3H_7	4-$CHMe_2$	4.13	4.05	0.08	4.56	−0.43	4.17	−0.04
61	H	O	2-MA	4-$CHMe_2$	4.9	4.82	0.08	4.75	0.15	4.65	0.25
62	H	O	2-MA	4-C_3H_7	4.32	4.40	−0.08	4.72	−0.40	4.64	−0.32
63	H	O	DMA	7-Me	4.92	5.00	−0.08	5.57	−0.65	5.42	−0.50
64	8-Cl	O	DMA	7-Me	6.84	6.11	0.73	5.91	0.93	6.25	0.59
65	9-Cl	O	DMA	7-Me	6.8	6.72	0.08	5.86	0.94	6.10	0.70
66	H	S	C_3H_7	7-Me	5.61	5.53	0.08	5.52	0.09	5.59	0.02
67	H	S	DMA	7-Me	7.11	7.19	−0.08	6.98	0.13	7.18	−0.07
68[t]	8-Cl	S	DMA	7-Me	7.92	7.39	0.53	7.31	0.61	7.56	0.36
69	9-Cl	S	DMA	7-Me	7.64	7.33	0.31	7.26	0.38	7.21	0.43
70	H	O	DMA	4,5-di-Me(*cis*)	4.25	4.33	−0.08	5.84	−1.59	4.22	0.03
71	H	S	DMA	4,5-di-Me(*cis*)	5.65	5.73	−0.08	7.23	−1.58	7.23	−1.58
72[t]	H	S	CPM	4,5-di-Me(*trans*)	4.87	5.59	−0.72	5.89	−1.02	5.51	−0.64
73	H	S	DMA	5,7-di-Me(*trans*)	7.38	6.70	0.68	7.20	0.18	7.22	0.16

(*continued on next page*)

Table 2 (*continued*)

	Substituents				$\log(1/IC_{50})$						
N	X	Z	R	X′	Exp[a]	SVM[b]	Diff	MLR[b]	Diff	ANN[b]	Diff
74	H	S	DMA	5,7-di-Me(*cis*)	5.94	6.70	−0.76	7.20	−1.26	7.22	−1.28
75	9-Cl	O	DMA	5,7-di-Me(R,R-*rans*)	6.64	6.72	−0.08	6.09	0.55	6.51	0.13
76[t]	9-Cl	S	DMA	5,7-di-Me(R,R-*rans*)	6.32	7.44	−1.12	7.49	−1.17	7.79	−1.47
77	9-Cl	O	DMA	5-Me(S)	6.74	6.66	0.08	5.82	0.92	5.96	0.78
78	H	O	C_3H_7	5-Me	4.22	4.14	0.08	4.08	0.14	4.09	0.13
79[t]	H	S	C_3H_7	5-Me	5.78	5.96	−0.18	5.49	0.29	5.40	0.38
80	H	O	2-MA	5-Me	4.46	4.51	−0.05	4.24	0.22	4.35	0.11
81	H	S	DMA	5-Me	7.01	7.27	−0.26	6.94	0.07	7.17	−0.16
82[t]	H	O	DMA	5-Me(S)	5.48	5.33	0.15	5.53	−0.05	5.02	0.46

[a] Experimental activity.
[b] Predicted activity by SVM, MLR and ANN, respectively.
[t] Test set.

skeleton (Fig. 1). The structures and anti-HIV-1 activities of these compounds were described previously [7]. The anti-HIV activity of the compounds has been expressed by the compound's ability to protect MT-4 cells against the cytopathic effect of the virus. The concentration of the compound leading to 50% effect has been measured and expressed as IC_{50}. The logarithm of the inverse of this parameter has been used as biological end points ($\log 1/IC_{50}$) in the QSAR studies.

3.2. Molecular descriptors used

A molecular descriptor is a numerical representation of the structure which describes a motif within the structure or the structure itself as a whole. It can be obtained in either empirical or nonempirical ways. In QSAR models established for the TIBO derivatives [7], various molecular descriptors were cited but the hydrophobic parameter log *P* is often used. Garg et al. [7] published an extensive QSAR study on TIBO derivatives and suggested that a hydrophobic X substituent will be beneficial to the activity and that it would be more advantageous if it is at the 8-position. This position shows also to have a steric effect. In order to take into account this effect, $B1(8-x)$ (Verloop's sterimol parameter of the X substituent at the position 8) is used as the molecular descriptor. They also suggested that at position 2Z = Sulfur was more favorable than Z = Oxygen. The indicator variable I_Z is used to account for this variation with a value of unity for the former and zero for the latter. They added that at position 6 a 3,3-dimethylallyl group would be preferred. This group is represented by the indicator variable I_R. Thus, in the present study, each molecule was described by these 4 descriptors, which are given by Garg et al. [7]. They characterize the hydrophobic, the steric and the electronic aspects, as listed below:

log *P*: the calculated octanol/water partition coefficient of the molecule.
$B1(8-x)$: Verloop's sterimol parameter (width parameter of the X substituent at the position 8).
$I_R = 1$ if $R =$ 3,3-dimethyallyl and $I_R = 0$ for others (see Fig. 1).
$I_Z = 1$ if Z = Sulfur and $I_Z = 0$ if Z = Oxygen (see Fig. 1).

4. Results and discussion

In order to develop QSAR models of anti-HIV activity of TIBO derivatives, the most commonly practiced stages (computation, prediction and the descriptor's contribution) have been achieved: The first one was aimed at selecting the parameters of the SVM. The second one was aimed at determining the predictive ability of the SVM. In the third session, we attempt an evaluation of the importance of the descriptors used.

4.1. Computation

Similar with other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter *C*, ε of ε-insensitive loss function, the kernel type $K(x,y)$ and its corresponding parameters. *C* is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If *C* is too small then insufficient stress will be placed on fitting the training data. If *C* is too large then the algorithm will overfit the training data.

The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of ε is critical from theory.

It is well known that the results of SVM approach lie largely on the choice of a kernel, which determines the sample distribution in the mapping space. Therefore the kernel functions should be decided first. One has several possibilities for the choice of this function, including linear, polynomial, sigmoid, and radial basis function. For regression tasks, a commonly used kernel function is the Gaussian radial basis function (RBF) [32] due to its good general performance. The RBF is formulated as:

$$\exp\left(-\gamma\|\mu-\nu\|^2\right)$$

Where γ is the parameter of the kernel, μ and ν are two independent variables, γ controls the amplitude of the Gaussian function and therefore, controls the generalization ability of SVM. We have to optimize γ and find the optimal one.

The support vector regression (SVR) algorithm includes three parameters to be optimized, ε in the ε-insensitive loss function, the regularization constant *C* and the Gaussian function parameter γ. To determine the optimal parameters, a grid search was performed based on leave-one-out cross-validation (LOOCV) on the original training set for all parameter combinations of *C* from 100 to 1600 with incremental steps of 100, γ ranging from 0.001 to 0.016 with incremental steps of 0.001 and ε from 0.01 to 0.16 with incremental steps of 0.01. Because the grid search was performed over three parameters, the cross-validation errors RMSE could not be shown in one plot to be visualized easily. Fig. 2a–c show the influence of each parameter with the other two fixed to the optimal values on the model performance. One observes from Fig. 2a that the SVR model reaches the best performance when $C = 1000$ (with $\gamma = 0.01$ and $\varepsilon = 0.08$). The curve of RMSE versus γ is depicted in Fig. 2b. It is observed that the best SVR model is

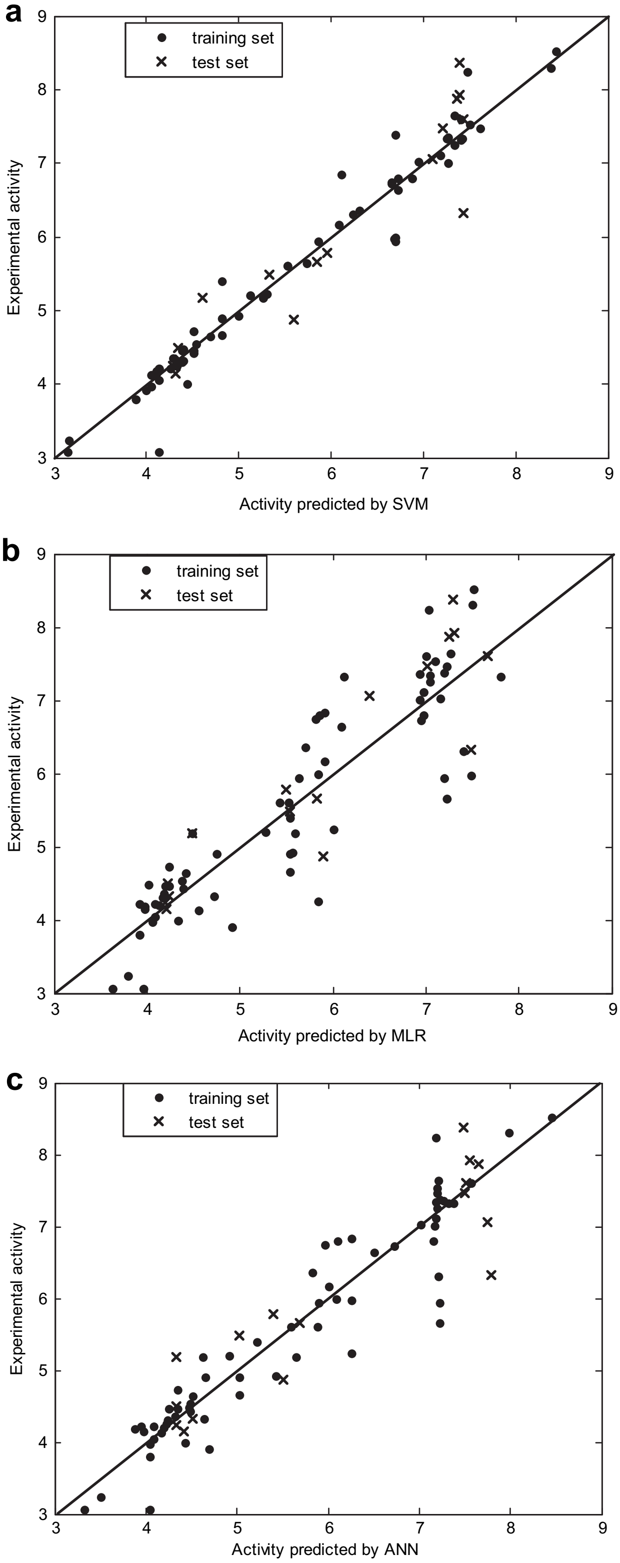

Fig. 3. a. log (1/IC$_{50}$) observed experimentally versus log (1/IC$_{50}$) predicted by SVM, b. log (1/IC$_{50}$) observed experimentally versus log (1/IC$_{50}$) predicted by MLR, c. log (1/IC$_{50}$) observed experimentally versus log (1/IC$_{50}$) predicted by ANN.

obtained with γ equaling to 0.01 (with $C = 1000$ and $\varepsilon = 0.08$). The influence of ε on the performance of SVR is shown in Fig. 2c. One sees that the optimal value of ε is 0.08 (with $C = 1000$ and $\gamma = 0.01$).

Table 3
Statistical parameters of different constructed QSAR models.

	Training test		Test set	
	R^2	RMSE	R^2	RMSE
SVM	0.96	0.286	0.89	0.489
(4-4-1) ANN	0.90	0.449	0.85	0.571
MLR	0.80	0.632	0.83	0.595

4.2. Prediction

Validation of the QSAR models was required to test the prediction and the generalization of the methods. Any model needs to be validated before it is used for "understanding" or predicting new events. Tropsha et al. [33] have addressed this problem by providing a set of guidelines for developing or/and assessing QSAR models. In order to be reliable and predictive, QSAR models should: (1) be statistically significant and robust, (2) be validated by making accurate predictions for external data sets that were not used in the model development, and (3) have their application boundaries defined.

For the present work, the proposed methodology was validated using several strategies: internal validation, Y-randomization and external validation using division of the entire data set into training and test sets. Furthermore, the domain of applicability which indicates the area of reliable predictions was defined.

4.2.1. Internal validation

LOOCV procedure was widely used to evaluate the internal validation of QSAR models. As the name suggests, LOOCV procedure involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

In our previous study [34] ANN method was applied to the same data set and the same four molecular descriptors. Using LOOCV procedure, nine ANN architectures of 4-x-1 ($x = 5$–13, x represents the number of hidden neurons) have been tested. The results of QSAR done by these ANN architectures, by MLR analysis and by SVM method are listed in Table 1. The quality of the fitting is estimated by the RMSE and by the statistical parameter q^2. As it can be seen in Table 1, high correlation coefficient ($q^2 = 0.90$) and low RMSE = 0.407 have been obtained by means of the SVM. According to this table, it is clear that the performance of SVM is better than those obtained by ANN and MLR techniques. Indeed, in every case, the SVM's correlation coefficient is greater and its standard deviation is lower than those of the ANN and MLR.

4.2.2. External validation

In their highly cited publication "Beware of q^2!", Golbraikh and Tropsha [35] have demonstrated the insufficiency of the training set statistics for developing externally predictive QSAR models and formulated the main principles of model validation. The purpose of the SVM model generation is not just to predict the training set, but also to verify whether the SVM model is capable of predicting external set. In many cases, more new chemicals being unavailable for prediction purpose, the original data set is divided into a training set and a test set. In the present work, the whole data set (82 compounds) was divided into a training set (66 compounds) for model development and a test set (16 compounds) for external prediction. The test set is selected such that each of its members is close to at least one point of the training set.

Table 4
Results of randomization test of the developed models.

Modeling technique	R^2 from non-random model	Mean value of R^2 from model trials
SVM	0.96	0.18
MLR	0.80	0.25
ANN	0.90	0.22

External validation for SVM, MLR and ANN was performed. Using the 66 compounds of the training set, several ANN architectures were tried. Among all these architecture, the best one is 4-4-1.

The predicted values by the SVM, MLR and ANN models are given in Table 2. The plot of predicted versus experimental values for data set is shown in Fig. 3a (SVM), Fig. 3b (MLR) and Fig. 3c (4-4-1 ANN). Among all these figures, the first one shows that the log(1/IC_{50}) values calculated by the SVM are very close to the experimental ones. The statistical parameters of the three models are shown in Table 3. As can be seen from this table, the statistical parameters of SVM model are better than the other ones.

4.2.3. Y-randomization test

A widely used approach to establish the robustness of a given QSAR model is the so-called Y-randomization [33]. In this approach, dependent variable vector $(y = \log(1/IC_{50}))$ is randomly shuffled and a new QSAR model is built using the original independent variables. If the new QSAR models have lower R^2 values for several trials, then the given QSAR model is thought to be robust.

In this work, ten random shuffles of the y vector were performed for SVM, MLR and ANN. The results are shown in Table 4. For each technique, the mean value of random models is significantly lower than the corresponding value of the non-random model. This suggests that the models are not obtained by chance.

4.2.4. Domain of applicability of the model

The domain of application [33] of a QSAR model must be defined if the model is to be used for screening new compounds. Predictions for only those compounds that fall into this domain may be considered reliable. Extent of Extrapolation [33] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i for each chemical, for which QSAR model is used to predict its activity:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \quad i = 1 \ldots n$$

Where x_i is the descriptor row-vector of the query compound, and X is the $n \times k - 1$ matrix of k model descriptor values for n training set compounds. The superscript T refers to the transpose of the matrix X and the vector x_i. The warning leverage h^* is, generally, fixed at $3k/n$, where n is the number of training compounds, and k is the number of model parameters.

Table 5
Leverage for the test set.

Molecule number	Leverage (h)	Molecule number	Leverage (h)
3	0.0448	53	0.0349
6	0.0542	57	0.0352
14	0.0922	59	0.0112
19	0.0440	68	0.0443
22	0.0468	72	0.0180
31	0.0606	76	0.0522
34	0.0356	79	0.0981
43	0.0360	82	0.0550

$h^* = 0.1818$.

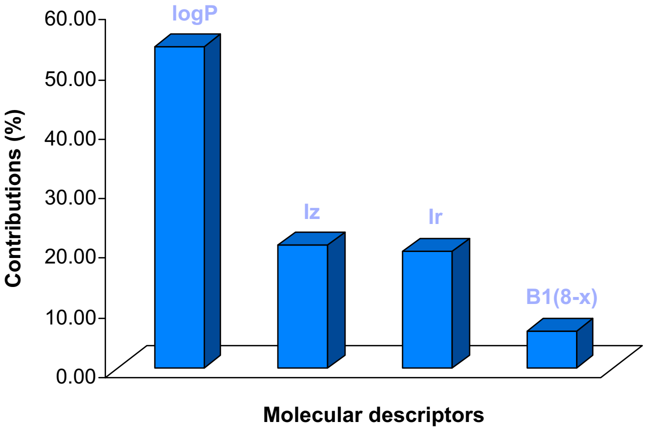

Fig. 4. Contributions of the 4 molecular descriptors to QSAR.

The leverage values for test set (16 compounds) were computed and are presented in Table 5. As seen in this table, the leverage of all compound in the test set are lower than h^* for all models. This means that all predicted values are acceptable.

4.3. Descriptor's contribution

One of the major goals of QSAR studies is the determination of the factors influencing the activity of the studied compounds. They contribute to the comprehension of modes of action of composed on their biological targets and guide the synthesis towards compounds with optimal activity. We thus saw necessary to evaluate there contribution of the molecular descriptors to the model established by the SVM. The contribution of each descriptor to the establishment of the QSAR was estimated from the trained SVM using a technique proposed by Cherqaoui et al. [36]. Fig. 4 indicates that the relative importance of the descriptors varied in the following order: $\log P > I_Z > I_R > B1(8 - x)$. The same classification was found in our previous work using the ANN [34].

The descriptor related to the hydrophobic property is the most important in the establishment of the QSAR of TIBO derivatives. Numerous pharmacological and biochemical process studies show that the interaction of organic compounds with the biological systems depends strongly on their hydrophobicity [37]. Clearly, a drug molecule must interact with several domains of macromolecules and membranes of different degrees of hydrophobicity to reach its target site [38,39]. Descriptors I_Z, I_R and $B1(8 - x)$ seem to be important in the establishment of the structure-anti-HIV-1 activity relationships. So, the inhibitory activity of TIBO is also governed by electronic (I_Z) and steric effects (I_R and $B1(8 - x)$).

5. Conclusion

Support vector machines were applied to build up the QSAR model for predicting the anti-HIV-1 activity of TIBO derivatives of 82 compounds, based on descriptors calculated from molecular structure. The results obtained show that the SVM technique was able to establish a satisfactory relationship between the molecular descriptors and the anti-HIV-1 activity. It has been shown in this study that SVM give a superior performance to those given by MLR and ANN techniques. The main factor controlling the anti-HIV activity of TIBO derivatives has been determined by SVM. Hydrophobicity of the compounds was thus found to take the most relevant part in the molecular description. The SVM is a very promising machine learning technique from many aspects and will gain more extensive application. Furthermore, the proposed approach can also be extended in other QSAR investigation.

Abbreviations

AIDS	Acquired immunodeficiency syndrome
ANN	Artificial Neural networks
AZT	3′-azido-2′,3′-dideoxy-thymidine (Zidovudine)
D4T	2′,3′-didehydro-2′,3′-dideoxy-thymidine (Stavudine)
DDC	2′,3′-dideoxycytidine (Zalcitabine)
DDI	2′,3′-dideoxyinosine (Didanosine)
ERM	Empirical risk minimization
HEPT	1-[2-hydroxyethoxy-methyl]-6-(phenylthio) thymine]
HIV-1	Human immunodeficiency virus type 1
LOOCV	Leave-one-out cross-validation
MLR	Multiple linear regression
NNRTIs	Non-Nucleoside Reverse Transcriptase Inhibitors
NRTIs	Nucleoside Reverse Transcriptase Inhibitors
QSAR	Quantitative structure-activity relationships
RBF	Radial basis function
RMSE	Root-mean-square error
RT	Reverse transcriptase
SRM	Structure risk minimization
SVM	Support Vector Machines
SVR	Support vector regression
TIBO	Tetrahydroimidazo [4,5,1-jk][1,4] benzodiazepinone

References

[1] A.L. Dunne, H. Siregar, J. Mills, S.M. Crowe, HIV replication in chronically infected macrophages is not inhibited by the Tat inhibitors Ro-5-3335 and Ro-24-7429. J. Leukoc. Biol. 56 (1994) 369–373.

[2] Y.J. Rosenberg, A.O. Anderson, R. Pabst, HIV-induced decline in blood CD4/CD8 ratios: viral killing or altered lymphocyte trafficking. Immunol. Today 19 (1998) 10–17.

[3] E. De Clercq, Toward improved anti-HIV chemotherapy: therapeutic strategies for intervention with HIV infection. J. Med. Chem. 38 (1995) 2491–2517.

[4] J. Balzarini, A. Karlsson, M.J. Camarasa, E. De Clercq, Knocking-out concentrations of HIV-1-specific inhibitors completely suppress HIV-1 infection and prevent the emergence of drug-resistant virus. Virology 96 (1993) 576–585.

[5] M.B. Vasudevachari, C. Battista, H.C. Lane, M.C. Psallidopoulos, B. Zhao, J. Cook, J.R. Palmer, D.L. Romero, W.G. Tarpley, N.P. Salzman, Prevention of the spread of HIV-1 infection with nonnucleoside reverse transcriptase inhibitors. Virology 190 (1992) 269–277.

[6] R. Pauwels, K. Andries, J. Desmyter, D. Schols, M.J. Kukla, H.J. Breslin, A. Raeymaechers, J.V. Gelder, R. Woestenborgs, J. Heykants, K. Schellekens, M.A.C. Janssen, E. DeClercq, P.A.J. Janssen, Potent and Selective Inhibition of HIV-1 replication in vitro by a novel series of TIBO derivatives. Nature 343 (1990) 470–474.

[7] R. Garg, S.P. Gupta, H. Gao, M.S. Babu, A.K. Debnath, Comparative quantitative structure-activity relationship studies on anti-HIV drugs. Chem. Rev. 99 (1999) 3525–3601.

[8] M.A. Sattwa, R. Kunal, Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives. Eur. J. Med. Chem. 44 (4) (2009) 1509–1524.

[9] B. Hemmateenejad, S.M. Tabaei, F. Namvaran, Computer-aided design of potential anti-HIV-1 non-nucleoside reverse transcriptase inhibitors by contraction of β-ring in TIBO derivatives. J. Mol. Struct. Theochem. 732 (1–3) (2005) 39–45.

[10] J. Huuskonen, QSAR modeling with the electrotopological state: TIBO derivatives. J. Chem. Inf. Comput. Sci. 41 (2001) 425–429.

[11] H.H. Maw, L.H. Hall, E-state modeling of HIV-1 protease inhibitor binding independent of 3D information. J. Chem. Inf. Comput. Sci. 42 (2002) 290–298.

[12] E. Gancia, G. Bravi, P. Mascagni, A. Zaliani, Global 3D-QSAR methods: MS-WHIM and autocorrelation. J. Comput. Aided Mol. Des. 14 (2000) 293–306.

[13] C. Klein, L. Lawtrakul, S. Hannongbua, P. Wolschann, Accessible charges in structure-activity relationships: a study of HEPT-based HIV-1 RT inhibitors. Sci. Pharm. 68 (2000) 25–40.

[14] R. Kiralj, M. Ferreira, A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors. I. The chemometric approach. J. Mol. Graph. Model. 21 (5) (2003) 435–448.

[15] J. Zupan, J. Gasteiger, Neural Networks for Chemists. An Introduction. Wiley-VCH, Weinheim, 1993.

[16] V.N. Vapnik, The Nature of Statistical Learning Theory. Springer, Berlin, 1995.

[17] J.C. Burges, A tutorial on support vector machines for pattern recognition. Data Min. Know. Discov. 2 (1998) 121–167.

[18] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schumme, D. Haussle, Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16 (2000) 906–914.

[19] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines. Mach. Learn. 46 (2002) 389–422.

[20] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggi, W. Gerald, M. Loda, E.S. Lander, T.R. Golub, Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 15149–15154.

[21] H.X. Liu, R.S. Zhang, F. Luan, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Diagnosing breast cancer based on support vector machines. J. Chem. Inf. Comput. Sci. 43 (2003) 900–907.

[22] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for predicting HIV protease cleavage sites in protein. J. Comput. Chem. 23 (2002) 267–274.

[23] S.J. Hua, Z.R. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308 (2001) 397–407.

[24] U. Norinder, Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. Neurocomputing 55 (2003) 337–346.

[25] R. Hu, J.P. Doucet, M. Delamar, R. Zhang, QSAR models for 2-amino-6-arylsulfonylbenzonitriles and congeners HIV-1 reverse transcriptase inhibitors based on linear and nonlinear regression methods. Eur. J. Med. Chem. 44 (5) (2009) 2158–2171.

[26] I. Massarelli, M. Imbriani, A. Coi, M. Saraceno, N. Carli, A.M. Bianucci, Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals. Eur. J. Med. Chem. 44 (9) (2009) 3658–3664.

[27] C.Y. Zhao, H.X. Zhang, X.Y. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology 217 (2006) 105–119.

[28] C. Cortes, V. Vapnik, Support-vector networks. Mach. Learn. 20 (1995) 273–297.

[29] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines. Cambridge University Press, Cambridge UK, 2000.

[30] J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations. Philos. Trans. Roy. Soc. London, Ser. A 209 (1909) 415–446.

[31] C.C. Chang, C.J. Lin, LIBSVM-A Library for support vector machine.http://www/csie.edu/tw/cjlin/libs/libsvm.

[32] C. Nianyi, L. Wencong, Y. Jie, L. Guozheng, Support Vector Machine in Chemistry. World Scientific Publishing Company, New York, 2004.

[33] A. Tropsha, P. Gramatica, V. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. Quant. Struct. Act. Relat. 22 (2003) 69–77.

[34] L. Douali, D. Villemin, D. Cherqaoui, Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by neural networks: TIBO derivatives. Int. J. Mol. Sci. 5 (2004) 48–55.

[35] A. Golbraikh, A. Tropsha, Beware of q^2! J. Mol. Graph. Model. 20 (2002) 269–276.

[36] D. Cherqaoui, M. Esseffar, D. Villemin, J.M. Cense, M. Chastrette, D. Zakarya, Structure musk odour relationships studies of tetralin and indan compounds using neural networks. N. J. Chem. 22 (1998) 839–843.

[37] C. Hansch, A. Leo, Exploring QSAR, Fundamentals and Applications in Chemistry and Biology. ACS, Washington DC, 1995.

[38] C. Hansch, J.P. Björkoth, A. Leo, Hydrophobicity and central nervous system agents: on the principle of minimal hydrophobicity in drug design. J. Pharm. Sci. 76 (1987) 663–687.

[39] J.W. McFarland, On the parabolic relationship between drug potency and hydrophobicity. J. Med. Chem. 13 (1970) 1192–1196.