Control Theory and Informatics
ISSN 2224-5774 (Paper) ISSN 2225-0492 (Online)
Vol.3, No.3, 2013

Pattern recognition system based on support vector machines: HIV-1 integrase inhibitors application

Rachid Darnag[1*] Brahim Minaoui[2] Mohamed Fakir[3]
1. Département de Physique, Laboratoire de Traitement de l'Information et de Télécommunication, Faculté des
Sciences et Technique BP 523 Université Sultan Moulay Slimane, Béni-Mellal, Morocco,
2. Département de Physique, Laboratoire de Traitement de l'Information et de Télécommunication, Faculté des
Sciences et Technique BP 523 Université Sultan Moulay Slimane, Béni-Mellal, Morocco
3. Département de l'Informatique, Faculté des Sciences et Technique BP 523 Université Sultan Moulay Slimane, Béni-Mellal, Morocco
* E-mail of the corresponding author:r.darnag@gmail.com

Abstract
Support Vector Machines (SVM) represent one of the most promising Machine Learning (ML) tools that can be applied to develop a predictive Quantitative Structure-Activity Relationship (QSAR) models using molecular descriptors. The performance and predictive power of support vector machines (SVM) for regression problems in quantitative structure-activity relationship were investigated. The SVM results are superior to those obtained by artificial neural network and multiple linear regression. These results indicate that the SVM model with the kernel radial basis function can be used as an alternative tool for regression problems in quantitative structure-activity relationship.
Keywords: Support Vector Machines; Artificial Neural Network; Quantitative Structure-Activity Relationship.

Introduction

The Quantitative Structure-Activity Relationship (QSAR) approach became very useful and largely widespread for the prediction of anti-HIV activity, 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". A certain number of computational techniques have been found useful for the establishment of the relationships between molecular structures and anti-HIV activity such as Multiple Linear Regression (MLR), Partial Least Square regression (PLS) and different types of Artificial Neural Networks (ANN) [1]. For these methods, linear model is 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, over fitting 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. The SVM is a relatively recent approach introduced by Vapnik [2] and Burges [3] in order to solve supervised classification and regression problems, or more colloquially learning from examples.

SVM have strong theoretical foundations and excellent empirical successes. They have been applied to tasks such as handwritten digit recognition, object recognition, text classification, cancer diagnosis [4,5], identification of HIV protease cleavages sites[6]. They have also been applied to the prediction of retention index of protein and the investigation of QSAR studies.

The acquired immunodeficiency syndrome (AIDS)[7] has been spreading continuously since it was first reported in 1981, and becomes one of the most hazardous diseases.[8] The human immunodeficiency virus type 1 (HIV-1),[9] first isolated from a patient with generalized lymphadenopathy in 1983, has been found to be the pathogenic retrovirus and causative agent of AIDS epidemic.[10]. The number of HIV infected subjects keeps alarmingly on the rise.[11] Considerable attention has been paid to understand the viral life cycle and the functional nuances of nine genes encoded by HIV- 1.[12] The protease (PR), reverse transcriptase (RT), and integrase [13] are regarded as the key enzymes in the duplication of HIV-1, thus structure-assisted design for these targets based on the knowledge of their three-dimensional structures may play a critical role in the discovery of novel anti-HIV drugs.

In the present paper, we present the applications of support vector regression (SVR) to investigate the relationship between structure and activity of 1,3,4-oxadiazole substituted naphthyridine derivatives based on molecular descriptors. The performance and predictive capability of support vector machines method are investigated and compared with other methods such as artificial neural network and multiple linear regression methods.

Methodology

Support vector machines

A SVM is a supervised learning technique from the field of machine learning applicable to both classification and regression. SVM developed by Cortes and Vapnik [14], as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance.

Originally it was worked out for linear two-class classification with margin, where margin means the minimal distance from the separating hyper plane to the closest data points. SVM learning machine seeks for an optimal separating hyper-plane, where the margin is maximal. An important and unique feature of this approach is that the solution is based only on those data points, which are at the margin. These points are called support vectors. The linear SVM can be extended to nonlinear one when first the problem is transformed into a feature space using a set of nonlinear basis functions. In the feature space which can be very high dimensional, the data points can be separated linearly. An important advantage of the SVM is that it is not necessary to implement this transformation and to determine the separating hyper-plane in the possibly very-high dimensional feature space, instead a kernel representation can be used, where the solution is written as a weighted sum of the values of certain kernel function evaluated at the support vectors.

All SVM model in our present study were implemented using the software Libsvm that is an efficient software for classification and regression developed by Chin-Chang and Chih-Jen Lin [15].

Artificial neural networks

ANN are artificial systems simulating the function of the human brain. Three components constitute a neural network: the processing elements or nodes, the topology of the connections between the nodes, and the learning rule by which new information is encoded in the network. While there are a number of different ANN models, the most frequently used type of ANN in QSAR is the three-layered feed-forward network [16]. In this type of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer.

According to the supervised learning adopted, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A back-propagation algorithm is used to minimize the error function. This algorithm has been described previously with a simple example of application [17] and a detail of this algorithm is given elsewhere [18].

Data set

In this QSAR study, biological and chemical data from 67 of 1,3,4- oxadiazole substituted naphthyridine derivatives were used, which have been presented in the work of Johns et al. [19]. HIV-1 integrase inhibitory activities used in the present study were expressed as $pIC_{50}= -\log_{10}(IC_{50})$, Where IC_{50} is the micro molar concentration of the compounds producing 50% reduction in the effect caused by the virus is stated as themeans of at least two experiments. In our study, each molecule was described by tree descriptors, which are given by Ravichandran et al. [20].

$^1\chi$: first-order connectivity index

LUMO: lowest unoccupied molecular orbital.

DE: dielectric energy

67×4 matrix was obtained. 67 represents the number of the molecules and 4 represents the dependent variable ($-\log 1/IC_{50}$) and the tree independent variables (the 3 mentioned descriptors).

Results and Discussion

Two different sessions have been achieved: computation and prediction. 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.

Computation

The performances of SVM depend on the combination of several parameters. They are capacity parameter C, ε of ε-insensitive loss function and the corresponding parameters of the kernel function. C is a regularization

parameter that controls the tradeoff 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. However, Wang et al. [21] indicated that prediction error was scarcely influenced by C. In order to make the learning process stable, a large value should be set up for C.

The selection of the kernel function and corresponding parameters is very important because they implicitly define the distribution of the training set samples in the high dimensional feature space and also the linear model constructed in the feature space. There are four possible choices of kernel functions available in the LibSVM package i.e., linear, polynomial, radial basis function, and sigmoid function. For regression tasks, the radial basis function kernel is often used because of its effectiveness and speed in training process. In this work the form of the radial basis function used is:

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

Where γ is a parameter of the kernel, μ and ν are the two independent variables.

The of the kernel function greatly affect the number of support vectors, which has a close relation with the performance of the SVM and training time. Many support vectors could produce over fitting and increase the training time. In addition, controls the amplitude of the RBF function, and therefore, controls the generalization ability of SVM.

The optimal value for 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 , there is the practical consideration of the number of resulting support vectors. -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 is critical from theory.

To determine the optimal parameters, a grid search was performed based on leave-one-out cross validation on the original data set for all parameter combinations of from 100 to 1000 with incremental steps of 50, ranging from 2 to 3.2 with incremental steps of 0.1 and from 0.04 to 0.16 with incremental steps of 0.01. The optimal values of C, γ and are 500, 2.8 and 0.09, respectively.

Prediction

The main goal of any QSAR modelling is that the developed model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds. Tropsha et al [22] emphasizes the importance of rigorous validation as a crucial, integral component of QSAR model development. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged.

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

Internal validation

The internal validation technique used is cross-validation (CV), CV is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modelling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). The leave-one-out (LOO) procedure was utilized, in this study, which produce a number of models by deleting one from the whole data set.

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. As it can be seen in Table 1, high correlation coefficient (Q^2 =0.90) and low RMSE =0.145 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.

Table 1. Q2 and RMSE of SVM, ANN and MLR using cross validation (CV)

Method	Q^2	RMSE
SVM	0.90	0.145
ANN	0.84	0.185
MLR	0.80	0.210

External validation

In order to estimate the predictive power of SVM, MLR and ANN, we must use a set of compounds which have not been used for training set (used for establishing the QSAR model). The models established in the computation procedure, by using the 55 cyclic-urea derivatives, are used to predict the activity of the remaining 12 compounds. The plot of predicted versus experimental values for data set is shown in Figure 1 (SVM), Figure 2 (ANN) and Figure 3 (MLR). Among all these figures, the first one shows that the activity values calculated by the SVM are very close to the experimental ones. The statistical parameters of the three models are shown in Table 2. As can be seen from this table, the statistical parameters of SVM model are better than the other ones.

Table 2. Statistical parameters of different constructed QSAR models

Method	Training test			Test set	
	R	RMSE		R	RMSE
SVM	0.94	0.184		0.96	0.166
ANN	0.90	0.244		0.92	0.191
MLR	0.88	0.264		0.90	0.205

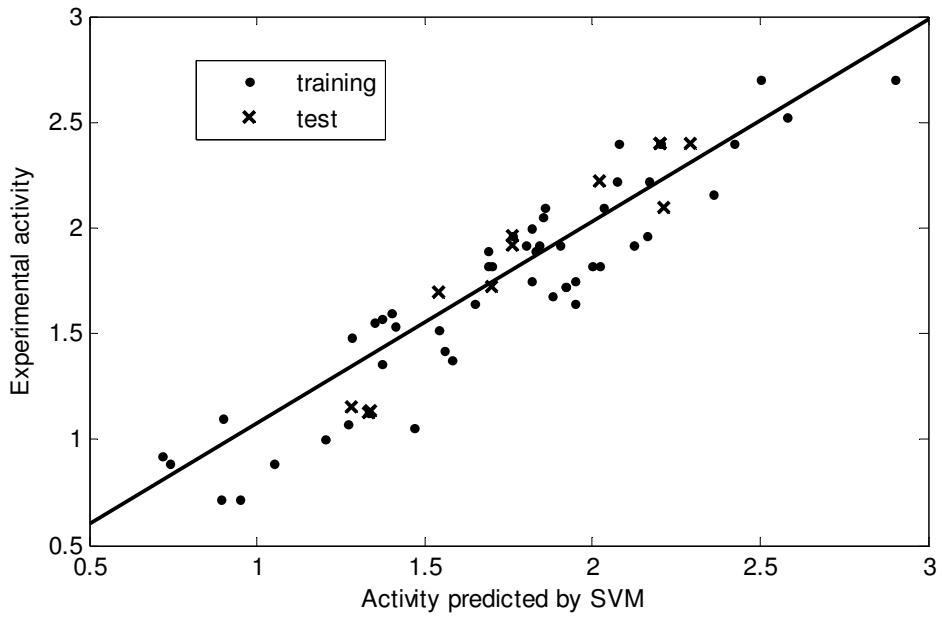

Figure 1. pIC_{50} observed experimentally versus pIC_{50} predicted by SVM

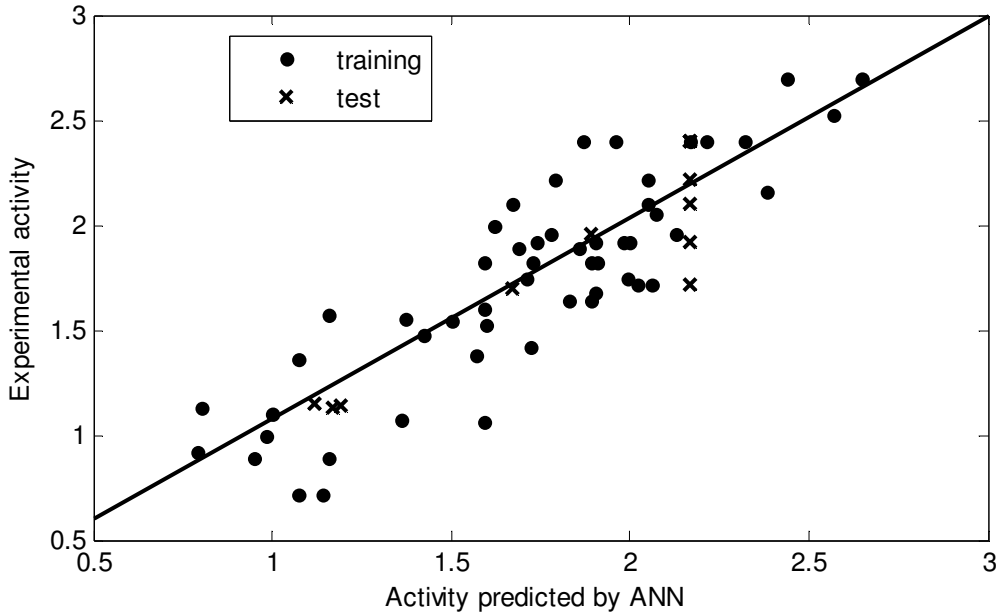

Figure 2. pIC_{50} observed experimentally versus pIC_{50} predicted by ANN

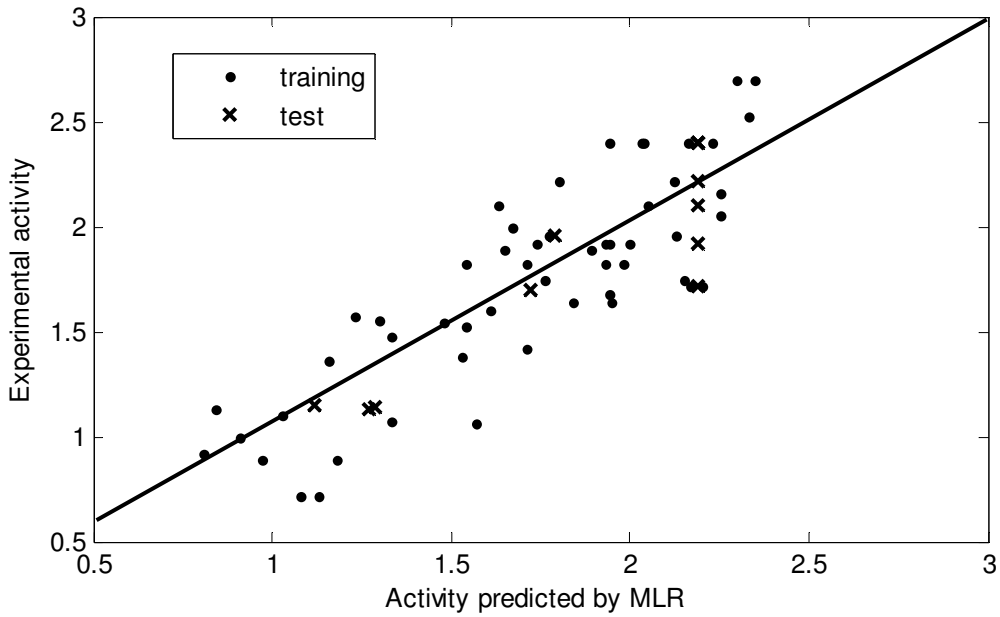

Figure 3. pIC_{50} observed experimentally versus pIC_{50} predicted by MLR

Y-randomization test

Y-randomisation is an attempt to observe the action of chance in fitting given data. In other words it is applied to exclude the possibility of chance correlation. This technique ensures the robustness of a QSAR model [23, 24]. The dependent variable vector [$y = \log(1/IC_{50})$] is randomly shuffled and a new QSAR model is developed using the original molecular descriptors. The new QSAR models (after several repetitions) are expected to have low R^2 values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modelling method and data.

In this work, ten random shuffles of the y vector were performed for SVM, ANN and MLR. The results are shown in Table 3. 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.

Table 3. Results of randomization test of the developed models

Modeling technique	R from non random model	Mean value of R from model trials
SVM	0.94	0.12
MLR	0.90	0.21
(10-5-1) ANN	0.88	0.29

Domain of applicability

The domain of application [25] 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 [39] 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:

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad i = 1...n$$

Where x_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The superscript T refers to the transpose of the matrix/vector. The warning leverage h^* is, generally, fixed at 3(k+1)/N, where N is the number of training compounds, and k is the number of model parameters. A leverage greater than the warning leverage h^* means that the predicted response is the result of substantial extrapolation of the model and, therefore, may not be reliable.

The Williams plot for the presented SVM model was showed in Figure 4. From this plot, the applicability domain is established inside a squared area within (±3s) standard deviations and a leverage threshold h^* of 0.22. As shown in the Williams plot (Figure 4), h_i values of all the compounds in the training and test sets are lower than the warning value (h^* =0.22). None of the compounds are particularly influential in the model space and the training set has great representativeness. For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units (2s). This means that all predicted values are acceptable.

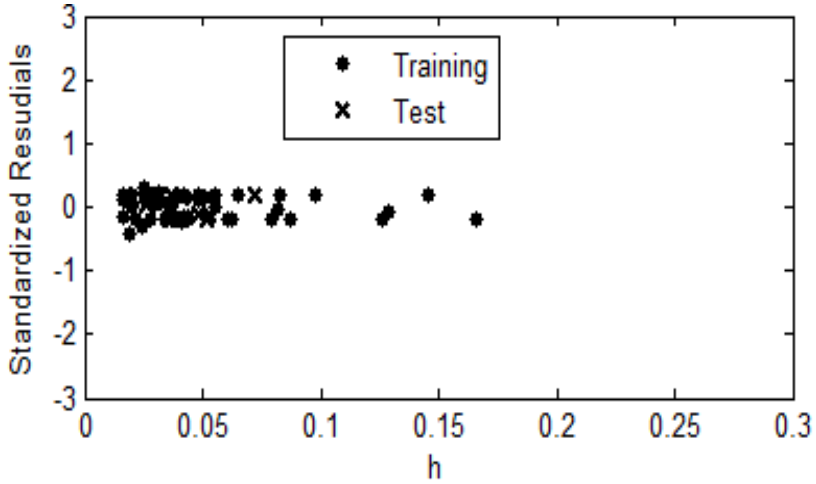

Figure 4. Williams plot of the current QSAR model

Conclusion

The support vector machine was used to develop a QSAR model for the prediction of the HIV-1 activity of 1,3,4- oxadiazole substituted naphthyridine derivatives. The results obtained show that the SVM technique was able to establish a satisfactory relationship between the molecular descriptors and the HIV-1 activity. of 1,3,4-oxadiazole substituted naphthyridine. The SVM approach would seem to have a great potential for determining quantitative structure-anti-HIV-1 activity relationships and as such be a valuable tool for the chemist.

References

1. Douali L., Villemin D., Cherqaoui D. Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives, Curr. Pharm. Des 2003, 9, 1817-1826.
2. The Nature of Statistical Learning Theory. Vapnik V N. (Eds) Springer, Berlin, 1995.
3. Burges J C. A tutorial on support vector machines for pattern recognition Data Min, Know. Discovery 1998, 2, 121–167.
4. Sweilam N H., Tharwat A A., Abdel Moniem N K. Support vector machine for diagnosis cancer disease: A comparative study, Egyptian Informatics Journal 2010, 11, 81-92.
5. Chen H L., Yang B., Liu J., You Liu D. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications 2011, 38, 9014-9022.
6. Wentong C., Xuefeng Y. Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR. Chemo. Intell. Labo. Syst 2009, 98, 130–135.
7. Menéndez-Arias L., Tözsér J. HIV-1 protease inhibitors: effects on HIV-2 replication and resistance. Trends Pharmacol Sci. 2008, 29, 42-49.
8. Walgate R., Degett J. EAGLES report on HIV/AIDS research. New. Biotechnology 25 (2008) 29.
9. Andersen J L., Le Rouzic E., Planelles V. HIV-1 Vpr: mechanisms of G2 arrest and apoptosis. Exp. Mol. Pathol. 2008, 85, 2-10.
10. Vangelista L., Secchi M., Lusso P. Rational design of novel HIV-1 entry inhibitors by RANTES engineering. Vaccine 2008, 26, 3008-3015.
11. Dessalew N. QSAR Study on Piperidinecarboxamides as Antiretroviral Agents: An Insight Into the Structural Basis for HIV Coreceptor Antagonist Activity. QSAR Comb. Sci. 2008, 27, 901-912.
12. Yuan H., Parrill A L. QSAR Development to Describe HIV-1 Integrase Inhibition. J. Mol. Struct.THEOCHEM 2000, 529, 273-282.
13. Chen X., Tsiang M., Yu F., Hung M., et al. Modeling, Analysis, and Validation of a Novel HIV Integrase Structure Provide Insights into the Binding Modes of Potent Integrase Inhibitors. J. Mol. Biol. 2008, 380, 504-519.
14. Cortes C., Vapnik V. Support vector networks. Mach. Learn. 1995, 20, 273-297.
15. Chang C C., Lin C J., LIBSVM-A Library for support vector machine. http://www/csie.edu/tw/cjlin/libs/libsvm
16. Neural Networks for Chemists. An Introduction. Zupan J., Gasteiger J. (Eds) VCH Publishers, Weinheim, 1993.
17. Cherqaoui D., Villemin D., Use of neural network to determine boiling point of alkanes. J. Chem. Soc. Faraday. Trans 1994, 90, 97-102.
18. Neural Networks Algorithms, Applications, and Programming Techniques. Freeman J A., Skapura D M. (Eds.), Addition Wesley Publishing Company, Reading, 1991.
19. Johns B A., Weatherhead J G., Allen S H., Thompson J.B., et al. 1,3,4-Oxadiazole substituted naphthyridines as HIV-1 integrase inhibitors. Part 2: SAR of the C5 position Bioorg. Med. Chem. Lett. 2009, 19, 1807-1810.
20. Ravichandran V., Sivadasan S., Karupiah S., Arumugam D S., QSAR study of substituted 1,3,4-oxadiazo;le naphthyridines as HIV-1 integrase inhibitors. European Journal of Medicinal Chemistry 2010, 45, 2791-2797.
21. Wang W J., Xu Z B., Lu W Z., Zhang X Y., Determination of the spread parameter in the Gaussian kernel for classification and regression. Neurocomputing 2003, 55, 643–663.
22. Tropsha A., Gramatica P., Gombar V., The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. Quant. Struct.-Act. Relat. 2003, 22, 1-9.
23. Golbraikh A., Tropsha A., Beware of q2!, J. Mol. Graph. Model. 2002, 20, 269–276.
24. Tropsha A., Golbraikh A., Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des 2007,13, 3494-504.
25. Eriksson L., Jaworska J., Worth A.P., Cronin, M.T., et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health. Perspect 2003, 111, 1361-1375.