



QSAR and classification models of a novel series of COX-2 selective inhibitors: 1, 5-diarylimidazoles based on support vector machines

H.X. Liu¹, R.S. Zhang^{1,2,*}, X.J. Yao^{1,3}, M.C. Liu¹, Z.D. Hu¹ & B.T. Fan³

¹Department of Chemistry, Lanzhou University, Lanzhou 730000, China; ²Department of Computer Science, Lanzhou University, Lanzhou 730000, China; ³Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, F-75005 Paris, France

Received 10 March 2004; accepted in revised form 9 August 2004

Key words: classification, COX-2 selective inhibitors, drug design, drug screening, QSAR, SVM

Summary

The support vector machine, which is a novel algorithm from the machine learning community, was used to develop quantitation and classification models which can be used as a potential screening mechanism for a novel series of COX-2 selective inhibitors. Each compound was represented by calculated structural descriptors that encode constitutional, topological, geometrical, electrostatic, and quantum-chemical features. The heuristic method was then used to search the descriptor space and select the descriptors responsible for activity. Quantitative modelling results in a nonlinear, seven-descriptor model based on SVMs with root mean-square errors of 0.107 and 0.136 for training and prediction sets, respectively. The best classification results are found using SVMs: the accuracy for training and test sets is 91.2% and 88.2%, respectively. This paper proposes a new and effective method for drug design and screening.

Introduction

Nonsteroidal antiinflammatory drugs (NSAIDs) [1, 2] are of immense benefit in the treatment of inflammatory diseases. The principal pharmacological effects of NSAIDs are due to their ability to inhibit prostaglandin (PG) synthesis by blocking cyclooxygenase (COX), which catalyzes the conversion of arachidonic acid to PGH₂ [3]. The therapeutic use of NSAIDs, especially in chronic diseases, has revealed their association with well-known side effects at the gastrointestinal level (mucosal damage, bleeding) [4, 5] and, less frequently, at the renal level [6]. The discovery of two isoforms [7], COX-1 and COX-2, helped in understanding the side effects associated with NSAIDs. COX-2 is an inducible form that is present only in inflammatory states. The inhibition of COX-1, the form constitutively present in many tissues such as stomach, kidney, and plate-

lets, by nonselective NSAIDs may be responsible for the secondary effects associated with their use [8, 9]. These observations suggest that selective COX-2 inhibitors could provide antiinflammatory, analgesic and antipyretic drugs devoid of the unwanted side effects such as ulcers and renal failure associated with the classical nonselective NSAIDs [10]. Selective COX-2 inhibition could also be an important strategy to prevent or treat a number of cancers [11, 12] and to delay or slow the clinical expression of Alzheimer's disease [13, 14]. Therefore, considerable efforts have been made to discover selective COX-2 inhibitors [15–28]. Overall, these selective COX-2 inhibitors have fulfilled the hope that they would exhibit a reduced risk in gastrointestinal events, although it is becoming increasingly apparent that they can cause nearly identical renal effects to those observed with nonselective NSAIDs. In order to achieve good activity and selectivity, recently, Carmen Almansa et al. synthesized a series of novel compounds and tested their activity and selectivity *in vitro* by obtaining their IC₅₀ values in human cell lines expressing COX-2

*To whom correspondence should be addressed. Phone: +86-931-891-2578; Fax: +86-931-891-2582; E-mail: ruison@public.lz.gs.cn

[29]. However, no quantitative structure–activity relationship (QSAR) model has been reported to date. To further design new drugs with high activity, it is very necessary and useful to investigate quantitative structure–activity relationships for this series of compounds.

Structure–activity relationship (SAR) analysis is one technique used to reduce the search for new drugs. A successful solution to this problem has the potential to provide significant economic benefit via increased process efficiency [30]. QSAR involves modeling a continuous activity for quantitative prediction of the activity of previously unseen compounds. The advances in quantitative structure–activity relationship (QSAR) studies have widened the scope of rationalizing drug design and the search for the mechanisms of drug actions. Artificial intelligence techniques have been applied to SAR analysis since the late 1980s, mainly in response to increased accuracy demands. Machine learning techniques have, in general, offered greater accuracy than have their statistical forebears, but there exist accompanying problems for the SAR analyst to consider. Neural networks, for example, offer high accuracy in most cases but can suffer from the reproducibility of results, due largely to random initialization of the network and variation of stopping criteria, and lack of information regarding the classification produced [31]. Genetic algorithms can suffer in a similar manner. The stochastic nature of both population initialization and the genetic operators used during training can make results hard to reproduce. Owing to the reasons outlined above, there is a continuing need for the application of more accurate and informative techniques to SAR analysis [30].

The support vector machine (SVM) is a new algorithm from the machine learning community. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application in pattern recognition and regression problems [30, 32–39].

The goal of this study is to develop robust QSAR models and binary classification models that predict and categorize inhibition values of 1,5-diarylimidazoles toward COX-2 based on the support vector machine and then serve as practical screening tools. The use of quantitative and classification models can augment and narrow the search for future drug compounds. Other classification and regression methods like k-nearest neighbour and the heuristic method were also respectively applied to build clas-

sification and quantitative models in order to identify the reliability of the support vector machines.

Experimental

Data set

The studied compounds are a series of 1,5-diarylimidazoles, whose structures are shown in Figure 1 and Table 1. The biological activity expressed as IC_{50} (in human cell lines expressing COX-2) was taken from Ref. 29. The IC_{50} values range from 0.002 to $>10 \mu\text{M}$ and are reported as the average of duplicate measurements. QSAR analysis was performed on smaller compound subsets of the 53 structures with IC_{50} values ranging from 0.002 to $0.1 \mu\text{M}$. For analysis purposes, $-\log (IC_{50})$ values were used as the dependent variables and are given in Table 1.



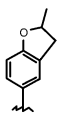

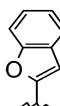
Descriptor generation and feature selection

The two-dimensional structures of the molecules were drawn with the ISIS DRAW program. All molecules were transferred into Hyperchem and pre-optimized using the MM+ molecular mechanics force field. A more precise optimization is done with the semi-empirical PM3 method in MOPAC. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.001. The resulting geometry was transferred into the CODESSA software, developed by the Katritzky group [40, 41], that can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors and has been successfully used in various QSPR and QSAR researches. Constitutional descriptors are related to the number of atoms and bonds in each molecule. Topological descriptors include valence and non-valence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition and the degree of branching of a molecule. Geometrical descriptors are calculated from 3D atomic coordinates of the molecule. These descriptors comprise moments of inertia, shadow indices, molecular volume, molecular surface area, and gravitation indices. Electrostatic descriptors reflect characteristics of the charge distribution of the molecule. The quantum chemical descriptors include information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels.

Table 1. The compounds and the predicted results of $-\log\text{IC}_{50}$ from the SVM regression and classification model.

Compound	R1	R2	$-\log\text{IC}_{50}$ (exp.)	$-\log\text{IC}_{50}$ (pred.)	Act./inact. (exp.)	Act./inact. (pred.)
1a	H	H			—	—
1b*	Cl	H	7.854	7.554	+	+
1c	Br	H	7.444	7.427	+	+
1d	Me	H	6.842	6.772	+	+
1f [#]	Et	H			—	—
1g [#]	Allyl	H			—	—
1h	CH ₂ Ph	H			—	—
1i	Pent	H			—	—
1j [#]	H	Cl			—	+
1k	H	COPh			—	—
1m [#]	H	CH ₂ OH			—	—
1n	Cl	CH ₂ OH			—	—
1o	Cl	Cl			—	+
1p	Cl	Me			—	—
2a	H	H			—	—
2b*,#	Cl	H	7.854	7.769	+	+
2c	Br	H	7.319	7.510	+	+
2d	CH ₃	H	7.292	7.414	+	—
2e [#]	CH ₂ OH	H			—	—
2f	CH ₂ F	H			—	—
2g	CHO	H			—	—
2h	COOMe	H			—	—
2i	CN	H	7.292	7.298	+	+
2j	COPh	H	7.523	7.517	+	+
2k	H	Cl			—	+
2l	H	Br			—	+
2m	Cl	Cl			—	—
2n [#]	Cl	Me			—	—
3a	3-F		7.187	7.887	+	+
3b	2-F		7.553	7.575	+	+
3c	H		6.910	7.397	+	+
3d	4-Cl		7.745	7.707	+	+
3e*,#	4-Me		7.796	7.247	+	+
3f*	4-OMe		7.959	7.667	+	+
3g	4-OEt		8.398	7.850	+	+
3h	4-OPr				—	—
3i	4-OPr ⁱ				—	+
3j	4-OCF ₃				—	+
3k	4-Pr		7.000	7.378	+	+
3l*,#	4-Pr ⁱ		7.409	7.532	+	+
3m [#]	4-SMe		7.959	7.805	+	+
3n [#]	4-SEt				—	—
3o	4-SO ₂ Et		6.728	7.132	+	+
3p	4-NH ₂				—	—
3q	4-AcNH		7.097	7.103	+	+
3r	4-NEt ₂		8.155	7.712	+	+
3s	2,4-di-F		7.824	7.948	+	+
3t	4-OMe-2-F		8.222	8.216	+	+
3u [#]	3,4-di-Cl		8.398	7.922	+	+
3v	4-OMe-3-F		8.222	7.781	+	+

Table 1. Continued.

Compound	R1	−logIC ₅₀ (exp.)	−logIC ₅₀ (pred.)	Act./inact. (exp.)	Act./inact. (pred.)
3w	4-Me-3-F	7.886	7.772	+	+
3x*	4-OMe-3-Me	7.824	7.449	+	+
3y	4-Me-3-OMe	7.959	7.805	+	+
3z*	4-Cl-3-Me	7.569	7.731	+	+
3aa	4-NMe ₂ -3-Cl	8.097	7.933	+	+
3ab	4-OMe-3-Cl	8.155	8.206	+	+
3ac	4-OEt-3-Cl	7.602	7.596	+	+
3ad	4-OEt-3-F	7.796	7.904	+	+
3ae [#]	4-F-3-OMe	8.155	8.102	+	+
3af	4-OMe-3,5-Cl			−	−
3ag	3,5-di-OEt	7.125	7.489	+	+
3ah	3,5-di-F			−	−
4a	3-pyridyl			−	−
4b	6-Me-3-pyridyl	6.547	6.541	+	+
4c*	6-Cl-3-pyridyl	7.367	6.855	+	+
4d	6-EtO-3-pyridyl	6.562	7.489	+	+
4e [#]	6-OH-3-pyridyl			−	−
4f	6-NEt ₂ -3-pyridyl			−	−
4g	4-pyridyl			−	−
4h		8.222	8.218	+	+
4i		8.046	8.052	+	+
4j		7.131	7.866	+	+
4k				−	−
4l				−	−
5a*	4-F	7.745	8.033	+	+
5b [#]	H	8.046	7.911	+	+
5c	4-Me	8.523	8.286	+	+
5d	4-OEt	8.699	8.502	+	+
5e	3,4-di-Cl	8.699	9.110	+	+
5f	4-OMe-3F	8.301	8.273	+	+
5g	4-F-3-OMe	8.097	7.933	+	+
5h [#]	4-OEt-3-Cl	7.699	7.870	+	+
5i	4-OMe-3-Cl	8.301	8.593	+	+
5j*	4-OEt-3-Cl	8.155	7.618	+	+
5k [#]	4-Cl-3-Py ^g	7.377	7.380	+	−

*The compounds in the test set for the support vector regression model; [#]the compounds in the test set for the support vector classification model. 'Act.' = active. 'inact.' = inactive.

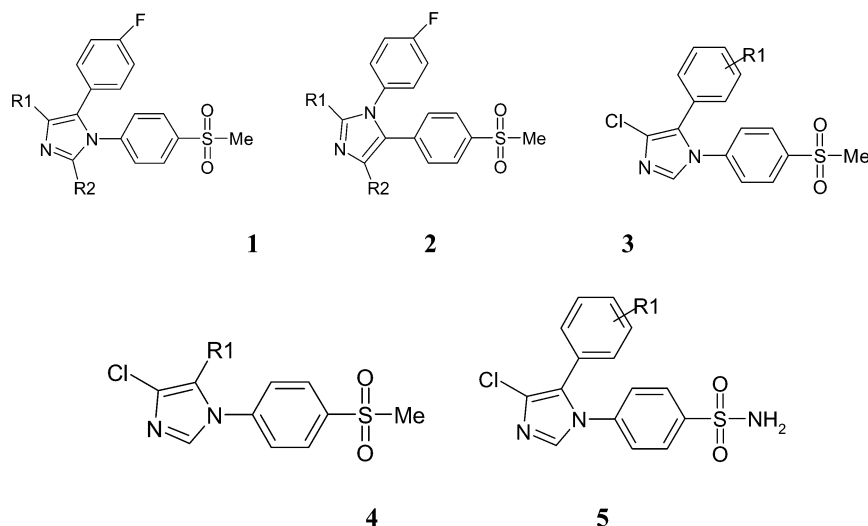


Figure 1. Generic structures for the 85 compounds used in this study.

Once molecular descriptors are generated, CODESSA uses the heuristic method to accomplish the pre-selection of descriptors. Its advantages are the high speed and no software restrictions on the size of the data set. The heuristic method can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which descriptors are insignificant (from the standpoint of a single-parameter correlation), and which descriptors are highly intercorrelated. This information will be helpful in reducing the number of descriptors involved in the search for the best QSPR model.

First of all, all descriptors are checked to ensure: (a) that values of each descriptor are available for each structure; and (b) that there is a variation in these values. Descriptors for which values are not available for every structure in the data set in question are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested and insignificant descriptors removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. All two-parameter regression models with remaining descriptors are subsequently developed and ranked by the regression correlation coefficient. A stepwise addition of further descriptor scales is performed to find the best multi-parameter regression

models with the optimum values of statistical criteria (highest values of R^2 , the cross-validated R_{cv}^2 , and the F -value).

Methodology

After the descriptors are selected, the next step is to build the quantitative and classification model by using some computational method. In this paper, for the quantitative model, the heuristic method and support vector machines were used; for the classification model, the k -nearest neighbour and support vector machines were examined. As the theories of the heuristic method and clustering analysis have been well described in many monographs and articles, we only give a brief description on the simple theory of the SVMs.

Support vector machines

The support vector machine (SVM), developed by Vapnik [42] as a novel type of learning machine, is gaining popularity due to many attractive features and promising empirical performance. Compared with traditional neural networks, SVM possesses prominent advantages: (1) Its strong theoretical background provides SVM with high generalization capability and can avoid local minima. (2) SVM always has a solution, which can be quickly obtained by a standard algorithm (quadratic programming). (3) SVM does not need to determine network topology in advance, which can be automatically obtained when the training process ends. (4) SVM builds a result based on a sparse subset of training samples, which reduce the

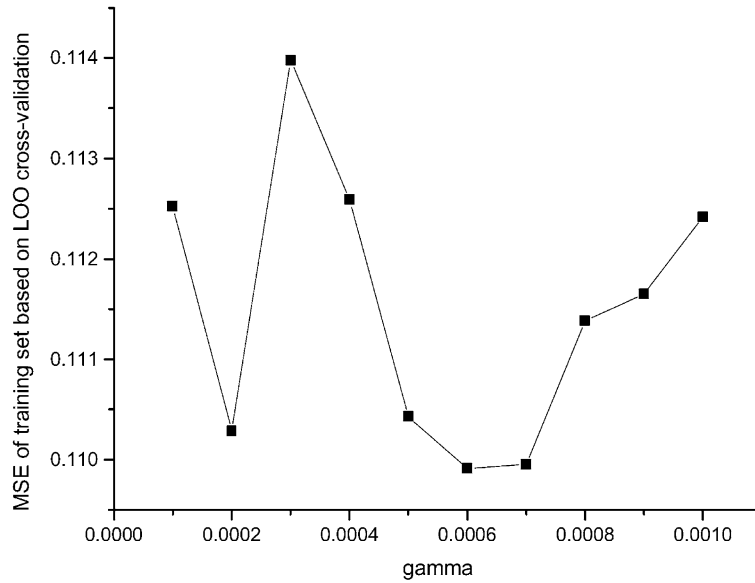


Figure 2. The gamma versus MSE error of the training set based on LOO cross-validation ($C=1000$, $\varepsilon = 0.01$).

workload. It can solve high-dimension problems and therefore avoid the ‘curse of dimensionality’ [43]. Originally, SVMs are developed for pattern recognition problems. And now, with the introduction of an ε -insensitive loss function, SVMs have been extended to solve nonlinear regression estimation and time-series prediction and excellent performances have been obtained [36, 44, 45].

There exist a number of excellent introductions into SVM, both printed [46–48] and electronically available [49]. For this reason, we will only briefly describe the main ideas of SVM classification and regression here.

For the classification problem, in brief, this involves the optimization of Lagrangian multipliers α_i with constraints $0 \leq \alpha_i \leq C$ and $\sum \alpha_i y_i = 0$ to yield a decision function

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^l y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \quad (1)$$

where y_i are input class labels that take a value of -1 or 1 , \mathbf{x}_i is a set of descriptors, and $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function, whose value is equal to the inner product of two vectors \mathbf{x} and \mathbf{x}_i in the feature space $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}_i)$. That is, $K(\mathbf{x}, \mathbf{x}_i) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i)$. The elegance of using a kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(\mathbf{x})$ explicitly. Any function that satisfies Mercer’s condition can be

used as the kernel function. The sign function $\text{sign}(u)$ returns 1 when $u > 0$, and -1 when $u \leq 0$.

For the regression problem, decision function (1) takes on the following form

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (2)$$

The constraints are the same as those of Equation (1).

For a given dataset, only the kernel function and the regularity parameter C must be selected to specify one SVM. In support vector classification and regression, the Gaussian kernel $K(u, v) = \exp(-|u - v|^2 / \delta^2)$ is most commonly used.

Algorithm implementation and computation environment

All calculation programs implementing SVM were written in R-file using Libsvm based on R script. The K -nearest neighbour algorithm was also performed by R software. All scripts were compiled using R1.7.1 compiler running on a Pentium IV PC with 256M RAM.

Table 2. The linear model between structure and activity ($R^2 = 0.7141$, $F=16.06$, $s^2=0.0934$).

Descriptor	Coefficient	error	T-test value
Constant	0.6826	1.4707	0.4641
RPCS relative positive charged SA (RPCS)	0.2785	0.0796	3.4981
HACA-2/TMSA (HACA2/T)	-818.8700	301.03	-2.7202
HACA-1 (HACA-1)	0.1239	0.0391	3.1683
Average structural information content (order 0) (ASIC0)	10.3010	2.5456	4.0467
Number of benzene rings (NBR)	0.7853	0.2181	3.6013
PNSA-1 Partial negative surface area (PNSA-1)	-0.0075	0.0022	-3.4368
Avg bond order of a N atom (ABON)	1.9507	0.9156	2.1305

Results and discussion

The heuristic method model

Through the heuristic method implemented in CODESSA, the best linear model with seven parameters was obtained, which is shown in Table 2. By interpretation of the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the activity of these compounds under the same mechanism of interaction. According to the t-test, the most important descriptor affecting the activity is a topological descriptor, average structural information content (order 0) (ASIC0). The ASIC0 [50, 51] is defined on the basis of the Shannon information theory. The calculated value for each molecule reflects how information rich the molecule is and in essence gives us information concerning how many atoms with similar connectivity patterns are in the molecule. 'Information rich' describes how many different atoms there are in the molecule. The average structural information content also depends on the number of atoms in the molecule, and it arranges the molecules in the order of compound size and then describes the difference in hydrophobic forces in the binding. However, this descriptor cannot give the differences among all molecules, the addition of another constitutional descriptor *number of benzene rings* (NBR) can differentiate the sulfones whose aryl substituent is phenyl or pyridyl. When the substituent is phenyl, the number of benzene rings is 2, otherwise 1. In addition, this descriptor also can reflect the size of the molecule to some degree and then can account for the hydrophobicity of compounds partially. The introduction of a quantum chemical descriptor Avg bond order of an N atom (ABON) can account for the difference between the sulfones and the sulfon-

amides. The combination of these three descriptors, comprising size and shape information about molecules, adequately represents hydrophobic and steric term effects on the activity of a molecule.

The remaining four descriptors are all the electrostatic descriptors, which reflect characteristics of the charge distribution of the molecule, and all are charged partial surface area (CPSA) descriptors encoding features responsible for polar interactions between molecules. The descriptors RPCS Relative positive charged SA and PNSA-1 Partial negative surface area describe the charge distribution in the molecule. They can be loosely related to the hydrogen bonding acceptor and donor ability and also to the reactivity. The descriptors HACA-1, *hydrogen acceptor charged area* and HACA-2/TMSA, *hydrogen bond acceptor charged surface area/total molecular surface area (HACA-1/TMSA)*, describe the hydrogen bonding acceptor properties of the compounds. Based on the above discussion, these four descriptors can account sufficiently for the electrostatic and hydrogen bonding influence on the activity of a molecule [50].

Analysis of the results obtained indicates that the selected molecular descriptors calculated solely from structures can describe the structural features of the compounds responsible for their biological activity. However, from Table 2, we can see that the linear relationship between structure and the experimental $-\log IC_{50}$ values is not very strong. Therefore, the descriptors listed in Table 1 were submitted to SVMs to determine whether nonlinearity in the training method would produce better results.

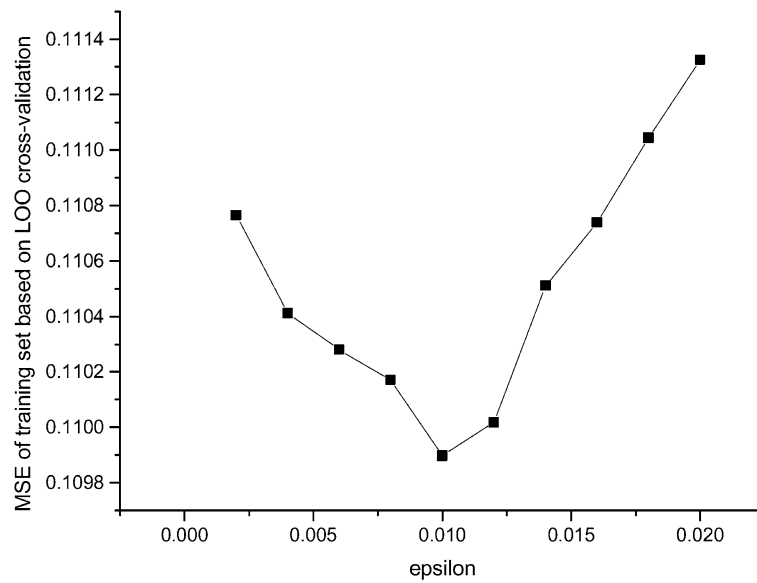


Figure 3. The epsilon versus MSE error of the training set based on LOO cross-validation ($C=1000$, $\gamma = 0.0006$).

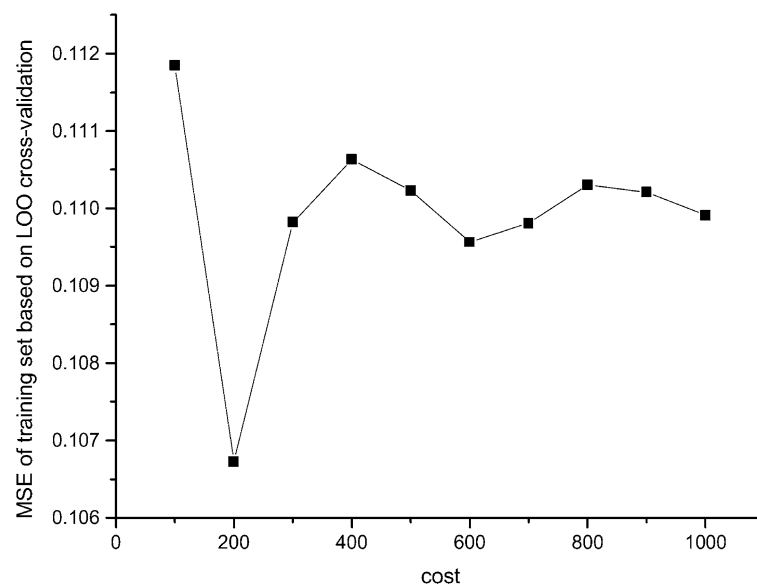


Figure 4. The cost versus MSE error of the training set based on LOO cross-validation ($\epsilon = 0.01$, $\gamma = 0.0006$).

Support vector regression model

Selection of the kernel function and parameters of the SVM

From the above discussion about the theory of SVM, the performances of SVM for regression depend on the combination of several parameters. These are capacity parameter C , ϵ of the ϵ -insensitive loss function, the kernel type K 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. But, Ref. 44 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 (e.g., $C = 1000$) firstly.

The kernel type is another important one. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function in R is as follows:

$$\exp\left(-\gamma * |u - v|^2\right)$$

where γ is a constant, the parameter of the kernel, u , v are two independent variables, and 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 value.

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 [44].

In order to find the optimized combination of several parameters, the data set was separated randomly into a training set of 43 compounds and a test set of 10 compounds and leave-one-out cross-validation of the whole training set was performed. The leave-one-out (LOO) procedure consists of removing one example from the training set, constructing the decision function on the basis only of the remaining training data and then testing on the removed example. In this fashion one tests all examples of the training data and measures the fraction of errors over the total number of training examples. The MSE was used as an error function, and it is computed according to the following equation

$$MSE = \frac{\sum_{i=1}^n (d_i - o_i)^2}{n}$$

where d_i are the teaching outputs (desired outputs) in the validation set, o_i are the actual outputs, and n is the number of samples in the validation set.

The detailed process of selecting parameters and the effects of every parameter on generalization performance of the corresponding model are shown in Figures 2–4. To obtain the optimal γ , support vector learning machines with different γ were trained, the γ varying from 0.0001 to 0.001. We calculated the MSE on different γ , according to the generalization ability of the model based on LOO cross-validation for the training set in order to determine the optimal one. The curve of MSE versus gamma is shown in Figure 2. The optimal γ was found to be 0.0006.

In order to find an optimal ε , the MSE on different ε was calculated. The curve of the MSE versus epsilon is shown in Figure 3. The performance of the SVM is worse firstly and then better as ε increases in Figure 3. The optimal ε was found to be 0.01.

The last important parameter is regularization parameter C, whose effect on the MSE is shown in Figure 4. From Figure 4, the performance of the model becomes better firstly and then worse and finally insensitive as C increases; its optimal value was 200.

The predicted result of SVMs

From the above discussion, γ , ε and C were fixed to 0.0006, 0.01 and 200, respectively. The predicted results of the optimal SVMs are shown in Figure 5 and in Table 1. The root mean square errors of the training set and the testing set are 0.107, 0.136, the mean relative errors are 2.44%, 4.15% respectively, and the prediction correlation coefficient $R=0.869$ was obtained for the whole data set. For the biological activity data of drug with high noise, it can be concluded that the predicted values are in good agreement with the experimental values from the above results.

Classification model

All 85 compounds were used for the classification analysis. The compounds whose IC_{50} were less than $1.0 \mu\text{m}$ are considered as active and the others as inactive. This results in a 53-member class of active compounds and a 32-member class of inactive compounds. Here, in order to verify the validation of the classification model, the whole data set with 85 compounds was divided into a test set with 17 compounds (including 9 active compounds and 8 inactive compounds) and a training set with 68 compounds (including 44 active compounds and 24 inactive compounds).

K-nearest neighbor classification model

K-nearest neighbor (kNN) classification is an algorithmically simple, supervised method which classifies an unknown compound based on the class membership of its k nearest neighbor compounds. To determine the nearest neighbors, an Euclidean distance matrix of all samples in the training set is scanned for the k shortest distances to the observation of interest. For this work, k was selected as 3. The model was able to correctly classify 64.7% of the training set compounds (44 of 68 compounds) and 64.7% of the prediction set compounds (11 of 17 compounds). From the above results,

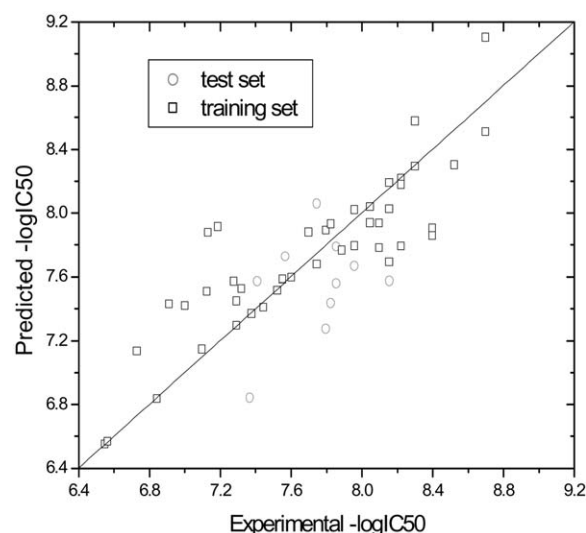


Figure 5. The predicted values of $-\log IC_{50}$ versus the experimental data ($\varepsilon = 0.01$, $\gamma = 0.0006$, $C=200$).

it can be seen that the results from the kNN model are not satisfied. In order to build up the more reliable classification model, the support vector machine for classification was introduced.

Support vector machine classification model

Similar with the support vector regression model, the performances of SVM for classification depend on the combination of several parameters. They are capacity parameter C , the kernel type K and its corresponding parameters. Through the same process with regression, the parameters selected ultimately are as follows: the kernel type is the Gaussian kernel, γ 0.006, C 1000. The corresponding results were obtained: the accuracy of predicting for the training set was 91.2%, for the test set 88.2%. Table 1 gives the detailed results from the SVM classification model.

Comparing the results from the SVM with those from the kNN , it can be seen that in this study, the performance of the kNN algorithm is far worse than that of SVM.

Conclusions

The above results indicate that the SVM is a very promising tool both for nonlinear approximation and classification. Besides, the SVM exhibits the better performance on the whole due to embodying the Structural Risk Minimization principle and some advantages over the other techniques of converging to

the global optimum, and not to a local optimum. The predictive results are satisfied not only for regression but also classification. Therefore it is a good approach for predicting the expected activity of drugs and aiding in drug design. At the same time, the models proposed could identify and provide some insight into what structural features are related to the biological activity of these compounds and afford some instruction for further designing new selective COX-2 inhibitors.

Acknowledgements

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Programme PRA SI 02-03). The authors also thank the R Development Core Team for affording the free R1.7.1 software.

References

- Mantri, P. and Witiak, D., *Curr. Med. Chem.*, 1 (1994) 328.
- Vane, J.R. and Botting, R., *Scand. J. Rheumatol. Suppl.*, 102 (1996) 9.
- Smith, W.L., Borgeat, P. and Fitzpatrick, F.A. The eicosanoids: COX, lipoxygenase, and epoxigenase pathways. In *Biochemistry of Lipids, Lipoproteins and Membranes* (Vance, D.E. and Vance, J., Eds.). Elsevier, New York, 1991, pp. 297–325.
- Sontag, S.J., *Drugs*, 32 (1986) 445.
- Allison, M.C., Howatson, A.G., Torrance, C.J., Lee, F.D. and Russell, R.Y.G., *N. Engl. J. Med.*, 327 (1992) 749.
- Clive, D.M. and Stoff, J.S., *N. Engl. J. Med.*, 310 (1984) 563.
- Xie, W., Chipman, J., Robertson, D.L., Erikson, R.L. and Simmons, D.L., *Proc. Natl. Acad. Sci. U.S.A.*, 88 (1991) 2692.
- Hla, T. and Neilson, K., *Proc. Natl. Acad. Sci. U.S.A.*, 89 (1992) 7384.
- Kujubu, D.A. and Herschman, H.R., *J. Biol. Chem.*, 267 (1992) 7991.
- Ballinger, A. and Smith, G., *Exp. Opin. Pharmacother.*, 2 (2001) 31.
- Subbaramaiah, K., Zakim, D., Weksler, B.B. and Dannenberg, A.J., *Proc. Soc. Exp. Biol. Med.*, 216 (1997) 201.
- Hsu, A.L., Ching, T.T., Wang, D.S., Song, X., Rangnekar, V.M. and Chen, C.S., *J. Biol. Chem.*, 275 (2000) 11397.
- Pasinetti, G.M., *J. Neurosci. Res.*, 54 (1998) 1.
- Hull, M., Lieb, K. and Fiebich, B.L., *Exp. Opin. Invest. Drugs*, 9 (2000) 671.
- Leblanc, Y., Black, W.C., Chan, C.C., Charleson, S., Delorme, D., Denis, D., Gauthier, J.Y., Grimm, E.L., Gardon, R., Guay, D., Hamel, P., Kargman, S., Lau, C.K., Mancini, J., Ouellet, M., Percival, D., Roy, P., Skorey, K., Tagari, P., Vickers, Wong, E., Xu, L. and Prasit, P., *Biorg. Med. Chem. Lett.*, 6 (1996) 731.
- Kalgutkar, A.S., *Exp. Opin. Ther.*, 9 (1999) 831.
- Reitz, D.B. and Isakson, P.C., *Curr. Pharm. Design*, 1 (1995) 211.
- Carter, J., *Exp. Opin. Ther. Pat.*, 8 (1997) 21.

19. Penning, T.D., Talley, J.J., Bertenshaw, S.R., Carter, J., Collins, P.W., Docter, S., Graneto, M.J., Lee, L.F., Malecha, W., Miyashiro, J.M., Rogers, R.S., Rogier, D.J., Yu, S., Anderson, G.D., Burton, E.G., Cogburn, J.N., Gregory, S., Koboldt, C.M., Perkins, W.E., Seibert, K., Veenhuizen, A., Zhang, Y.Y. and Isakson, P.C., *J. Med. Chem.*, 40 (1997) 1347.
20. Prasit, P., Wang, Z., Brideau, C., Chan, C.-C., Charleson, Cromlish, W., Ethier, D., Evans, J.F., Ford-Hutchinson, A. Gauthier, J.Y., Gordon, R., Guay, J., Gresser, M., Kargman, Kennedy, B., Leblanc, Y., Léger, S., Mancini, J., McNeill, G. Ouellet, M., Percival, M.D., Perrier, H., Riendeau, D., Rodger, Y., Tagari, P., Thérien, M., Vickers, P., Wong, E., Xu, L.-Young, R.N. and Zamboni, R., *Bioorg. Med. Chem. Lett.*, 9 (1999) 1773.
21. Talley, J.J., Brown, D.L., Carter, J.S., Graneto, M., Koboldt, C.M., Masferrer, J.L., Perkins, W.E., Rogers, R., Shaffer, A.F., Zhang, Y.Y., Zweifel, B.S. and Seibert, K., *J. Med. Chem.*, 43 (2000) 775.
22. Talley, J.J., Bertenshaw, S.R., Brown, D.L., Carter, J.S., Graneto, M.J., Kellogg, M.S., Koboldt, M., Yuan, J., Zhang, Y.Y. and Seibert, K., *J. Med. Chem.*, 43 (2000) 1661.
23. Riendeau, D., Percival, M.D., Brideau, C., Charleson, S., Dubé, D., Ethier, D., Falgoutyret, J.P., Friesen, R.W., Gordon, R., Greig, G., Guay, J., Mancini, J., Oellet, M., Wong, E., Xu, L., Boyce, S., Visco, D., Girard, Y., Prasit, P., Zamboni, R., Rodger, I.W., Gresser, M., Ford, Hutchinson, A.W., Young, R.N. and Can, C.C., *J. Pharmacol. Exp. Ther.*, 296 (2001) 558.
24. Balsamo, A., Coletta, I., Domiano, P., Guglielmotti, A., Landolfi, C., Mancini, F., Milanese, C., Orlandini, E., Rapposelli, S., Pinza, M. and Macchia, B., *Eur. J. Med. Chem.*, 37 (2002) 391.
25. Kalgutkar, A.S., Rowlinson, S.W., Crews, B.C. and Marnett, L.J., *Bioorg. Med. Chem. Lett.*, 12 (2002) 521.
26. Rao, P.N.P., Amini, M., Li, H.Y., Habeeb, A.G. and Knaus, E.E., *Bioorg. Med. Chem. Lett.*, 13 (2003) 2205.
27. Pal, M., Veeramaneni, V.R., Nagabelli, M., Kalleda, S.R., Misra, P., Casturib, S.R. and Yeleswarapua, K.R., *Bioorg. Med. Chem. Lett.*, 13 (2003) 1639.
28. Hu, W.H., Guo, Z.R., Chu, F.M., Bai, A.P., Yi, X., Cheng, G. F. and Li, J., *Bioorg. Med. Chem.*, 11 (2003) 1153.
29. Almansa, C., Alfón, J., Arriba, A.F., Cavalcanti, F.L., Escamilla, I., Gómez, L.A., Miralles, A., Soliva, R., Bartrolí, J., Carceller, E., Merlos, M. and Julián, G.R., *J. Med. Chem.* 46 (2003) 3463.
30. Burbidge, R., Trotter, M., Buxton, B. and Holden, S., *Comput. Chem.*, 26 (2001) 5.
31. Manallack, D.T. and Livingstone, D.J., *Eur. J. Med. Chem.*, 34 (1999) 95.
32. Bao, L. and Sun, Z.R., *FEBS Lett.*, 521 (2002) 109.
33. Belousov, A.I., Verzakov, S.A. and Von Frese J., *Chemometr. Intell. Lab. Syst.*, 64 (2002) 15.
34. Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C., *Comput. Chem.*, 26 (2002) 293.
35. Morris, C.W., Autret, A. and Boddy, L., *Ecol. Model.*, 146 (2001) 57.
36. Song, M., Breneman, C.M., Bi, J., Sukumar, N., Bennett, K.P., Cramer, S. and Tugcu, N., *J. Chem. Inf. Comput. Sci.*, 42 (2002) 1347.
37. Liu, H.X., Zhang, R.S., Luan, F., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., *J. Chem. Inf. Comput. Sci.*, 43 (2003) 900.
38. Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., *J. Chem. Inf. Comput. Sci.*, 43 (2003) 1288.
39. Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., *J. Chem. Inf. Comput. Sci.*, 44 (2004) 161.
40. Katritzky, A.R., Lobanov, V.S. and Karelson, M., 1995. CODESSA: Training Manual. University of Florida, Gainesville, Florida.
41. Katritzky, A.R., Lobanov, V.S. and Karelson, M. 1994. CODESSA: Reference Manual. University of Florida, Gainesville, Florida.
42. Cortes, C. and Vapnik, V., *Machine Learning*, 20 (1995) 273.
43. Gunn, S.R., Brown, M. and Bossley, K.M., *Lecture Notes Comput. Sci.*, 1280 (1997) 313.
44. Wang, W.J., Xu, Z.B., Lu, W.Z. and Zhang, X.Y., *Neurocomputing*, 55 (2003) 643.
45. Tugcu, N., Song, M., Breneman, C.M., Sukumar, N., Bennett, K.P. and Cramer, S.M., *Anal. Chem.*, 75 (2003) 3563.
46. Vapnik, V. *Statistical Learning Theory*, Wiley, New York, 1998.
47. Schölkopf, B., Burges, C. and Smola, A., *Advances in Kernel Methods – Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
48. Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.
49. URL: <http://www.kernel-machines.org/>.
50. Basak, S.C., Balaban, A.T., Grunwald, G.D. and Gute, B.D., *J. Chem. Inf. Comput. Sci.*, 40 (2000) 891.
51. Katritzky, A.R. and Tatham, D.B., *J. Chem. Inf. Comput. Sci.*, 41 (2001) 0062.