European Journal of Medicinal Chemistry 44 (2009) 4090–4097

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

QSAR study of Akt/protein kinase B (PKB) inhibitors using support vector machine

Xiaowu Dong, Chaoyi Jiang, Haiyun Hu, Jingying Yan, Jing Chen, Yongzhou Hu*

ZJU-ENS Joint Laboratory of Medicinal Chemistry, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 17 November 2008
Received in revised form
28 March 2009
Accepted 30 April 2009
Available online 15 May 2009

Keywords:
Akt/protein kinase B (PKB) inhibitors
Support vector machine (SVM)
QSAR
Support vector regression (SVR)
Support vector classification (SVC)

ABSTRACT

A three-class support vector classification (SVC) model with high prediction accuracy for the training, test and overall data sets (95.2%, 88.6% and 93.1%, respectively) was developed based on the molecular descriptors of 148 Akt/protein kinase B (PKB) inhibitors. Then, support vector regression (SVR) method was applied to set up a more accurate model with good correlation coefficient (r^2) for the training, test and overall data sets (0.882, 0.762 and 0.840, respectively). Enrichment factors (EF) and receiver operating curves (ROC) studies of database screening were also performed either using the SVR model alone or assisted with the SVC model, the results of which demonstrated that the established models could be useful and reliable tools in identifying structurally diverse compounds with Akt inhibitory activity.

1. Introduction

Kinase inhibitors have gained great interest due to the remarkable success of Gleevec and other tyrosine kinase inhibitors in clinical oncology [1–3]. As one of the major downstream targets of several tyrosine kinases in signal transduction pathways, Akt (also known as protein kinase B) is a central node of PI3K/Akt signaling pathway and is thought to be among the most frequently mutated or overexpressed signaling abnormality in human cancer [4]. Activated Akt phosphorylates a variety of protein substrates, including GSK-3β, FKHRL1, BAD and mTOR, regulating cell proliferation, protein translation, cell survival and apoptosis, and progression through the cell cycle [5,6]. Several Akt inhibitors, including diphenylquinoxalines [7], bispyridinylethylenes [8], isoquinoline (indazole)–pyridines [9,10] and indazole–diones [11], have shown the ability to sensitize tumor cells to apoptotic stimuli and to slow tumor growth *in vivo* during the past few years. Much attention has been attached to the design and synthesis of novel compounds for Akt modulation through X-ray interaction and molecular docking studies, and several interaction modes have been reported [12,13]. However, the results were not applicable to molecules with diverse structures. Alternatively, the untested molecule might be "*in silico*" evaluated based on structural information from identified QSAR models. QSAR analysis is an effective method in rational drug design and exploring the activation mechanism of drugs [14,15]. Many statistical methods have been involved in QSAR research [16,17], such as multiple linear regression (MLR), artificial neural network (ANN), support vector machine (SVM), etc. As we known, linear method is mainly limited to a complex biological system; although the flexibility of neural networks enables them to study more complex nonlinear relationships based on experimental data, some inherent problems (such as overtraining) still wait to be addressed; the SVM is a new algorithm developed from the machine learning community and is commonly used for QSAR studies. Previously, an SVM-based classification model for vasorelaxation agents was identified by our group and applied to design novel vasodilators [18].

Despite intensive studies for Akt inhibitors, few relating QSAR studies are available up till now. Since correlation between the structural information of diverse compounds and their Akt inhibitory activities may be feasible and useful for designing novel inhibitors. SVM study was applied to set up Akt inhibitors QSAR model based on the descriptors of the diverse data set in our present work. Among various descriptors calculated from molecules using Dragon software, 17 descriptors were selected successively by correlation analysis, stepwise-MLR method and SVM feature selection tool. Using the selected descriptors, we developed a novel QSAR model applying SVC method to classify three-class (highly, moderately and weakly active) of Akt inhibitors with diverse structures. Furthermore, SVR method was employed to set up prediction model to explore the IC_{50} values of the inhibitors in order to further increase the accuracy of estimated activities. Finally, enrichment factor and receiver operating characteristic

* Corresponding author. Tel./fax: +86 571 88208460.
E-mail address: huyz@zju.edu.cn (Y. Hu).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.04.050

(ROC) studies were performed either using SVR model alone or assisted with SVC model in order to validate the reliability of the established models.

2. Materials and methods

2.1. Data preparation

For the QSAR studies, 148 Akt inhibitors were selected from literature [7–13, 19,20]. Structures of these compounds are listed in Fig. 1. In our study, $-\log(IC_{50})$ values were used as the dependent variables, given in Table 1. Among these inhibitors, 104 molecules were selected as the training set considering both structural diversity and wide coverage of biological activity ranges (from 0.16 to 32,250 nM), while the remaining compounds (44 molecules) also covering wide range of activity spanning and structural skeletons constituted the test set. For estimation (prediction) purposes, the activity values were classified based on an Akt inhibitory activity scale: highly active (+++, $IC_{50} < 100$ nM); moderately active (++, 100 nM $< IC_{50} <$ 5000 nM), and weak active compounds (+, $IC_{50} >$ 5000 nM). And the 1264 negatives compounds used in enrichment and ROC studies were retrieved from Available Chemical Directory (ACD) database (Symyx Technologies, Santa Clara, CA) using "Random Percent Filter protocol" by Pipeline Pilot software (SciTegic, Inc., San Diego, CA). All the compounds were optimized in Discovery Studio 2.0 software (Accelrys, Inc. San Diego, CA).

2.2. Descriptors calculation and selection

The resulted geometry of molecules were inputted into Dragon software [21], which can calculate constitutional descriptors, topological descriptors, walk and path counts, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalue descriptors, etc. After the calculation of the molecular descriptors, those that stayed constant for all molecules were eliminated, and pairs of variables with a correlation coefficient greater than 0.75 were classified as intercorrelated with one of each correlated pair deleted. Then stepwise multiple linear regression (Stepwise-MLR) and SVM feature selection tool [22] were used to select the most relevant descriptors.

2.3. SVM model development

The major advantage of support vector machine (SVM) [23] lies in its adoption of the structure risk minimization (SRM) principle which has been proved to be superior to the traditional empirical risk minimization (ERM) principle employed by conventional neural networks.

One-against-one method is proposed for solving multiclass problem in support vector machine classification (SVC) [22]. This method constructs $k(k-1)/2$ hyperplanes where each one is built using the training data chosen out of k classes. The decision function for class pair ij is defined by:

$$f_{ij}(x) = \langle \phi(x) w^{ij} \rangle + b^{ij}.$$

It is found by solving the following optimization problem:

$$\min \frac{1}{2}\left(\left\| \omega^{ij} \right\|^2 + C \sum_n \xi_n \right)$$

$$\text{subject to} \quad \begin{array}{ll} \langle \omega^{ij} \phi(x_n) \rangle + b^{ij} \geq 1 - \xi_n^{ij}, & \text{if } y_n = i \\ \langle \omega^{ij} \phi(x_n) \rangle + b^{ij} \geq -1 + \xi_n^{ij}, & \text{if } y_n = j \end{array} \qquad \xi_t^{ij} \geq 0,$$

for all n examples in classes i and j.

Since $f_{ij}(x) = -f_{ji}(x)$, there exist $k(k-1)/2$ different decision functions for a k-class problem. This method fits perfectly to the known characteristics of the SVM, where the borderlines between two classes are computed directly. The most popular method for the class identification of the one-against-one method is the "max wins" algorithm. In the "max wins" algorithm each classifier casts one vote for its preferred class, and the final result is the class with the most votes:

$$\text{The class of } x = \arg\max_i \sum_{j \neq i, j=1}^{k} \text{sign}(f_{ij}(x)),$$

where sign () is the sign function, i.e. its value is 1 when f_{ij} is positive and 0 otherwise. When more than one class has the same number of votes, i.e. a tie situation arises, each point in the unclassifiable (tie) region is assigned to the closest class using the real valued decision functions as:

$$\text{The class of } x = \arg\max_i \sum_{j \neq i, j=1}^{k} f_{ij}(x).$$

In support vector regression (SVR) [22], the basic idea is to map the data x into a higher-dimensional feature space F via a nonlinear mapping and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}_i^n$ (x_i is the input vector, d_i is the desired value, and n is the total number of data patterns). SVR approximates the function in the following form:

$$y = \sum_{i=1}^{l} w_i \phi(x_i) + b, \tag{1}$$

where $\phi(x)$ is the high dimensional feature space, which is nonlinearly mapped from the input space x, and w_i and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\xi(d_i, y_i) + \frac{1}{2}\|w\|^2, \tag{2}$$

where

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

The first term $C1/N\sum_{i=1}^{N} L_\xi(d_i, y_i)$ is the so-called empirical error (risk), which is measured by the ε-insensitive loss function (3). The second term $1/2\|w\|^2$, on the other hand, is called the regularized term. ε is called the tube size of SVM, and C is the regularization constant determining the trade-off between the empirical error and the regularized term. Introduction of slack variables 'ξ' leads Eq. (2) to the following constrained function:

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

Thus, decision function of Eq. (1) changes to the following form:

$$f\left(x, \alpha_i, \alpha_i^*\right) = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)K(x, x_i) + b, \tag{5}$$

where α_i and α_i^* are the introduced Lagrange multipliers and $K(x, x_i)$ is the kernel function. The value is equal to the inner product of two vectors x and x_i in the feature space $\Phi(x)$ and $\Phi(x_i)$, that is, $K(x,x_i) = \Phi(x)^T\Phi(x_i)$. And the radial basis function (RBF) kernel $K(\bar{x}_i, \bar{x}_j) = \exp(-\gamma\|\bar{x}_i - \bar{x}_j\|^2)$ is commonly used.

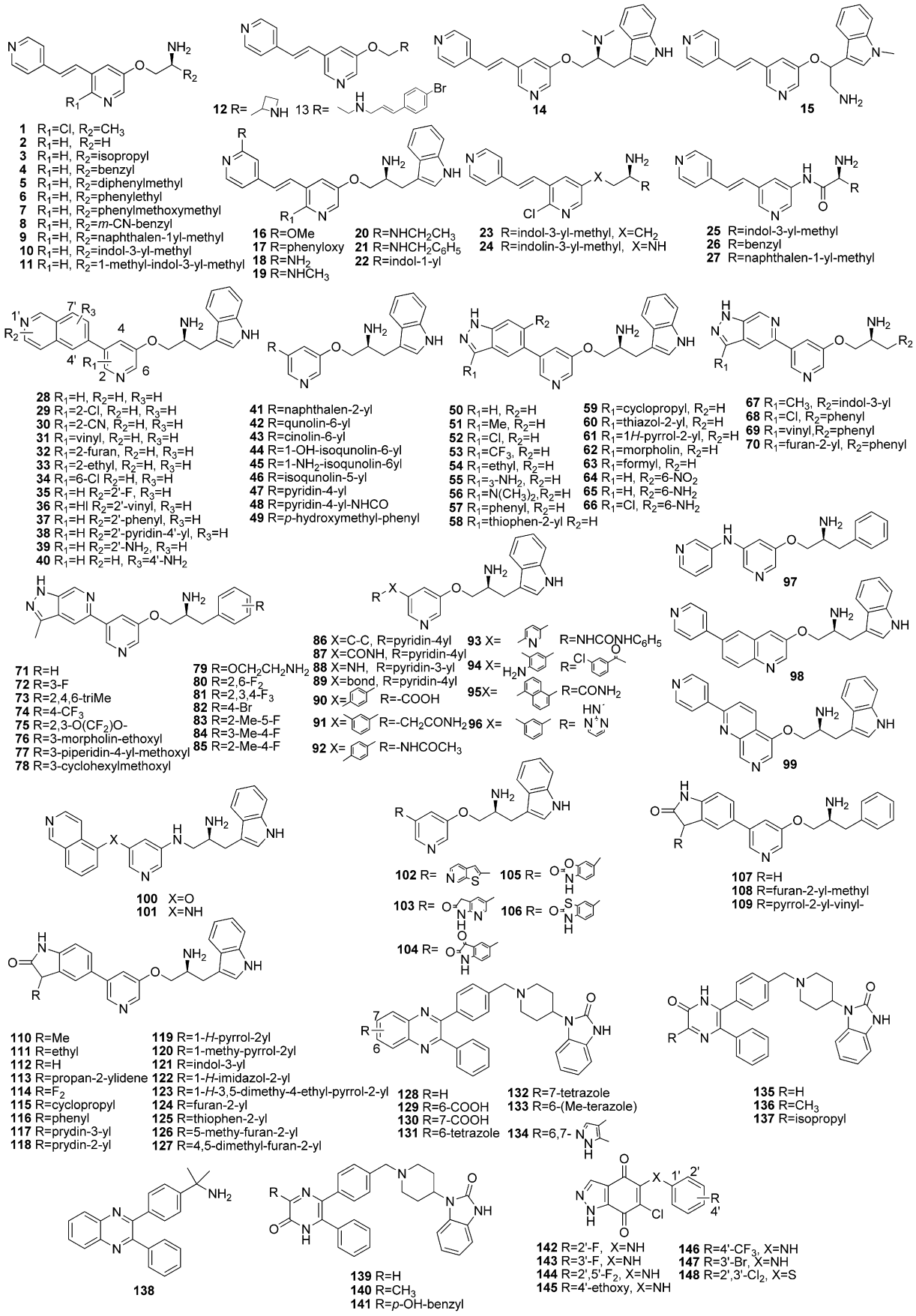

Fig. 1. The structures of Akt inhibitors used for the SVM model development.

Table 1
Compounds in training set and test set for SVC and SVR models, and their corresponding experimental and theoretical activities.

Compound no.	IC_{50} (nM)	Class[a]	−Log IC_{50}	SVC[a]	SVR	Compound no.	IC_{50} (nM)	Class[a]	−Log IC_{50}	SVC[a]	SVR
1	5290	+	5.28	+	5.81	**75**[b]	14	+++	7.85	+++	7.84
2	14710	+	4.83	+	4.87	**76**[c]	751	++	6.12	++	6.64
3	19320	+	4.71	+	5.05	**77**	6.8	+++	8.17	+++	7.48
4	690	++	6.16	++	5.39	**78**	1360	++	5.87	++	6.21
5	273	++	6.56	++	6.41	**79**	47	+++	7.33	+++	7.03
6	8020	+	5.10	++	5.44	**80**	65.5	+++	7.18	+++	7.43
7[b]	4040	++	5.39	++	4.65	**81**[b]	12	+++	7.92	+++	7.50
8	320	++	6.49	++	5.62	**82**[c]	5	+++	8.30	+++	7.91
9	324	++	6.49	++	5.83	**83**	3.9	+++	8.41	+++	7.66
10	360	++	6.44	++	6.38	**84**	2.5	+++	8.60	+++	8.26
11	500	++	6.30	++	6.64	**85**[b]	6	+++	8.22	+++	7.68
12[b]	10870	+	4.96	+	5.12	**86**	210	++	6.68	+++	7.39
13	32250	+	4.49	+	4.83	**87**	2020	++	5.69	++	5.85
14	1100	++	5.96	++	6.47	**88**	2160	++	5.67	++	6.73
15[b]	760	++	6.12	++	6.51	**89**[b]	260	++	6.59	++	6.78
16	2670	++	5.57	++	5.48	**90**	31280	+	4.50	++	4.84
17[b]	6450	+	5.19	++	5.30	**91**	1390	++	5.86	++	6.20
18	121	++	6.92	++	6.58	**92**[b]	227	++	6.64	++	5.94
19[b]	176	++	6.75	++	6.52	**93**	1120	++	5.95	++	5.56
20	163	++	6.79	++	6.62	**94**	1450	++	5.84	++	5.50
21	147	++	6.83	++	6.49	**95**[b]	2170	++	5.66	++	5.98
22[b]	3340	++	5.48	++	5.95	**96**	3670	++	5.44	++	6.37
23	198	++	6.70	++	6.78	**97**	6970	+	5.16	+	5.73
24[c]	79	+++	7.10	+++	7.44	**98**	300	++	6.52	++	7.00
25	278	++	6.56	++	6.33	**99**[b]	190	++	6.72	++	6.63
26	3200	++	5.49	++	5.45	**100**	210	++	6.68	++	7.02
27	668	++	6.18	++	5.84	**101**[b]	270	++	6.57	++	7.27
28	2	+++	8.70	+++	8.36	**102**[b]	179	++	6.75	++	6.78
29	1.8	+++	8.74	+++	8.44	**103**	7.4	+++	8.13	+++	7.79
30[b]	4.6	+++	8.70	+++	8.89	**104**[b,c]	5.4	+++	8.27	+++	8.04
31	2	+++	8.70	+++	8.21	**105**	4.3	+++	8.37	+++	8.55
32[b]	1.5	+++	8.82	+++	9.14	**106**[b]	99	+++	7.00	+++	8.19
33[c]	0.8	+++	9.10	+++	9.07	**107**	16	+++	7.80	+++	7.55
34[b,c]	12	+++	7.92	+++	8.48	**108**[b]	1.4	+++	8.85	+++	7.28
35	3.5	+++	8.46	+++	8.58	**109**[b]	0.6	+++	9.22	+++	7.49
36	22	+++	7.66	+++	8.00	**110**	4	+++	8.40	+++	8.46
37	305	++	6.52	++	7.45	**111**[c]	52.7	+++	7.28	+++	7.61
38[b]	122	++	6.91	+++	7.41	**112**[b]	3.2	+++	8.49	+++	8.34
39[b]	3.4	+++	8.47	+++	8.40	**113**	80	+++	7.10	+++	8.34
40[c]	15	+++	7.82	+++	8.44	**114**	1.5	+++	8.82	+++	8.48
41[b]	1117	++	5.95	++	7.16	**115**	3.3	+++	8.48	+++	8.41
42	312	++	6.51	++	7.00	**116**[b]	5.9	+++	8.23	+++	7.65
43[c]	215	++	6.67	++	6.87	**117**	10.4	+++	7.98	+++	7.64
44	1022	++	5.99	++	6.28	**118**	10.3	+++	7.99	+++	7.65
45	188	++	6.73	++	7.25	**119**[b]	1.5	+++	8.82	+++	8.50
46[b]	331	++	6.48	+++	6.99	**120**	8.2	+++	8.09	+++	8.50
47[b,c]	245	++	6.61	++	6.95	**121**	153	++	6.82	++	7.16
48	1659	++	5.78	++	6.12	**122**	7.2	+++	8.14	+++	8.48
49	4022	++	5.40	++	6.27	**123**	40.3	+++	7.39	+++	7.05
50	1.5	+++	8.82	+++	8.15	**124**	0.17	+++	9.77	+++	8.77
51[c]	0.16	+++	9.80	+++	9.46	**125**	0.9	+++	9.05	+++	8.81
52	1	+++	9.00	+++	8.46	**126**	0.7	+++	9.15	+++	8.66
53[b]	1.8	+++	8.74	+++	8.39	**127**	10	+++	8.00	+++	8.34
54	1.15	+++	8.94	+++	8.41	**128**[c]	290	++	6.54	++	6.88
55	2.9	+++	8.54	+++	8.43	**129**[b]	240	++	6.62	++	5.97
56[b]	18	+++	7.74	+++	8.30	**130**[c]	166	++	6.78	++	6.44
57	3.9	+++	8.41	++	7.61	**131**	63	+++	7.20	+++	6.95
58	1.3	+++	8.89	+++	8.89	**132**[c]	20	+++	7.70	+++	7.36
59[b,c]	2.6	+++	8.59	+++	8.69	**133**[b]	1089	++	5.96	+++	6.63
60[b,c]	9.8	+++	8.01	+++	8.87	**134**[b]	85	+++	7.07	+++	7.90
61[c]	1.2	+++	8.92	+++	8.58	**135**	1500	++	5.82	++	6.00
62[c]	18	+++	7.74	+++	7.69	**136**	1003	++	6.00	++	5.92
63	1848	++	5.73	++	6.07	**137**	21200	+	4.67	+	5.01
64[b]	953	++	6.02	+++	6.73	**138**	3400	++	5.47()	++	5.81
65	51	+++	7.29	+++	8.25	**139**[b]	3029	++	5.52	++	5.96
66	7.1	+++	8.15	+++	8.49	**140**[b]	760	++	6.12	++	5.96
67	0.34	+++	9.47	+++	8.05	**141**	21670	+	4.66	+	4.64
68	59	+++	7.23	+++	7.27	**142**	15800	+	4.80	+	4.98
69	13	+++	7.89	++	6.93	**143**	11800	+	4.93	+	4.96
70[b,c]	6.1	+++	8.21	+++	7.57	**144**[b]	16000	+	4.80	+	5.09
71	14	+++	7.85	+++	7.32	**145**	4900	++	5.31	++	4.97
72[b]	8.4	+++	8.08	+++	7.48	**146**	12400	+	4.91	+	5.25
73[c]	39	+++	7.41	+++	7.75	**147**	13900	+	4.86	+	5.19
74	18	+++	7.74	+++	7.58	**148**[b]	12300	+	4.91	+	5.10

[a] Akt inhibitory activity scale: highly active (+++, $IC_{50} < 100$ nM); moderately active (++, 100 nM $< IC_{50} <$ 5000 nM), and weakly active compounds (+, $IC_{50} > 5000$ nM).
[b] The compounds were used as test set.
[c] The compounds were used as active compounds (positive) in enrichment factor and ROC studies.

2.4. Models validation

Validation of the models was required to test the predictive ability and generalization of the methods by internal data set (cross-validation) as well as external data. In this work, each data set was divided into training set for model development and test set for external prediction. For the training set, both the SVC and SVR models underwent a leave-one-out (LOO) procedure. The stability of the correlations of SVC and SVR models were tested against the accuracy and cross-validated coefficient R^2_{cv}, respectively, which describes the stability of a model obtained by focusing on the sensitivity of the model to the elimination of any single data point.

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} d_i}{n}$$

where the value d_i is 1 when the output class is consistent with the actual class and 0 otherwise, and n is the number of compounds in the analyzed set.

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left[y_i - \left(\sum_{i=1}^{n} y_i\right)\Big/ n\right]^2},$$

where y_i is the desired output and $\hat{y}_i$ the actual output of the model, and n is the number of compounds in the analyzed set.

2.5. Enrichment factor and ROC studies

In the lead-discovery studies, the QSAR models should identify active leads against Akt protein kinase in the database screening. Therefore, the generated models were further validated by screening an in-house database which was retrieved randomly from ACD database that spiked some known inhibitors, and the enrichment factor (EF) and receiver operating characteristic (ROC) studies were performed.

$$\text{Enrichment factor (EF)} = \frac{\text{TP}}{\text{TP}+\text{FP}}\frac{N}{n},$$

where TP is the number of true positive compounds, FP is the number of false positive compounds, N is the number of the total compounds and n is the number of the total positive compounds.

Table 2
The involved molecular descriptors and their corresponding definition.

Symbol	Class	Definition
Se	Sum of atomic Sanderson electronegativities	Constitutional descriptors
Qindex	Quadratic index topological descriptors	Topological descriptors
Rww	Reciprocal hyper-detour index topological descriptors	Topological descriptors
D/Dr05	Distance/detour ring index of order 5	Topological descriptors
T(N···O)	Sum of topological distances between N···O	Topological descriptors
X0	Connectivity index chi-0	Connectivity indices
RDSQ	Reciprocal distance squared Randic-type index	Connectivity indices
HOMT	HOMA total	Geometrical descriptors
QYYm	Qyy COMMA2 value/weighted by atomic masses	Geometrical descriptors
G(N···N)	Sum of geometrical distances between N···N	Geometrical descriptors
G(O···O)	Sum of geometrical distances between O···O	Geometrical descriptors
RDF025u	Radial distribution function – 2.5	RDF descriptors
RDF035u	Radial distribution function – 3.5	RDF descriptors
Mor03u	Mor03u 3D-MoRSE – signal 03	3D-MoRSE descriptors
L2u	2nd component size directional WHIM index	WHIM descriptors
Au	A total size index	WHIM descriptors
Ui	Unsaturation index	Molecular properties

In addition, the receiver operating characteristic (ROC) study [24] was performed to calculate sensitivity (Se) and specificity (Sp) from a comparison between *in vitro* and *in silico*.

$$\text{Se} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

where TP is the number of true positive compounds and FN is the number of false negative compounds.

$$\text{Sp} = 1 - \frac{\text{TN}}{\text{TN}+\text{FP}},$$

where TN is the number of true negative compounds, FP is the number of false positive compounds.

The ROC curve is a function of $(1 - \text{Sp})$ versus Se, and the area under the ROC curve (AUC) is the important way of measuring the performance of the test.

$$\text{AUC} = \sum_{x=2}^{N} \text{Se}(x)[(1-\text{Sp})(x) - (1-\text{Sp})(x-1)],$$

where $\text{Se}(x)$ is the percent of the true positives versus the total positives at rank position x, $(1 - \text{Sp})(x)$ is the percent of the false positives versus the total negatives at rank position x.

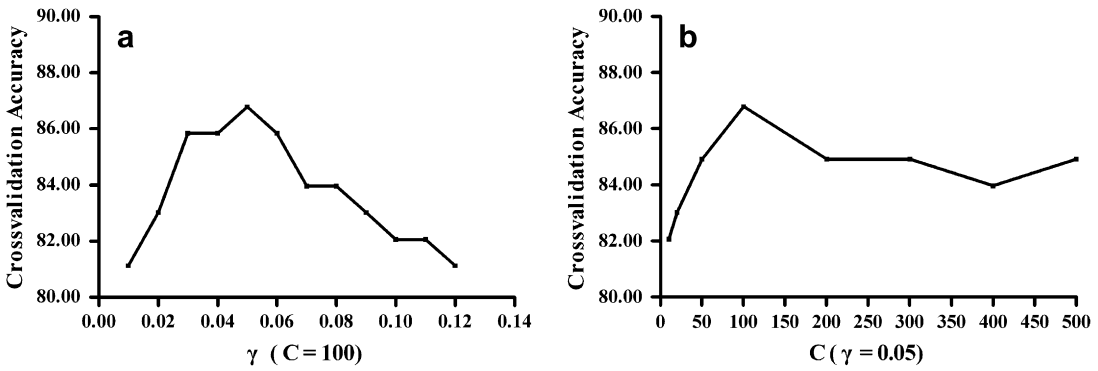

Fig. 2. Selection of gamma (γ) and cost (C) criteria for a training data set in three-class SVM model development. (a) γ versus accuracy on LOO cross-validation. (b) C versus accuracy on LOO cross-validation.

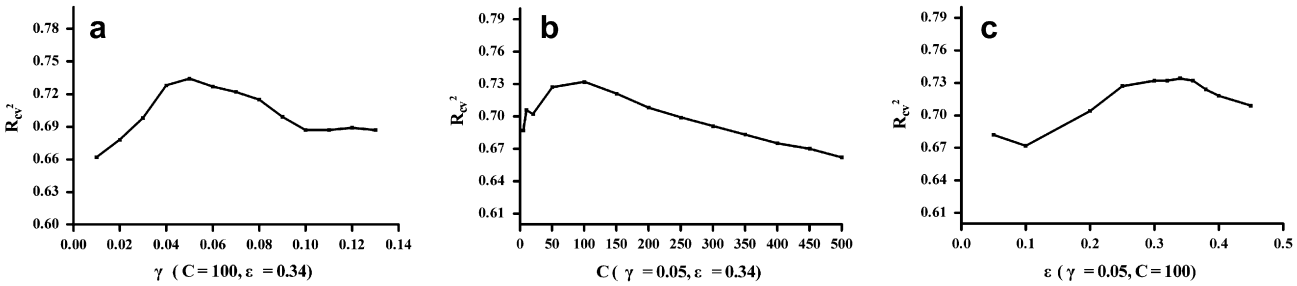

Fig. 3. Selection of γ, C and ε for the training data set in the SVR model development. (a) γ versus R^2_{cv} on LOO cross-validation. (b) C versus R^2_{cv} on LOO cross-validation. (c) ε versus R^2_{cv} on LOO cross-validation.

3. Results and discussion

As shown above, the descriptors of the compounds were calculated by Dragon software. Successively, inter-relationship analysis, stepwise-MLR method and SVM features selection tool were employed to select appropriate descriptors for QSAR models development. The physical-chemical meanings of the selected 17 descriptors are listed in Table 2.

3.1. SVC model

Three-class support vector classification (SVC) was performed using the selected 17 descriptors. For purpose of modeling, a value of "+++" was assigned to those with highly active properties, a value of "++" was assigned to those with moderately active properties, and a value of "+" was assigned to those with weakly active properties. Similar to other multivariate statistical models, the performances of SVC depend on the combination of several parameters. For classification tasks, radial basis function (RBF) kernel function is commonly used because of its good generalization performance and a few useful parameters. Once the kernel function has been decided, width of RBF (γ) and capacity parameter (C) should be optimized. Here, the optimal parameters are found by grid search (GS) method. In the grid search, we considered parameter lg2γ from −10 to 10 with 1 as the increment. Parameter lg2C was chosen from −5 and 10 with 1 as the increment. The result of this grid search is an error-surface spanned by the model parameters. A robust model is obtained by selecting those parameters that give the lowest error in a smooth area. To find the optimized combination of the parameters C and γ, a process of LOO cross-validation of the training set was performed as shown in Fig. 2. The best choices for C and γ were 100 and 0.05, respectively, and the corresponding support vector number of 53. Furthermore, the predictive ability of the model was tested by the compounds from test set.

The results obtained in the classification of the compounds that make up the training set, test set and overall set are shown in Fig. 4a. The accuracy in prediction for the training, test and overall data sets are 95.2%, 88.6% and 93.1%, respectively, and cross-validation accuracy of this model is 86.8%. It was suggested that the three-class SVC model derived in this study would be valuable and reliable in classifying structurally diverse compounds with Akt inhibitory activity.

3.2. SVR model

Based on the derived SVC model, novel designed Akt inhibitors could be "*in silico*" classified into three active classes (highly, moderately and weakly active), which were very valuable in development of new Akt inhibitors. More often, however, more detailed data, such as IC_{50} were requested. Therefore, in order to further increase the accuracy of estimated activities, the SVR method was employed to set up a QSAR model to explore the IC_{50} of the inhibitors. In this study, the RBF kernel was used as kernel function. Thus capacity parameter C, ε of ε-insensitive loss function and the corresponding parameters γ of RBF kernel need to be optimized. Here, the optimal parameters were initially found by grid search method performed as SVC model development, and a robust model was obtained. In order to find the optimized combination of the parameters C, ε, and γ, a process of LOO cross-validation of training set was performed as shown in Fig. 3. The best choices for C, ε and γ, were 100, 0.34 and 0.05, respectively, and the corresponding support vector number is 71.

The predicted results are listed in Table 1, and the plots of predicted versus experimental values for training and test sets are recorded in Fig. 4b. The square correlation coefficient for the training, test and overall data sets were 0.882, 0.762 and 0.840,

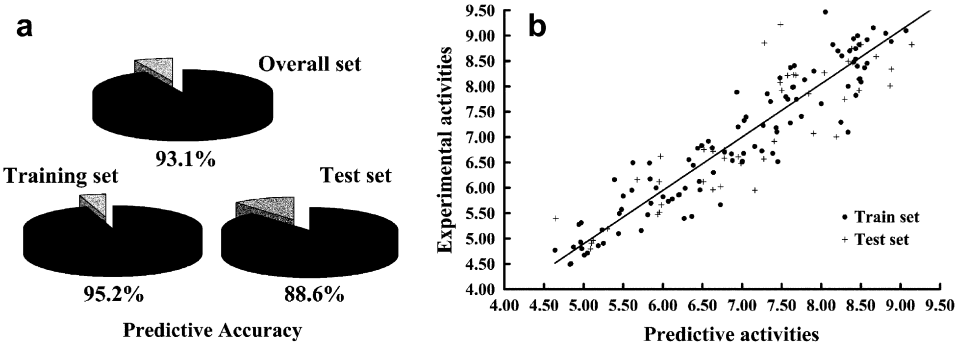

Fig. 4. (a) Description of the overall, training and test sets' classification accuracy percentages of the SVC model; (b) Predicted versus experimental Akt inhibitory activities (−Log IC_{50}) of the SVR model.

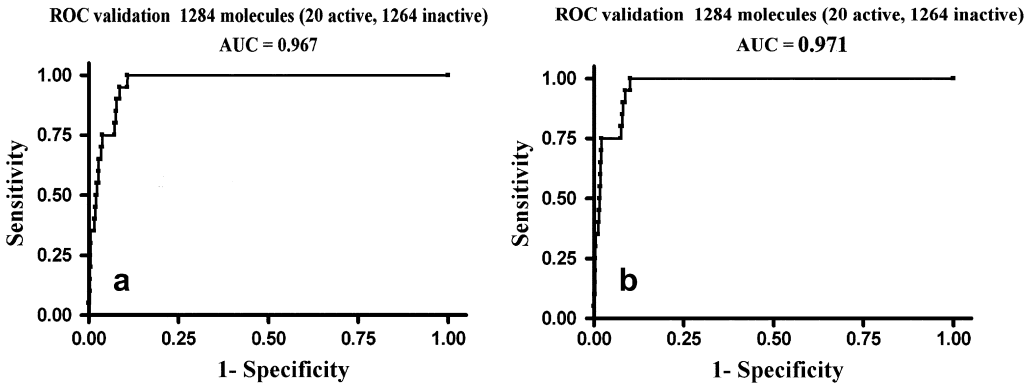

Fig. 5. Receiver operating characteristic (ROC) curve. (a) ROC curve of the database screening by SVR model; (b) ROC curve of the database screening by SVR assisted with SVC model.

respectively, and cross-validation coefficient of this model was 0.734. The prediction results indicated that SVR is a potent and promising tool for QSAR analysis and more detailed IC_{50} values could be obtained comparing with SVC method.

3.3. Combination of SVC and SVR model

As shown in Table 1, the predicted reliability was demonstrated by the consistent results between SVR and SVC models, with compounds **6**, **38**, **69** and **86** as exceptions. A total of 12 compounds were misclassified base on calculated $-\text{Log}\,IC_{50}$ in SVR model. Except for compounds **6**, **38**, **69** and **86**, the results of compounds **7**, **37**, **41**, **45**, **97**, **101**, **121** and **131** could be corrected by SVC model. For example, compound **7**, whose actual $-\text{Log}\,IC_{50}$ was 5.39 (moderately active class), was misclassified into weakly active class (predicted $-\text{Log}\,IC_{50}$ 4.65) by SVR method, while the result of SVC model indicated that compound **7** was a moderately active compound. Therefore, combination utilization of SVR and SVC methods could further decrease the error rate.

3.4. Enrichment factor and ROC studies

The SVR model was further validated for picking up active molecules from the database against Akt protein kinase. A spiked database including 1264 inactive compounds and 20 known inhibitors (15 highly active and 5 moderately active molecules, Table 1) was involved in this validation experiment. The established model picked up all 20 known inhibitors giving enrichment factors as 16.46, 9.35 and 8.33 at 3%, 8% and 12%, respectively. And the ROC curves of the database screening by SVR model are shown in Fig. 5a and AUC was 0.967. Furthermore, the generated SVC model was used to predict the same database, and compounds whose SVR model result was not consistent with that of SVC model were rearranged. Then the enrichment factor was enhanced to 23.04, 9.35 and 8.34 at 3%, 8% and 12%, while the AUC of ROC curve was increased to 0.971 (Fig. 5). These results demonstrated that the SVR model derived could be a useful and reliable tool in identifying Akt inhibitors, and the predictive ability of SVR assisted with SVC model was improved.

4. Conclusions

In this work, we applied SVR and SVC methods to explore the activities of 148 Akt inhibitors and structural descriptors. Both methods showed potent and reliable properties in QSAR model development. Besides, when the SVR model was assisted with SVC model, the predicted results were further more reliable and accurate. Then enrichment factor and ROC studies were performed either using SVR model alone or assisted with SVC model, the results of which demonstrated that the models established could be a useful and reliable tool in identifying Akt inhibitors.

Acknowledgement

The authors are grateful to the support provided by the Young Talent Fostering Program of Zhejiang Province (2008R40G2010063) and the support from College of pharmaceutical sciences Zhejiang University for providing Dragon 5.0 software.

Reference

[1] B.J. Druker, Adv. Cancer Res. 91 (2004) 1–30.
[2] D.S. Krause, R.A. Van Etten, N. Engl. J. Med. 353 (2005) 172–187.
[3] J.S. Ross, D.P. Schenkein, R. Pietrusko, M. Rolfe, G.P. Linette, J. Stec, N.E. Stagliano, G.S. Ginsburg, W.F. Symmans, L. Pusztai, G.N. Hortobagyi, Am. J. Clin. Pathol. 122 (2004) 598–609.
[4] G. Powis, N. Ihle, D.L. Kirkpatrick, Clin. Cancer Res. 12 (2006) 2964–2966.
[5] X. Liu, Y. Shi, E.K. Han, Z. Chen, S.H. Rosenberg, V. Giranda, Y. Luo, S.C. Ng, Neoplasia 3 (2001) 278–286.
[6] J.Q. Cheng, B. Ruggeri, W.M. Klein, G. Sonoda, D.A. Altomare, D.K. Watson, J.R. Testa, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 3636.
[7] C.W. Lindsley, Z. Zhao, W.H. Leister, R.G. Robinson, S.F. Barnett, D. Defeo-Jones, R.E. Jones, G.D. Hartman, J.R. Huff, H.E. Huberb, M.E. Duggan, Bioorg. Med. Chem. Lett. 15 (2005) 761–764.
[8] Q. Li, T. Li, G. Zhu, J. Gong, A. Claibone, C. Dalton, Y. Luo, E.F. Johnson, Y. Shi, X. Liu, V. Klinghofer, J.L. Bauch, K.C. Marsh, J.J. Bousk, S. Arries, R.D. Jong, T. Oltersdorf, V.S. Stoll, C.G. Jakob, S.H. Rosenberg, V.L. Giranda, Bioorg. Med. Chem. Lett. 16 (2006) 1679–1685.
[9] G. Zhu, J. Gong, A. Claiborne, K.W. Woods, V.B. Gandhi, S. Thomas, Y. Luo, X. Liu, Y. Shi, R. Guan, S.R. Magnone, V. Klinghofer, E.F. Johnson, J. Bouska, A. Shoemaker, A. Oleksijew, V.S. Stoll, R.D. Jong, T. Oltersdorf, Q. Li, S.H. Rosenberg, V.L. Giranda, Bioorg. Med. Chem. Lett. 16 (2006) 3150–3155.
[10] K.W. Woods, J.P. Fischer, A. Claiborne, T. Li, S.A. Thomas, G. Zhu, R.B. Diebold, X. Liu, Y. Shi, V. Klinghofer, E.K. Han, R. Guan, S.R. Magnone, E.F. Johnson, J.J. Bouska, A.M. Olson, R.D. Jong, T. Oltersdorf, Y. Luo, S.H. Rosenberg, V.L. Giranda, Q. Li, Bioorg. Med. Chem. 14 (2006) 6832–6846.
[11] J.H. Ko, S.W. Yeon, J.S. Ryu, T.Y. Kim, E.H. Song, H.J. You, R.E. Parka, C.K. Ryu, Bioorg. Med. Chem. Lett. 16 (2006) 6001–6005.
[12] Q. Li, K.W. Woods, S. Thomas, G. Zhu, G. Packard, J. Fisher, T. Li, J. Gong, J. Dinges, X. Song, J. Abrams, Y. Luo, E.F. Johnson, Y. Shi, X. Liu, V. Klinghofer, R. Jong, T. Oltersdorf, V.S. Stoll, C.G. Jakob, S.H. Rosenberga, V.L. Giranda, Bioorg. Med. Chem. Lett. 16 (2006) 2000–2007.
[13] G. Zhu, V.B. Gandhi, J. Gong, S. Thomas, K.W. Woods, X. Song, T. Li, R.B. Diebold, Y. Luo, X. Liu, R. Guan, V. Klinghofer, E.F. Johnson, J. Bouska, A. Olson, K.C. Marsh, V.S. Stoll, M. Mamo, J. Polakowski, T.J. Campbell, R.L. Martin, G.A. Gintant, T.D. Penning, Q. Li, S.H. Rosenberg, V.L. Giranda, J. Med. Chem. 50 (2007) 2990–3003.
[14] E. Estrada, Mini Rev. Med. Chem. 8 (2008) 213–221.

[15] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, J. Chem. Inf. Comput. Sci. 43 (2003) 1882–1889.
[16] C.W. Yap, H. Li, Z.L. Ji, Y.Z. Chen, Mini Rev. Med. Chem. 7 (2007) 1097–1107.
[17] A.Z. Dudek, T. Arodz, J. Galvez, Comb. Chem. High Throughput Screen. 9 (2006) 213–228.
[18] X. Dong, Y. Liu, J. Yan, C. Jiang, J. Chen, T. Liu, Y. Hu, Bioorg. Med. Chem. 16 (2008) 8151–8160.
[19] G. Zhu, V.B. Gandhi, J. Gong, Y. Luo, X. Liu, Y. Shi, R. Guan, S.R. Magnone, V. Klinghofer, E.F. Johnson, J. Bouska, A. Shoemaker, A. Oleksijew, K. Jarvis, C. Park, R.D. Jong, T. Oltersdorf, Q. Li, S.H. Rosenberg, V.L. Giranda, Bioorg. Med. Chem. Lett. 16 (2006) 3424.
[20] G. Zhu, J. Gong, V.B. Gandhi, K.W. Woods, Y. Luo, X. Liu, R. Guan, V. Klinghofer, E.F. Johnson, V.S. Stoll, M. Mamo, Q. Li, S.H. Rosenberg, V.L. Giranda, Bioorg. Med. Chem. 15 (2007) 2441.
[21] Talete srl, DRAGON for windows (software for molecular descriptor calculation) <http://talete.mi.it>, 2006.
[22] C. Chang, C. Lin, LIBSVM: A Library for Support Vector Machines Software (2001) available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
[23] V. Vapnik, A. Lerner, Automat. Remote Contr. 24 (1963) 774–780.
[24] N. Triballeau, F. Acher, I. Brabet, J.P. Pin, H.O. Bertrand, J. Med. Chem. 48 (2005) 2534–2547.