Taylor & Francis
Taylor & Francis Group

Journal of Enzyme Inhibition and Medicinal Chemistry

ISSN: 1475-6366 (Print) 1475-6374 (Online) Journal homepage: informahealthcare.com/journals/ienz20

QSAR studies on 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors using robust PCA and both linear and nonlinear models

Min Sun, Youguang Zheng, Hongtao Wei, Junqing Chen & Min Ji

To cite this article: Min Sun, Youguang Zheng, Hongtao Wei, Junqing Chen & Min Ji (2009) QSAR studies on 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors using robust PCA and both linear and nonlinear models, Journal of Enzyme Inhibition and Medicinal Chemistry, 24:5, 1109-1116, DOI: 10.1080/14756360802632906

To link to this article: https://doi.org/10.1080/14756360802632906

Published online: 07 Apr 2009.

Journal of Enzyme Inhibition and Medicinal Chemistry, 2009; 24(5): 1109–1116

informa
healthcare

RESEARCH ARTICLE

QSAR studies on 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors using robust PCA and both linear and nonlinear models

Min Sun, Youguang Zheng, Hongtao Wei, Junqing Chen, and Min Ji

Pharmaceutical Engineering Institute, School of Chemistry and Chemical Engineering, Southeast University, 2 Sipailou, Nanjing 210096, China

Abstract

Quantitative structure-activity relationship (QSAR) studies have been carried out on 4-anilino-3-quinolinecarbonitriles, a set of novel Src kinase inhibitors, with the aim of dissecting the structural requirements for Src inhibitory activities. After outlier identification using robust principal component analysis (robust PCA), linear models based on forward selection combined with multiple linear regression, (FS-MLR), enhanced replacement method followed by multiple linear regression (ERM) and a nonlinear model using support vector regression (SVR) were constructed and compared. All models were rigorously validated using leave-one-out cross-validation (LOOCV), 5-fold cross-validations (5-CV) and shuffling external validation (SEVs). ERM seems to outperform both FS-MLR and SVR evidenced by better prediction performance (n=35, $R^2_{training}$ = 0.918, R^2_{pred} = 0.928). Robustness and predictive ability of ERM model were also evaluated. The generated QASR model revealed that the Src inhibitory activity of 4-anilino-3-quinolinecarbonitriles could be associated with the size of substituents in the C7 position and the steric hindrance effect. The results of the present study may be of great help in designing novel 4-anilino-3-quinolinecarbonitriles with more potent Src kinase inhibitory activity.

Keywords: *QSAR; robust PCA; ERM; 4-anilino-3-quinolinecarbonitriles; Src kinase inhibitors*

Introduction

The Src family kinases (SFKs), a family of nonreceptor tyrosine kinases, including Src, Yes, Lck, Fyn, Lyn, Fgr, Hck, Blk and Yrk, are involved in the regulation of a wide variety of normal cellular transduction pathways, such as cell growth, differentiation, survival, adhesion and migration [1], and are maintained in an inactive conformation in the absence of extracellular and intracellular stimuli. However, considerable evidences implicates elevated expression and/or activity of Src kinases in many human cancers (e.g., colon, rectal, or stomach cancer) [2,3], osteoporosis [4], cardiovascular disorders [5] and immune system dysfunction [6]. Thus, this family of protein tyrosine kinase now exists as intriguing targets for both basic research and drug discovery.

Currently, numerous efforts have been devoted to the design of Src kinase inhibitors, with most attentions through an ATP-competitive inhibition mechanism. Several Src kinase inhibitors have been identified to date. These include various heteroaromatic compounds, such as pyrazolopyrimidines, pyrrolopyrimidines, pyridopyrimidines, quinazolines, quinolines, indolinones, isoquinoline and others [7]. A series of 4-anilino-3-quinolinecarbonitriles developed by Boschelli and co-workers [8–10] exhibited potent Src kinase inhibiting activity. Considering the important role of Src kinase in regulating normal cellular functions and recent interest in development of such inhibitors, a QSAR investigation of these compounds is carried out. The objective of this study is to analyze the physicochemical and structural requirements of these inhibitors to exhibit optimal inhibitory potency of Src kinase which will in turn help in rationalizing the design of these molecules as Src kinase inhibitors, as well as to provide a strategy for predicting activities of novel 4-anilino-3-quinolinecarbonitriles with high accuracy.

Address for Correspondens: Min Ji Tel.: (+86)(25) 83272349 Fax: (+86)(25) 86636901 E-mail: jimin@seu.edu.cn

(Received 10 September 2008; revised 31 October 2008; accepted 10 November 2008)

ISSN 1475-6366 print/ISSN 1475-6374 online © 2009 Informa UK Ltd
DOI: 10.1080/14756360802632906

Materials and methods

Dataset and descriptors

The data set consisting of 37 4-anilino-3-quinolinecarbonitriles together with their Src kinase inhibitory activities, expressed as $\log(1/IC_{50})$, was obtained from references [8–10]. The structure of the compounds were drawn using MDL® ISIS/Draw (Symyx Technologies, Inc.) implemented in the ISIS 2.5 package and pre-optimized using molecular mechanics force field (MM+) encoded in HyperChem (Version 8.04, Hypercube, Inc.). The final refined equilibrium molecular geometrics were obtained using the semiempirical method PM3 (Parametric Method-3). We chose a gradient norm limit of 0.01 kcal/A° for the geometry optimization. More than one thousand meaningful descriptors were calculated for each compound using E-Dragon version 1.1 [11], encoding different aspects of the molecular structures. These descriptors consist of constitutional, topological, electronic, thermodynamic, geometric descriptors, etc. Descriptors with same entries for most of the training compounds were removed from the pool of variables considered. Pairs of variables with correlation coefficients greater than 0.90 were considered as inter-correlated, and one of them in each correlated pair was deleted. Finally, the resulting data matrix was utilized for further analysis.

Robust PCA

Although principal component analysis (PCA) is a very popular dimension reduction technique, the results are highly affected by anomalous observations in the data. To avoid the sensitivity towards outliers, various robust PCA algorithms [12,13] have recently been developed. The algorithm of ROBPCA utilized in this study combines projection pursuit techniques with robust covariance estimation in lower dimensions and could be concluded as three stages: first, the data matrix is processed by reducing its data space to the affine subspace spanned by the number of observations; then a measure of outlyingness is computed for each data point, which is obtained by projecting the high-dimensional data points on many univariate directions; the last stage of ROBPCA consists of selecting the number of principal components (k) to retain and projecting the data points onto the k-dimensional subspace spanned by the k largest eigenvectors and of computing their center and shape by means of the reweighted MCD estimator. The eigenvectors of this scatter matrix then determine the robust principal components, and the MCD estimation estimate serves as a robust center. For visualization, we also represent the result of the PCA analysis by means of a diagnostic plot based on orthogonal distance and score distances. The orthogonal distance measures the distance between an observation and its projection in the k-dimensional PCA subspace, while the score distance is measured within the PCA subspace. Thus robust PCA might serve as a valuable tool for outlier detection [13,14], and any observations with large orthogonal distance or score distance would be identified as potential outliers. More information about ROBPCA algorithm could be obtained in reference [15]. ROBPCA is carried out by using the Matlab Toolbox [16] for Robust Calibration.

FS-MLR

For simplicity and interpretability, multiple linear regression (MLR) [17] was employed as the modeling method and a multiple-term linear equation is built step-by-step using forward selection (FS) strategy. To illustrate the process concisely, two descriptor pools need to be defined first. Pool1 indicates the descriptors which have been selected into the MLR model and pool2 deposits the remaining descriptors. In each step, the performance of each descriptor in pool2 in combination with those in pool1 is evaluated and the best one would be transferred from pool2 to pool1. The search process is terminated when stepping is no longer possible or when a specified maximum number of steps has been reached. FS-MLR model is achieved using JMP (Version 5.1, SAS.) with parameters of 'Prob to Enter' and 'Prob to leave' set default as 0.250 and 0.100, respectively.

ERM

Enhanced Replacement Method (ERM) is a modified version of replacement method (RM) proposed by Andrew G. Mercader et al [18,19] for variable selection in linear models. For RM, it approaches the minimum of standard error of regression (S) by judiciously taking into account the relative error of the coefficients of the least-squares model given by a set of d descriptors. ERM follows the same RM philosophy but exhibits less propensity for remaining in local minima and at the same time is less dependent on the initial solution. More information of this algorithm could be obtained in reference [18]. ERM is run by using Matlab (Version 7.2b, The Mathworks, Inc.).

Support vector regression (SVR)

Support vector machine (SVM), developed by Vapnik and Cortes [20], as a novel type of machine learning method, is gaining increasing popularity due to many attractive features and promising empirical performances. Besides the basic aim of SVM for data classification, an extension of this algorithm named support vector regression (SVR) has been developed to address regression problems. Briefly, a regression task usually involves training and test data which consist of some data instances. Each instance in the training set contains one "target value" (property value) and several "attributes" (features). The goal of SVM is to produce a model which predicts target value of data instances in the test set which are given only the attributes. In this study, SVR was performed with RBF as the kernel function. The parameters C and γ were set default, with C=1 and γ=1/k, where k means the number of attributes in the input data. All calculations in this work were carried out by using Matlab (Version 7.2b, The Mathworks, Inc.) and the SVM toolbox was developed by Chih et al. [21] The calculations were performed on a 1.80GHz Intel Pentium Dual E2160 with 2G RAM under windows XP.

Cross-validation

Cross-validation techniques [22,23] including leave-one-out cross-validation (LOOCV) and 5-fold cross-validation (5-CV) were employed to evaluate the performance of both linear an nonlinear models. In LOOCV, only one sample is selected as the test set for each time, and the other samples are used as training set to predict the selected sample. The process is repeated until all the samples have been removed once. While for 5-CV, the whole dataset is classified into 5 subsets. Each time, samples in one of the subsets are selected as the test set, while remaining samples are used as training set to predict the test set. This procedure is repeated for 5 times until each subset has been removed once as the test set. However, because the 5-CV results vary for each run due to random partitioning of the data set, the whole process is repeated for 20 times to eliminate the effect of random sample partitioning in this study. The average result of the multiple cross-validation runs provides an unbiased assessment of the model performance in predicting unknown compounds. The models were evaluated by measuring the prediction R^2 (explained variance), *RMS* (root-mean-square error), and *RSE* (relative standard error), with the formulations shown as follows:

$$R^2 = 1 - \frac{\sum_{n=1}^{n}(y_{\exp} - y_{pred})^2}{\sum_{i=1}^{n}(y_{\exp} - y^{-})^2}$$

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(y_{\exp} - y_{pred})^2}{n}}$$

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{\exp} - y_{pred})^2}{\sum_{i=1}^{n} y^2{}_{pred}}}$$

In the above equations, y_{exp} and y_{pred} are experimental and predicted log(1/IC$_{50}$) values, respectively; n is the number of samples in the data set of interest. d is the number of variables.

External validation, shuffling external validations and training set selection

It needs to be emphasized that no matter how robust, significant and validated a QSAR model may be, it can not be expected to reliably predict the modeled activity for the entire universe of compounds. Therefore, the performance of the selected descriptors was further evaluated by external validation. However, it is also well known that a QSAR model's ability to predict the properties of unknown chemicals depends largely on the nature of the training set and a model's predictive accuracy and confidence for different unknown chemicals varies according to how well the training set represents the unknown chemicals. Thus 28 representative compounds were carefully selected as the training set using principal component analysis (PCA), taking sample distribution into consideration.

Moreover, considering the fact that the results of external validation are to some extent highly unstable due to the different selection of training sets, we also employed shuffling external validations (SEVs) to eliminate as most as possible the effect of different training sets and to evaluate model performance in a more objective way. Concisely, in each shuffling, 28 compounds are randomly selected as training set and the others as test set, ensuring that activity of compounds in the training set covering the range of 5.400 to 9.120. This procedure is repeated 20 times to eliminate the effects of random selection of training samples, with the averaged results used for model evaluation.

Y-randomization and predictive ability analysis of ERM model

Y-randomization analysis [24] is implemented for further ensuring the robustness of ERM model. The dependent variable (log(1/IC$_{50}$) values) is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to have low R^2, high *RMS* and *S*. If the opposite happens, then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

According to the Tropsha et al. [25], the predictive power of a QSAR model can be conveniently estimated by the following equations:

$$R^2_{cv,ext} > 0.5 \quad (1)$$

$$R'^2 > 0.6 \quad (2)$$

$$\frac{(R'^2 - R_0^2)}{R'^2} < 0.1 \text{ or } \frac{(R'^2 - R_0'^2)}{R'^2} < 0.1 \quad (3)$$

$$0.85 <= k <= 1.15 \text{ or } 0.85 <= k' <= 1.15 \quad (4)$$

Calculations relating to $R^2_{cv,ext}$, $R_o{}^2$ and the slope k and k' are based on regression of observed values against predicted values and vice versa. They were discussed in detail in reference [25,26].

Results and discussion

As a first step, robust principal component analysis (robust PCA) was performed on a complete set of 37 4-anilino-3-quinolinecarbonitriles to ensure whether potential outliers exist in this data set. The resulting plot of orthogonal distance versus score distance is illustrated in Figure 1, where large deviations from the cluster center for samples 36 and 37 indicate the potential outlying nature of these compounds. Restated, large orthogonal distance indicates the large

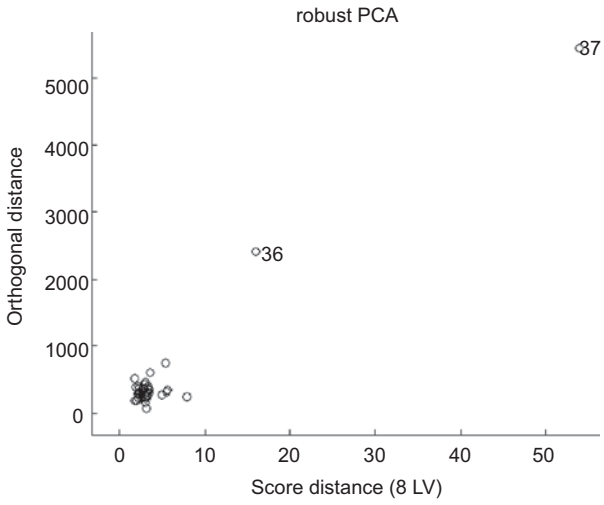


Figure 1. Orthogonal distance versus score distance for 37 samples using robust PCA.

deviation between these two compounds to their projection in the *k*-dimensional PCA subspace, while large score distance implies a large distance between the projections of them to that of the other samples in the *k*–dimensional PCA space. Thus samples 36 and 37 are removed from further analysis as potential outliers.

The biological activity values [IC_{50} (nM)] reported in the literature were converted to molar units [mol/l] and then further to –log scale and subsequently used as the response variable for the QSAR analysis. The $\log(1/IC_{50})$ values, along with the structure of all compounds including outliers are presented in Table 1. The compounds excluding outliers were divided into training and test sets containing 28 and 7 molecules respectively, with the detailed distribution shown in Figure 2. The training set has been used for QSAR model development and the test set was used to test the ability of developed QSAR model in predicting the Src kinase inhibiting activity.

Linear models (FS-MLR, ERM)

For simplicity and interpretability, multiple linear regression model was developed using both forward selection (FS) and enhanced replacement method (ERM) as variable selection strategy. For each model, a specific set of four descriptors were finally involved. The resulting regression model combined with forward selection was as follows:

$$\log(1/IC_{50}) = 30.235(\pm4.241) - 3.971(\pm0.428)*J3D - 12.870(\pm2.347)*PCR + 2.120(\pm0.357)*Mor25v + 193.540(\pm43.886)JGI6$$

$n = 28$, $R^2 = 0.886$, $RMS = 0.268$, $RSE = 0.033$ (training set)
$n = 7$, $R^2 = 0.892$, $RMS = 0.263$, $RSE = 0.033$ (test set)

ERM in combination with MLR results in a better regression model, with the equation shown as follows:

$$\log(1/IC_{50}) = 3.165(\pm0.767) - 0.051(\pm0.008) * RDF060e + 447.060(\pm47.477) * JGI6 - 1.931(\pm0.281)*Mor23v - 348.267(\pm85.804)*JGI9$$

$n = 28$, $R^2 = 0.918$, $RMS = 0.228$, $RSE = 0.028$ (training set)
$n = 7$, $R^2 = 0.928$, $RMS = 0.241$, $RSE = 0.030$ (test set)

In the models above, *n* is the number of compounds, R^2 is explained variance, *RMS* is root mean square error and *RSE* is relative standard error. The figures given in the parentheses with ± sign in the model are 95% confidence limits. It should be noted that the same training and test sets are utilized for FS-MLR and ERM, which ensures the comparability of both models. Since ERM outperforms FS as a variable selection strategy evidenced by significantly higher value of R^2, only the results obtained using ERM would be illustrated in detail. For visualization, a graphical representation of the experimental versus predicted $\log(1/IC_{50})$ values, as well as the residuals, is illustrated in Figures 3 and 4, respectively. The detailed information of selected descriptors is shown in Table 2. As a confirmation, the model mentioned above was also utilized to predict samples 36 and 37. The abnormal large residuals for both samples shown in Figure 4 confirmed their outlyingness to a large extent.

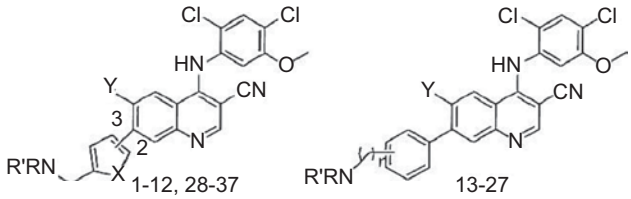
Nonlinear model (SVR)

Support vector regression (SVR) is an extension of support vector machine (SVM), with the aim of addressing regression problems. For SVR, forward variable selection method was also employed. By stepwise addition of the most important descriptors, the best SVR model was achieved when another four descriptors (RDF150e, Mor18u, C-025, C-034) were involved, with $R^2_{training} = 0.855$ and $R^2_{test} = 0.804$ for samples in training and test sets, respectively. Descriptors utilized in SVR are also illustrated clearly in Table 2.

Models validation and comparison

Cross-validation techniques including LOOCV and 5-CV were employed to evaluate the performance of these models. However, it needs to be emphasized that the real performance of any model could only be revealed using an external validation set. Therefore, the performance of these models was further evaluated by external validation. The detailed validation results are shown in Table 3, demonstrating that despite the much more sophisticated algorithm of SVR, ERM significantly outperforms SVR and FS-MLR with better performances. Considering that such inferiority of SVR could also be due to the selection bias of training samples, shuffling external validations (SEVs) was also implemented. The superiority of this method to the traditional external validation could be concluded as follows: first, compared to the random selection of training samples utilized in traditional external validation, this method takes sample distribution into consideration, ensuring samples in the training set covering the activity range of 5.400 to 9.120 in each shuffling; Moreover, average of 20 shuffles is chosen as the final external validation result, excluding to a large extent the bias resulted from different selection of training samples. The results of SEVs demonstrated in Table 3 confirmed the significantly better performance of ERM.

Table 1. Compound, experimental and calculated log(1/IC_{50}) values, as well as corresponding residuals based on ERM model.



ID	isomer	X	Y	n	NRR'	IC_{50}(nM)	log(1/IC_{50}) Obsd.	log(1/IC_{50}) Calcd.	residual
1[a]	3,5	S	H	1	morpholine	2.7	8.569	8.801	0.232
2	2,5	S	H	1	morpholine	2.5	8.602	8.455	−0.148
3	2,4	S	H	1	morpholine	5.7	8.244	8.420	0.176
4	3,2	S	H	1	morpholine	440	6.357	6.381	0.024
5[a]	3,4	S	H	1	morpholine	240	6.620	6.647	0.027
6[a]	3,5	S	H	1	*N*-Me-piperazine	3.8	8.420	8.549	0.129
7	3,5	S	H	1	*N*-OH-piperazine	1.4	8.854	8.314	−0.540
8	2,5	S	H	1	*N*-Me-piperazine	3.8	8.420	8.174	−0.246
9[a]	2,5	S	H	1	*N*-OH-piperazine	2	8.699	8.608	−0.091
10	2,5	S	H	1	piperidine	4.2	8.377	8.435	0.058
11	2,5	S	H	1	thiomorpholine	4.4	8.357	8.566	0.209
12[a]	C-6 isomer of 1					280	6.553	6.529	−0.024
13	1,3		OMe	1	morpholine	150	6.824	7.136	0.312
14	1,4		OMe	1	morpholine	15	7.824	7.697	−0.126
15	1,4		OMe	1	*N*-Et-piperazine	7	8.155	8.310	0.155
16	1,3		OMe	2	morpholine	74	7.131	7.259	0.128
17	1,4		OMe	2	morpholine	11	7.959	8.161	0.203
18	1,3		H	1	morpholine	28	7.553	7.348	−0.205
19	1,4		H	1	morpholine	3.3	8.481	8.154	−0.327
20	1,4		H	1	*N*-Et-piperazine	3	8.523	8.343	−0.180
21	1,3		H	2	morpholine	29	7.538	7.588	0.050
22	1,4		H	2	morpholine	14	7.854	7.773	−0.081
23	1,4		H	2	*N*-Et-piperazine	3.7	8.432	8.070	−0.362
24[a]	1,3		H	1	*N*-Me-piperazine	14	7.854	7.956	0.102
25	1,4		H	1	*N*-Me-piperazine	3.8	8.420	8.604	0.184
26	1,4		H	1	*N*-$(CH_2)_2$OH-piperazine	4.7	8.328	8.713	0.385
27	1,2		H	1	morpholine	4000	5.398	5.601	0.203
28	3,5	S	H	1	*N*-Me-piperazine	3.8	8.420	8.565	0.145
29	2,5	S	H	1	*N*-Me-piperazine	2.3	8.638	8.694	0.056
30	3,5	O	H	1	*N*-Me-piperazine	2.7	8.569	8.753	0.185
31	2,5	O	H	1	*N*-Me-piperazine	7.5	8.125	7.802	−0.323
32	3,5	O	OMe	1	*N*-Me-piperazine	0.78	9.108	9.021	−0.087
33	des 5-OMe of isomer of 30					11	7.959	8.249	0.290
34	C-6 isomer of 30					84	7.076	6.938	−0.138
35[a]	3,5	O	OMe	1	morpholine	1.5	8.824	8.257	−0.567
36[b]	3,5	O	OMe	1	NMe2	0.75	9.125	7.339	−1.786
37[b]	3,5	O	OMe	1	*N*-Ph-piperazine	3.6	8.444	9.545	1.101

'a' indicates samples in the test set, while 'b' indicates outliers.

Thus only descriptors selected in ERM model would be extensively discussed.

Y-randomization and predictive ability analysis of ERM model

Y-randomization analysis is implemented for further ensuring the robustness of ERM model, with detailed results shown in Table 4. The low R^2 and high *RMS* indicate that the good results in our models are not due to a chance correlation or structural dependency of the training set. Finally, the ERM model also passed the predictive ability analysis, with detailed results illustrated in Table 5.

Explanation of molecular descriptors

Considering the significantly better performance, only descriptors selected in ERM model would be extensively discussed in this study. The inter-correlation of these descriptors was evaluated and illustrated in Table 6, indicating

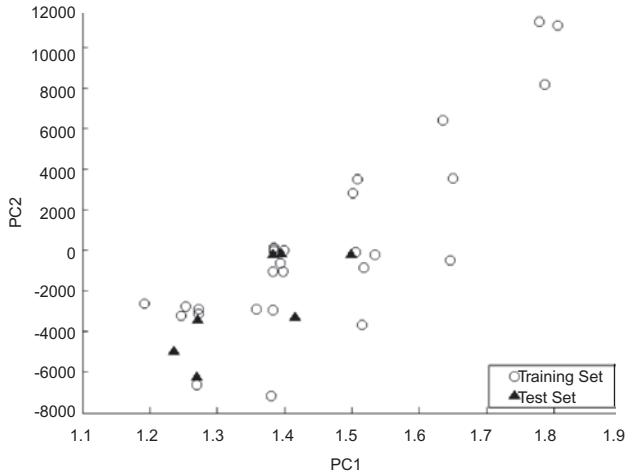


Figure 2. Score plot of PCA for training and test samples.

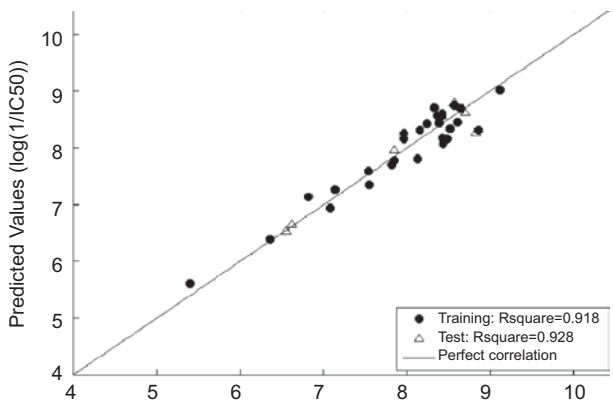


Figure 3. Plot of predicted versus experimental log(1/IC_{50}) values for ERM model.

no significant information overlapping among them. For evaluating the significance of each descriptor, we ranked the descriptors in ERM model according to their effect on increasing the value of *S* when removed from the model. In this case, the order found is:

JGI6 > Mor23v > RDF060e > JGI9

The most important descriptor JGI6 and the least one JGI9 belong to the family of topological charge index JGI. [27,28] Galvez Charge Indices GGIk and JGIk are defined as:

$$GGIk = \sum_{i=1,\, j=i+1}^{i=N-1,\, j=N} |CT_{ij}|\,\delta(k, D_{ij})$$

$$JCIk = \frac{GGIk}{(N-1)}$$

where N is the number of vertices (atoms different to hydrogen) in the molecular graph, and *k* the length of each path. $CT_{ij} = m_{ij} - m_{ji}$. '*m*' stands for the elements of the *M* matrix, $M = A \times D^*$, A is the adjacency (N×N) matrix of the molecular graph, D^* is the inverse square distance matrix, in which their diagonal entries are assigned as 0, and δ is Kronecker's delta. Thus, JGIk represents the average of the CT_{ij} terms, with $D_{ij}=k$, being D_{ij} the entries of the topological distance matrix (*D*). In the Charge Indices terms, the presence of heteroatoms is taken into account by introducing their

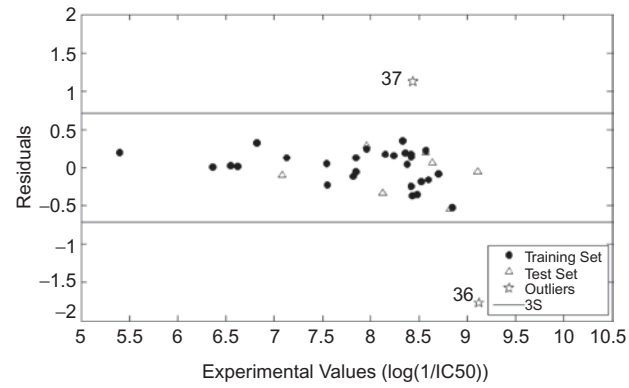


Figure 4. Plot of residuals against the experimental log(1/IC_{50}) values using ERM for samples in training and test sets, as well as outliers.

Table 2. Molecular descriptors selected in models.

Type	Molecular descriptor	Description
3D-MoRSE descriptors	Mor23v	3D-MoRSE - signal 23 / weighted by atomic van der Waals volumes
	Mor18u	3D-MoRSE - signal 18 / unweighted
	Mor25v	3D-MoRSE - signal 25 / weighted by atomic van der Waals volumes
RDF descriptors	RDF060e	Radial Distribution Function –6.0 / weighted by atomic Sanderson electronegativities
	RDF150e	Radial Distribution Function –15.0 / weighted by atomic Sanderson electronegativities
Topological charge indices	JGI6	mean topological charge index of order6
	JGI9	mean topological charge index of order9
Geometrical descriptors	J3D	3D-Balaban index
Walk and path counts	PCR	ratio of multiple path count over path count
Atom-centred fragments	C-025	R–CR–R
	C-034	R–CR–X

RDF, radial distribution function.

Table 3. Statistical results of performance validation.

				External validation		
		LOOCV	5-CV	training	test	SEVs
FS-MLR	R^2	0.861	0.854	0.886	0.892	0.832
	RMS	0.304	0.311	0.268	0.263	0.285
	RSE	0.038	0.039	0.033	0.033	0.036
ERM	R^2	0.890	0.880	0.918	0.928	0.872
	RMS	0.271	0.282	0.228	0.241	0.239
	RSE	0.034	0.035	0.028	0.030	0.030
SVR	R^2	0.698	0.638	0.855	0.804	0.607
	RMS	0.448	0.491	0.302	0.399	0.449
	RSE	0.056	0.061	0.038	0.050	0.056

Table 4. Y-randomization result of ERM model.

	LOOCV	5-CV	External validation training	External validation test
R^2	−0.297	−0.337	0.147	−0.291
RMS	0.929	0.942	0.646	1.284
RSE	0.117	0.117	0.081	0.161

Table 5. Predictive ability analysis result of ERM model.

	ERM	Target value
$R^2_{cv,ext}$	0.935	>0.5
R'^2	0.921	>0.6
$(R'^2-R_0^2)/R'^2$	−0.086	<0.1
$(R'^2-R_0'^2)/R'^2$	−0.086	<0.1
k	1.000	$0.85<=k<=1.15$
k'	0.999	$0.85<=k'<=1.15$

Table 6. Correlation matrix for descriptors selected in ERM model.

	RDF060e	JGI6	Mor23v	JGI9
RDF060e	1.000	−0.248	−0.077	0.210
JGI6		1.000	0.157	0.477
Mor23v			1.000	0.018
JGI9				1.000

electronegativity values in the corresponding entry of the main diagonal of the adjacency matrix. These indices represent a strictly topological quantity plausibly correlating with the charge distribution inside the molecule. In other words, the topological distance of substituents in C7 position plays an important role in determining the Src inhibitory activity. The positive coefficient of JGI6 indicates that the more the substituents with path lengths of 6, the higher the Src inhibitory activity might be, while the negative sign of JGI9 indicates an opposite effect, when path length is 9. This distribution is an important property, which conditions the behavior of many physiochemical and biological properties.

The 3D-MoRSE type of descriptor is obtained considering a molecular transform derived from an equation used in electron diffraction studies. [27,29] The electron diffraction does not directly yield atomic coordinates, but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations. These codes are defined in order to reflect the contribution at a prescribed scattering angle of an atomic property such as mass (m), polarizability (p), electronegativity (e) or volume (v) to the property under investigation, and so enabling to differentiate the nature of atoms. Mor23v, with the scattering angle of 23 $Å^{-1}$, is weighted by atomic volumes, and the negative coefficient might indicate the adverse molecular volume in improving the Src inhibitory activity.

RDF060e [27,30] belongs to the family of radial distribution function, which can act as a structure coding technique referred to as radial distribution function code (RDF code) to transform the 3D coordinates of the atoms of molecules into a structure code that has a fixed number of descriptors irrespective of the size of a molecule. Radial distribution function provides, besides information about interatomic distances in a whole molecule, the opportunity to gain access to other valuable information, for example, bond distance, ring types, planar and nonplanar systems and atoms types. RDF060e has a negative influence in the studied property, possibly decreasing the Src kinase inhibiting activity. This descriptor is weighted with atomic Sanderson electronegativities, and most significantly, this descriptor is corresponding to a sphere radius of 6.0 angstroms. Formally, the radial distribution function of an ensemble of n atoms can be interpreted as the probability distribution to find an atom in a spherical volume of radius R. In this sense, according to our ERM model, a spherical molecular volume with this dimension could have certain restrictions to the addition of substituents. This interpretation suggests that substituent in C6 position of 4-anilino-3-quinolinecarbonitriles might contribute negatively to the Src inhibitory activity when bulky substituents exist in C7 positions at the same time. This observation agrees with the explanation reported by Berger et al [9] .

Conclusions

Summarizing the above discussion, the present study gives rise to QSAR model with good statistical significance and predictive capacity for Src kinase inhibitory activity of 4-anilino-3-quinolinecarbonitriles. The result of this study suggests that the variables like RDF060e, Mor23v, JGI6 and JGI9 index play an important role in defining such inhibitory activity. The analysis, based on validation procedures, offers not only an accurate strategy for predicting Src inhibitory activity of novel 4-anilino-3-quinolinecarbonitriles, but also a useful guidance to synthesize novel analogs with potent activity against Src kinase.

Acknowledgements

***Declaration of interest*:** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Thomas SM, Brugge JS. Cellular functions regulated by Src family kinases. *Annu Rev Cell Dev Biol* (1997); 13: 513–609.
2. Cartwright CA, Meisler AI, Eckhart W. Activation of the pp60c-src protein kinase is an early event in colonic carcinogenesis. *Proc Natl Acad Sci USA* (1990); 87: 558–562.
3. Mao WG, Irby R, Coppola D, Fu L, Wloch M, Turner J, Yu H, Garcia R, Jove R, Yeatman TJ. Activation of c-Src by receptor tyrosine kinases in human colon cancer cells with high metastatic potential. *Oncogene* (1997); 15: 3083–3090.
4. Soriano P, Montgomery C, Geske R, Bradley A. Targeted disruption of the c-src proto-oncogene leads to osteopetrosis in mice. *Cell* (1991); 64: 693–702.
5. Paul R, Zhang ZG, Eliceiri BP, Jiang Q, Boccia AD, Zhang RL, Chopp M, Cheresh DA. Src deficiency or blockade of Src activity in mice provides cerebral protection following stroke. *Nat Med* (2001); 7: 222–227.

6. Kamens JS, Ratnofsky SE, Hirst GC. Lck inhibitors as a therapeutic approach to autoimmune disease and transplant rejection. *Curr Opin Invest Drugs* (2001); 2: 1213–1219.
7. Parang K, Sun GQ. Recent advances in the discovery of Src kinase inhibitors. *Expert Opin Ther Pat* (2005); 15:1183–1207.
8. Boschelli DH, Wang DY, Ye F, Yamashita A, Zhang N, Powell D, Weber J, Boschelli F. Inhibition of Src kinase activity by 4-anilino-7-thienyl-3-quinolinecarbonitriles. *Bioorg Med Chem Lett* (2003); 12: 2011–2014.
9. Berger D, Dutia M, Powell D, Wissner A, De Morin F, Raifeld Y, Weber J, Boschelli F. Substituted 4-anilino-7-phenyl-3-quinolinecarbonitriles as Src kinase inhibitors. *Bioorg Med Chem Lett* (2003); 12: 2989–2992.
10. Boschelli DH, Wu B, Ye F, Wang Y, Golas JM, Boschelli F. Synthesis and Src kinase inhibitory activity of a series of 4-[(2,4-dichloro-5-methoxyphenyl)amino]-7-furyl-3-quinolinecarbonitriles. *J Med Chem* (2006); 49: 7868–7876.
11. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV. Virtual computational chemistry laboratory - design and description. *J Comput Aid Mol Des* (2005); 19: 453–63.
12. Hubert M, Rousseeuw PJ, Verboven S. A fast method for robust principal components with applications to chemometrics. *Chemom Intell Lab Syst* (2002); 75: 101–111.
13. Hubert M, Engelen S. Robust PCA and classification in biosciences. *Bioinformatics* (2004); 20: 1728–1736.
14. Jackson DA, Chen Y. Robust principal component analysis and outlier detection with ecological data. *Environmetrics* (2004); 15: 129–139.
15. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal component analysis. *Technometrics* (2005); 47:64–79.
16. Verboven S, Hubert M. LIBRA: a MATLAB library for robust analysis. *Chemom Intell Lab Syst* (2005); 75: 127–136.
17. Sharma BK, Sharma SK, Singh P, Sharma S. Quantitative structure-activity relationship study of novel, potent, orally active, selective VEGFR-2 and PDGFR alpha tyrosine kinase inhibitors: Derivatives of N-Phenly-N '-{4-(4-quinolyloxy)phenyl)urea as antitumor agents. *J Enz Inhib Med Chem* (2008); 23: 168–173.
18. Mercader AG, Duchowicz PR, Fernández FM, Castro EA. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemom Intell Lab Syst* (2008); 92: 138–144.
19. Mercader AG, Duchowicz PR, Fernández FM, Castro EA, Bennardi DO, Autino J C, Romanelli GP. QSAR prediction of inhibition of aldose reductase for flavonoids. *Bioorg Med Chem* (2008); 16: 7470–7476.
20. Cortes C, Vapnik V. Support- vector networks. *Mach Learn* (1995); 20: 273–297.
21. Chang CC, Lin CJ. LIBSVM, a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
22. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* (1983); 78: 316–331.
23. Osten DW. Selection of optimal regression models via cross-validation. *J Chemom* (1988); 2: 39–48.
24. Wold S, Eriksson L. In: van de Waterbeemd H, editor. Chemometrics Methods in Molecular Design. Wiley-VCH: Weinheim. (1995); 309–318.
25. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* (2003); 22: 69–77.
26. Golbraikh A, Tropsha A. Beware of q2!. *J Mol Graph Model* (2002); 20: 269–276.
27. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley-VCH: Weinheim. (2000).
28. Fernandez M, Caballero J, Helguera AM, Castro EA, Gonzalez MN. Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. *Bioorg Med Chem* (2005); 13: 3269–3277.
29. Schuur J, Selzer P, Gasteiger J. The coding of three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J Chem Inf Model* (1996); 36: 334–344.
30. Gonzalez MP, Teran C, Teijeira M, Helguera AM. Radial distribution function descriptors: an alternative for predicting A2A adenosine receptors agonists. *Eur J Med Chem* (2006); 41: 56–62.