Mol Divers (2013) 17:75–83
DOI 10.1007/s11030-012-9411-0

FULL-LENGTH PAPER

Discriminating of ATP competitive Src kinase inhibitors and decoys using self-organizing map and support vector machine

Aixia Yan · Xiaoying Hu · Kai Wang · Jing Sun

Received: 10 July 2012 / Accepted: 15 October 2012 / Published online: 2 November 2012
© Springer Science+Business Media Dordrecht 2012

Abstract A data set containing 686 Src kinase inhibitors and 1,941 Src kinase non-binding decoys was collected and used to build two classification models to distinguish inhibitors from decoys. The data set was randomly split into a training set (458 inhibitors and 972 decoys) and a test set (228 inhibitors and 969 decoys). Each molecule was represented by five global molecular descriptors and 18 2D property autocorrelation descriptors calculated using the program ADRIANA.Code. Two machine learning methods, a Kohonen's self-organizing map (SOM) and a support vector machine (SVM), were utilized for the training and classification. For the test set, classification accuracy (ACC) of 99.92 % and Matthews correlation coefficient (MCC) of 0.98 were achieved for the SOM model; ACC of 99.33 % and MCC of 0.98 were obtained for the SVM model. Some molecular properties, such as molecular weight, number of atoms in a molecule, hydrogen bond properties, polarizabilities, electronegativities, and hydrophobicities, were found to be important for the inhibition of Src kinase.

Keywords Protein tyrosine kinase Src · ATP competitive Src inhibitors · Classification · Kohonen's self-organizing map (SOM) · Support vector machine (SVM)

Electronic supplementary material The online version of this article (doi:10.1007/s11030-012-9411-0) contains supplementary material, which is available to authorized users.

A. Yan (✉) · X. Hu · K. Wang · J. Sun
State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, Beijing University of Chemical Technology, P.O. Box 53, 15 BeiSanHuan East Road, Beijing, 100029, People's Republic of China
e-mail: aixia_yan@yahoo.com; yanax@mail.buct.edu.cn

Introduction

Protein tyrosine kinases (PTKs) catalyze the phosphorylation of a tyrosine residue on substrate proteins, which play crucial roles in many signal transduction pathways that regulate a variety of cellular functions, such as differentiation, proliferation, and apoptosis [1]. PTKs are classified into two types: receptor protein tyrosine kinase and non-receptor tyrosine kinase. Src-family kinase is a class of non-receptor PTKs and comprises several highly homologous proteins including Src, Yes, Fyn, Lyn, Hck, Blk, Fgr, and Yrk [2]. Here, we focused on Src kinase, a member of the Src kinase family. It has been demonstrated by the preclinical and clinical studies that Src signaling is involved in several processes relevant to prostate, breast cancer, osteoporosis, stroke, and myocardial infarction [3]. Thus, Src kinase has been identified as a potent drug target against cancer, osteoporosis, stroke, etc. As such, the inhibition of Src-involved signaling represents a reasonable strategy for prostate and breast cancer, and some other diseases.

Being a druggable target, Src kinase and its inhibitors have been of great interest in drug discovery. More than 2,000 small molecules have been identified as potent Src kinase inhibitors [4]. These identified Src kinase inhibitors can be roughly grouped into several categories: purines [5], oxindoles [6], quinazolines [7], pyrrolopyrimidines [8], pyridopyrimidines [9], quinolines [10] and others, as shown in Fig. 1. In the Src kinase inhibitor pool, three agents, dasatinib, saracatinib, and bosutinib, have reached clinical stage [11].

The available Src kinase inhibitors provide rich structural and activity information that can be used for further structure-activity relationship (SAR) analysis. In medicinal chemistry and chemoinformatics, computational methods, such as machine learning methods, have been widely and successfully utilized for SAR analysis. Some investigations,

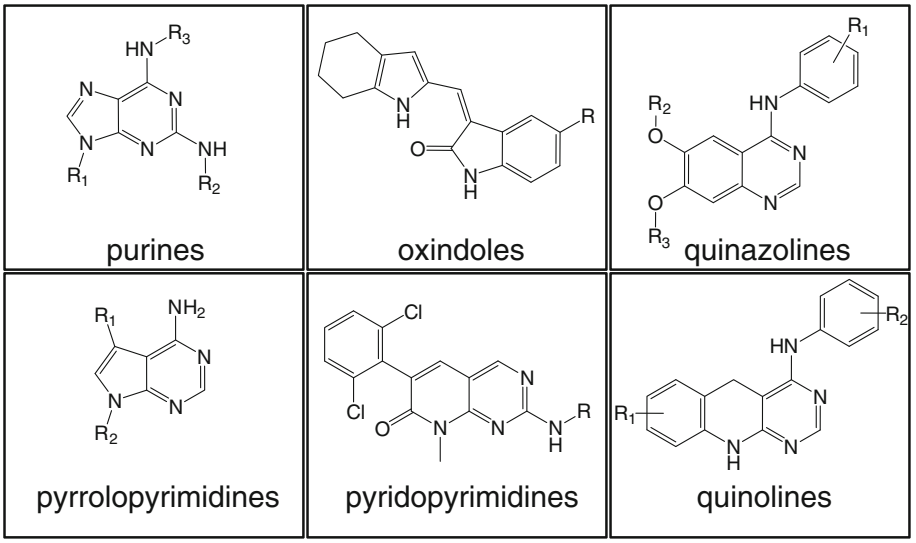

Fig. 1 Representative types of ATP competitive inhibitors against Src kinase

i.e., classification analysis or QSAR analysis about Src kinase inhibitors, have been done with attractive results. However, they only focused on one or two series of Src kinase inhibitors, which contain less than 100 compounds [12,13]. As there are far more available inhibitors against Src kinase, it would be interesting to carry out an SAR analysis using a larger data set.

Src kinase consists of three active domains, SH2 (Src homology 2), SH3 and protein–tyrosine kinase domain (ATP-binding site) [1]. ATP-binding site is the most popular binding site for the inhibition of kinases. It is also found that inhibitors binding to SH2 and SH3 domain are considered as potent anti-tumor agents [14]. Thus, there are three potent binding sites for Src kinase inhibitors: SH2 domain, SH3 domain and ATP-binding site. Here, we only focused on the inhibitors that bind to the ATP-binding site. Some representative ATP competitive inhibitors against Src kinase are shown in Fig. 2.

In this study, classification models were built using two computational methods, a Kohonen's self-organization map (SOM) and a support vector machine (SVM), to distinguish Src kinase ATP competitive inhibitors from non-binding decoys. Impressive results were obtained.

Datasets and methods

Datasets

More than 1000 Src kinase inhibitors were collected. In order to be binding site specific, only ATP competitive inhibitors with at least 50 μM potency on the basis of IC_{50} values were selected and used for further SAR analysis. Thus, 686 Src kinase inhibitors (see SrcInhibitorsDecoys.xls in the Electronic Supplementary Material), of which 671 inhibitors have the potency value of less than 10 μM, were collected from the literature [10,15–19] and BindingDB [4]. These inhibitors were considered active (class label: 1) in the calculations.

Src decoys, which have similar physical properties but dissimilar topology with Src kinase ligands, are unlikely active against Src kinase [20]. Thus, they were considered as inactive compounds against Src kinase (class label: -1). In this study, 1941 Src decoys (see SrcInhibitorsDecoys.xls in the Electronic Supplementary Material) were selected from the DUD database [20].

For the further classification analysis using SOM and SVM, all the Src kinase inhibitors and decoys were randomly divided into a training set containing 1430 compounds (458 inhibitors and 972 decoys) and a test set containing 1,197 compounds (228 inhibitors and 969 decoys) (as shown in Table 1). The training set was used to build classification models, and the test set was used for prediction.

Structure representation

Molecular physicochemical properties were used for structure representation for both the inhibitors and decoys. 131 molecular descriptors, which include 19 global molecular descriptors, 8 shape descriptors, and 104 2D autocorrelation vectors, were calculated using the program ADRIANA.Code [21,22].

A global molecular descriptor represents each chemical structure by a structural, chemical, or physicochemical feature or property of the molecule expressed by a single value. A size and shape descriptor represents a molecule by its 3D structure, and hydrogen atoms are taken into account.

The 2D autocorrelation vectors [23] were calculated based on the following eight atomic properties: atom identity, σ charge (SigChg), π charge (PiChg), total charges (TotChg), σ electronegativity (SigEN), π electronegativity (PiEN), lone-pair electronegativity (LpEN), and atomic polarizability (Apolariz). The 2D autocorrelation vectors for each of the above eight physicochemical atomic properties were calculated using the Eq. (1):

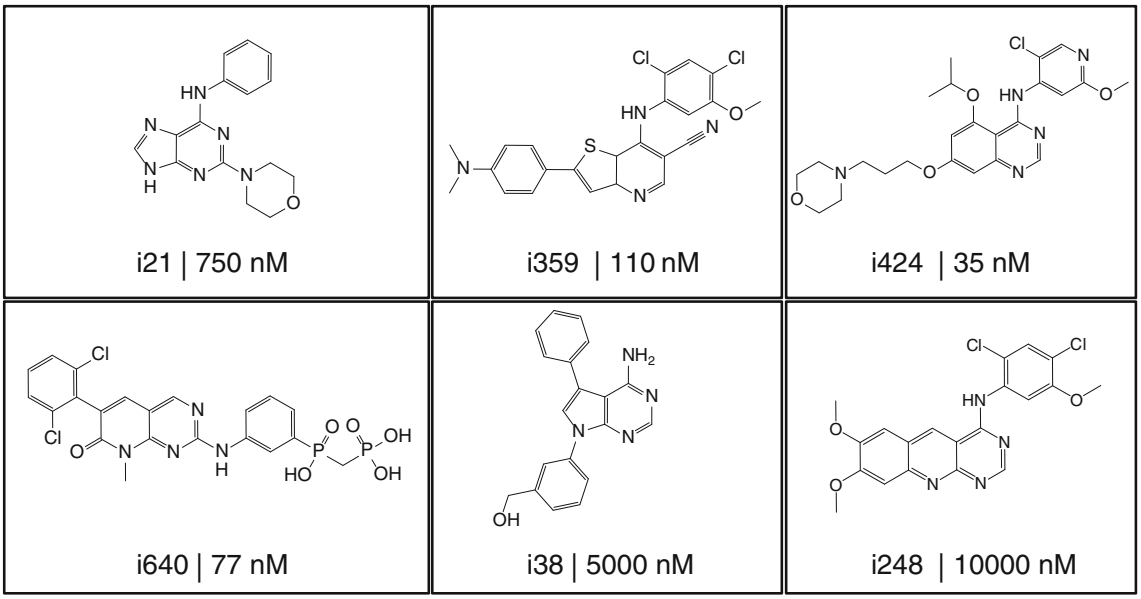

Fig. 2 Representative ATP competitive inhibitors of Src kinase. The compound ID and the IC_{50} values are also shown

Table 1 Data sets involved in this study

Data sets	# compounds	# Compounds (training set)	# Compounds (test set)
Inhibitors	686	458	228
Decoys	1, 941	972	969
Total	2, 627	1,430	1, 197

Number of compounds (# compounds) of the inhibitors, decoys, training set, and test set are shown

$$A(d) = \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}} p_i p_j \delta_{ij} (d - d_{ij}), \tag{1}$$

where $A(d)$ is the topological autocorrelation coefficient referring to atom pairs (i, j), which are separated by d bonds. $p_i(p_j)$ is an atomic property, e.g., the σ electronegativity on atom $i(j)$. Thus, for each molecule, a series of coefficients for different topological distances d, a so-called autocorrelation vector is obtained. In this work, 13 distances from $d = 0$ to $d = 12$ were considered and for each molecule, 104 2D autocorrelation vectors were calculated.

Molecular descriptor selection

Weka [24] was used for the selection of molecular descriptors. Weka is a collection of machine learning algorithms for data mining including a number of methods for data preprocessing, attribute selection, classification, etc. SVMAttributeEval [25] is a method contained in Weka for attribute selection. Attributes were evaluated by an SVM classifier and then they were ranked based on the square of the weight assigned by the SVM. The molecular descriptors with higher ranking score are more important for SVM training and prediction. To investigate the best combination of the highly ranked descriptors, different descriptor sets, each of which consists of the top N descriptors, are used for SVM training and prediction. The descriptor set that gives the best prediction performance would be the best combination of the descriptors.

Here, the SVMAttributeEval method implemented in Weka was used for molecular descriptor selection. By setting the N to a range of [5, 50], 46 combinations of the top N ranked descriptors were investigated using SVM method. The final combination containing 23 descriptors (shown in Table 2) was selected and used for classification analysis. The correlation coefficients between the descriptors and activity of the compounds are also shown in Table 2. Before training, the 23 descriptors were scaled to a [0.1,0.9] range via Eq. (2):

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \times 0.8 + 0.1 \tag{2}$$

where x_i is the original value, x_i^* is the scaled value, $x_{\min}$ and $x_{\max}$ are the corresponding minimum and maximum values of the descriptor variables, respectively.

Kohonen's self-organizing map

A Kohonen's self-organizing map (SOM) is a neural network model introduced by Kohonen for the construction of a nonlinear projection of objects from a high-dimensional space into a lower dimensional space [26]. The similarity perception of objects is an essential feature of a SOM. In a SOM, the neurons are arranged in a two-dimensional array to generate

Table 2 Selected 23 molecular descriptors and the correlation coefficient (CC) with the activity

Descriptor	Description	CC
Weight	Molecular weight	0.41
NAtoms	Number of atoms in the molecule	0.51
HDon_N	Number of hydrogen bonding donors derived from the sum of N-H groups in the molecule	−0.22
HAcc_N	Number of hydrogen bonding acceptors derived from the sum of nitrogen atoms in the molecule	0.43
NViolationsExtRo5	Number of violations of the extended Lipinski's rule of 5 (Weight > 500, XlogP > 5, HDon > 5, HAcc > 10, number of rotatable bonds > 10)	0.52
2DACorr_SigChg_1	2D autocorrelation weighted by σ atom charges, where $d = 0$	0.11
2DACorr_SigChg_2	2D autocorrelation weighted by σ atom charges, where $d = 1$	0.16
2DACorr_SigChg_5	2D autocorrelation weighted by σ atom charges, where $d = 4$	0.38
2DACorr_TotChg_3	2D autocorrelation weighted by total atom charges (sum of σ and π charges), where $d = 2$	0.12
2DACorr_SigEN_5	2D autocorrelation weighted by σ atom electronegativities, where $d = 4$.	0.38
2DACorr_SigEN_6	2D autocorrelation weighted by σ atom electronegativities, where $d = 5$	0.37
2DACorr_PiEN_4	2D autocorrelation weighted by π atom electronegativities, where $d = 3$	0.44
2DACorr_PiEN_8	2D autocorrelation weighted by π atom electronegativities, where $d = 7$	0.38
2DACorr_PiEN_11	2D autocorrelation weighted by π atom electronegativities, where $d = 10$.	−0.20
2DACorr_LpEN_3	2D autocorrelation weighted by lone-pair electronegativities, where $d = 2$.	−0.54
2DACorr_LpEN_10	2D autocorrelation weighted by lone-pair electronegativities, where $d = 9$.	0.09
2DACorr_Ident_9	2D autocorrelation weighted by atom idendities, where $d = 8$.	0.43
2DACorr_Polariz_1	2D autocorrelation weighted by effective atom polarizabilities, where $d = 0$.	0.61
2DACorr_Polariz_2	2D autocorrelation weighted by effective atom polarizabilities, where $d = 1$.	0.62
2DACorr_Polariz_3	2D autocorrelation weighted by effective atom polarizabilities, where $d = 2$	0.62
2DACorr_Polariz_4	2D autocorrelation weighted by effective atom polarizabilities, where $d = 3$	0.63
2DACorr_Polariz_7	2D autocorrelation weighted by effective atom polarizabilities, where $d = 6$	0.63
2DACorr_Polariz_9	2D autocorrelation weighted by effective atom polarizabilities, where $d = 8$	0.47

a two-dimensional Kohonen map such that similarity in the data is preserved. In other words, if two input data vectors are similar, they will be mapped into the same neuron or neurons near to each other in the Kohonen map. The SOM has been widely used in classification and similarity perception [27].

Represented by 23 descriptors, each molecule of the Src kinase inhibitors and decoys is an event in a 23-dimensional space, spanned by 23 descriptors. The 23-dimensional space is projected into a two-dimensional space using a SOM. The software used for the generation of the Kohonen map is SONNIA [28].

Support vector machine

A SVM [29] is a useful technique for data classification. A number of excellent introductions into SVM are available [25,30,31]. The SVM method originated as an implementation of Vapnik's Structural Risk Minimization (SRM) principle from statistical learning theory. The special property of an SVM is that it simultaneously minimizes the empirical classification error and maximizes the geometric margin. Thus, an SVM is also known as a maximum margin classifier.

In this study, the LIBSVM developed by Chang and Lin [32] was used for SVM analysis. The commonly used kernel, Radial Basis Function (RBF) kernel (Eq. 3), was used to implicitly convert the data into a higher-dimensional space. The parameter C (Eq. 4) and γ can be chosen by the auto-searching python script "grid.py" through a cross-validation method. Here, 23 descriptors representing the structures are used as input vectors.

$$k(x, y) = \exp(-\gamma \|x - y\|^2) \tag{3}$$

$$\min_{w,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i$$

$$\text{subject to } y_i(W^T \phi(X_i) + b) \geq 1 - \xi_i \tag{4}$$

$$\xi_i \geq 0$$

Results and discussion

Diversity of the Src kinase inhibitors

The 686 Src inhibitors used consist of different structural series, such as quinolinecarbonitriles, quinolines, benzotriazine, pyrazolopyrimidine, etc., as shown in Fig. 1. To

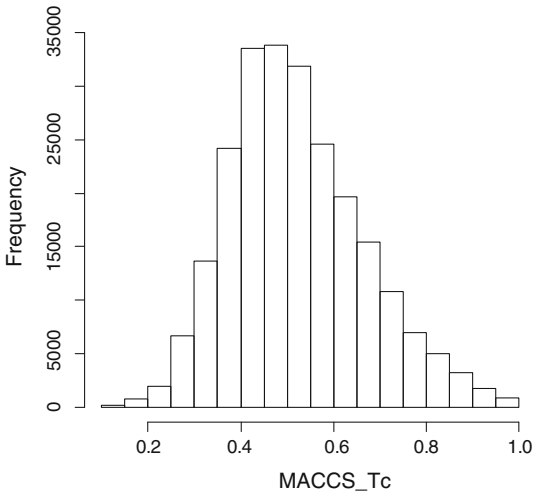

Fig. 3 Diversity of Src kinase inhibitors involved in this study. The frequency of the Tanimoto coefficient of compound pairs on the basis of MACCS fingerprint (MACCS_Tc) is shown

evaluate the diversity of the involved inhibitors, the pairwise Tanimoto coefficient (Tc) was calculated on the basis of MACCS structural keys [33]. The lower the Tc is, the more dissimilar the compounds are. Figure 3 reports the distribution of the molecular similarity. About 97 % (228,229 out of 234,956) of the compound pairs have the Tc value of less than 0.85, which is the normal threshold of molecular similarity. Thus, it is indicated that the data set is diverse.

Classification results using SOM

A SOM was applied for the classification analysis of Src kinase ATP competitive inhibitors and decoys. The initial learning rate was 0.7 and a rate factor was 0.95. The initial weights were randomly initialized, and the training was performed for a period of 1,600 epochs in an unsupervised manner. These parameters were utilized for the SOM learning of the training set and the combination of the training and test set.

The training set containing 458 Src kinase inhibitors and 972 decoys was projected into a planar 37 × 24 rectangular Kohonen map. The resulting map is shown in Fig. 4a. It can be seen from this Kohonen map that the Src kinase inhibitors (represented as unfilled triangles) mainly locate in three areas, and they are separated from the decoys (represented as unfilled circles) clearly. On the basis of most frequent occupation of a neuron, classification accuracy (ACC) of 99.16 % was obtained, as shown in Table 3.

Because of the imbalance of the data, i.e., the number of inactive compounds (decoys) is more than twice the number of active compounds (Src kinase inhibitors), the ACC is not sufficient for the evaluation of the performance of the classification model. Thus, the sensitivity (SE) of the Src

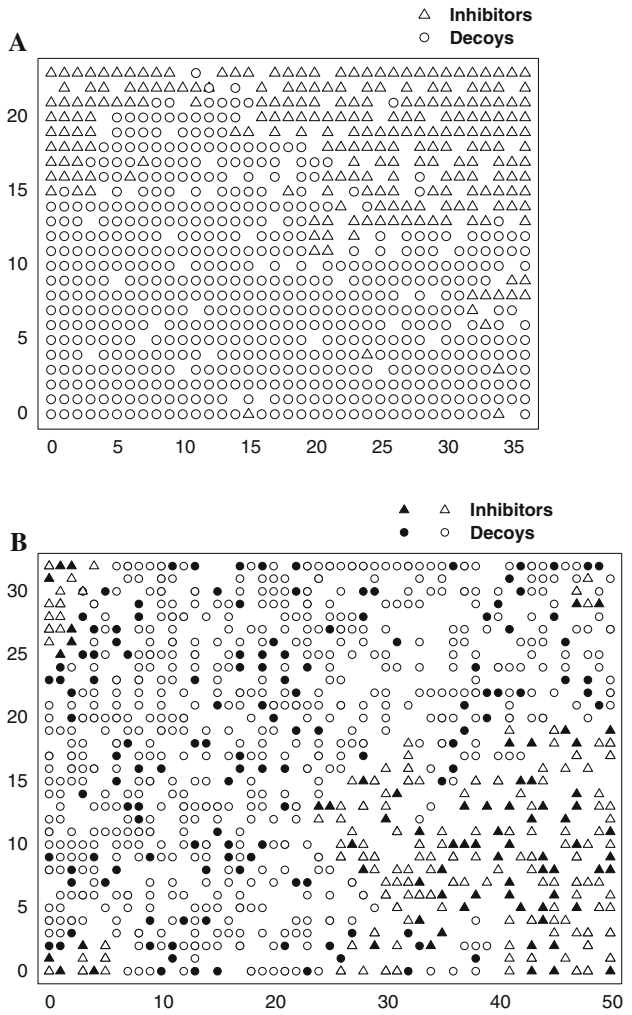

Fig. 4 Rectangular Kohonen maps for Src kinase inhibitors and decoys on the basis of most frequent occupation. **a** Location of the TR-inhibitors (inhibitors in the training set, *unfilled triangles*) and TR-decoys (*unfilled circles*) in the 37 × 24 rectangular Kohonen map. **b** Location of TR-inhibitors (*unfilled triangles*), TE-inhibitors (inhibitors in the test set, *black triangles*), TR-decoys (*unfilled circles*) and TE-decoys (*black circles*) in the 51 × 33 rectangular Kohonen map

Table 3 The performance of the SOM classification model

Datasets	ACC (%)	SE (%)	SP (%)	MCC
Training set	99.16	98.03	99.69	0.97
Test set	99.92	99.12	99.79	0.98

The classification accuracies (ACC), sensitivity of the inhibitors (SE), specificity of the decoys (SP), and the Matthews correlation coefficient (MCC) are shown

kinase inhibitors, specificity (SP) of the decoys, and the Matthews correlation coefficient (MCC) [34] of the classification model were calculated, via the following equations:

$$\mathrm{SE} = (\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})) \times 100\,\% \quad (5)$$

$$\mathrm{SP} = (\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})) \times 100\,\% \quad (6)$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}}. \quad (7)$$

Here, the number of true positives (Src kinase inhibitors that were correctly classified, TP), false positives (Src kinase decoys that were wrongly classified, FP), true negatives (Src kinase decoys that were correctly classified, TN), and false negatives (Src kinase inhibitors that were wrongly classified, FN) were of concern. The classification model performs better if its SE, SP, and MCC are higher.

For the training set, SE of 98.03 %, SP of 99.69 % and MCC of 0.97 were obtained on the basis of most frequent occupation of a neuron. The results are also shown in Table 3.

To further evaluate the classification ability of the SOM model, the training and test set were merged and were projected into a planar 51 × 33 rectangular Kohonen map. The resulting Kohonen map (shown in Fig. 4b) shows that the Src kinase inhibitors in the training set (TR-inhibitors, represented as unfilled triangles) are separated from TR-decoys (represented as unfilled circles) very well, which is consistent with the learning results using the training set alone. The Src kinase inhibitors in the test set (TE-inhibitors, represented as black triangles) are projected into the four corners of the rectangular Kohonen map, where TR-inhibitors locate. The TE-inhibitors are also clearly separated from the TE-decoys (represented as black circles). For the test set, ACC of 99.92 %, SE of 99.12 %, SP of 99.79 % and MCC of 0.98 were achieved on the basis of most frequent occupation of a neuron (see Table 3).

To evaluate the robustness of the SOM classification model, a fivefold cross validation (CV) for the training set was investigated. The Src inhibitors and decoys in the training set were randomly divided into five subsets with equal sizes, respectively. So that for each subset, the ratio of inhibitors to decoys is consistent. For each trial, one subset was considered as a test set, and the remaining four subsets were considered as a training set. For the fivefold CV, five trials were conducted and each compound was tested once. The average ACC for the fivefold CV of 97.06 % was obtained on the basis of most frequent occupation of the neurons, which indicates that the SOM model is stable.

Classification results using SVM

The training set was used to build an SVM classification model. SVM parameters C of 8 and γ of 8 were selected for learning after firstly auto selection using 'grid.py' and then manually optimization. The test set containing 228 Src kinase inhibitors and 969 decoys was used for the prediction. The accuracies of 100 % and 99.33 % for the training and test set were achieved, respectively (as shown in Table 4).

Table 4 The performance of the SVM classification model

Datasets	ACC (%)	5-CV (%)	10-CV (%)	LOO-CV (%)	SE (%)	SP (%)	MCC
Training set	100	98.74	98.81	98.81	100	100	1
Test set	99.33	–	–	–	99.12	99.38	0.98

The classification accuracies (ACC), fivefold Cross-validation (5-CV), tenfold CV (10-CV), Leave-One-Out CV (LOO-CV), sensitivity of the inhibitors (SE), specificity of the decoys (SP), and the Matthews correlation coefficient (MCC) are shown

The classification model was evaluated by the extensive CV. Fivefold, tenfold and Leave-One-Out (LOO) CV analysis were performed for the training set. In an n-fold CV, the dataset is randomly divided into n partitions of similar size. $n-1$ partitions are used for SVM learning, while the remaining partition is held for validation (testing). The prediction accuracies of n-fold and LOO CV are shown in Table 4. All the CV accuracies are above 98 %, which indicates that the classification model built in this study is stable.

For the training set, SE of 100 %, SP of 100 %, and MCC of 1 were obtained; for the test set, SE of 99.12 %, SP of 99.38 %, and MCC of 0.98 were achieved. The results are also shown in Table 4. Thus, it can be seen that the classification model built in this study is good.

The classification results of the SVM method are comparable to those of SOM's. Using the two methods, the Src kinase inhibitors were distinguished from decoys with high accuracy. This indicates that the selected 23 molecular descriptors are powerful.

Y-scrambling

In order to investigate the possibility of chance correlation of the classification model, Y-scrambling [35] was performed. For the training set, the activity class is randomly shuffled to change its true order. Thus, the meaningful connections between the descriptors and the activity class are destroyed. The shuffled training set was used to build an SVM model (scrambled model) with the same parameter settings as the classification model, and the original test set was used to test the scrambled model. The scrambling process was repeated 100 times. The performance of the scrambled models was then compared to that of the classification model (unscrambled model) built using the original data. If the unscrambled model was not a result of chance correlation, the MCC and prediction accuracy (ACC) of the unscrambled model should be much higher than those of scrambled models.

Here, the performance of the scrambled and unscrambled models is shown in Fig. 5. For the scrambled models, all the MCCs are smaller than 0.3, and all the ACCs are less than 0.8. For the unscrambled model, both the MCC and the ACC

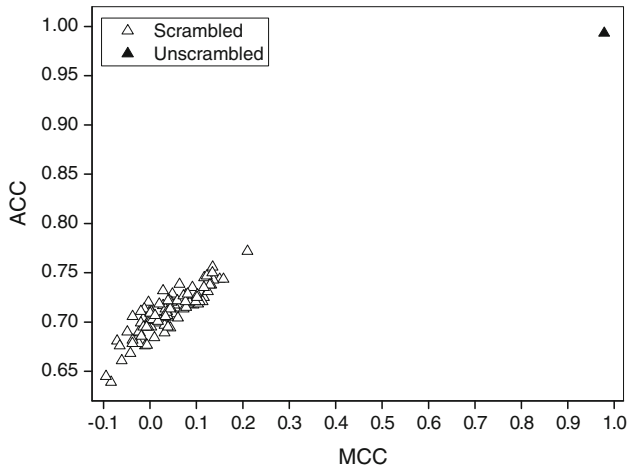

Fig. 5 Performances of scrambled and unscrambled models. For the scrambled models (*white triangles*), the Matthews correlation coefficients (MCCs) are less than 0.8 and the accuracies (ACC) are less than 0.3. For the unscrambled model (*black triangle*), the MCC and the ACC are close to 1

are close to 1. Thus, it is proved that the classification model built in this study was not a result of chance correlation.

Analysis of the selected descriptors

Based on the crystal structure of the Src kinase complex with ATP, the adenine moiety of ATP is anchored in the active site of Src kinase by two H-bond interactions, one with Tyr340 and the other with Met341 [36]. Therefore, H-bond interactions are essential for ATP competitive inhibition. ATP competitive Src kinase inhibitors were designed on the basis of the adenine moiety and the H-bonds network. Most Src competitive inhibitors contain an adenine-like substructure, as shown in Fig. 1, which can form hydrophobic interactions with residues Met341, Leu273, Val281, Ala293, and Leu393 [36]. Figure 6 shows representative interactions between Src kinase and its ATP competitive inhibitor, which were obtained by a docking simulation carried out with compound 42 [37] and the ATP-binding pocket of Src kinase (PDB: 2H8H) using MOE [33].

In this study, 23 (see Table 2) molecular descriptors were selected and used to build classification models. These descriptors can be roughly grouped into hydrogen bonding-related descriptors (HAcc_N and HDor_N), polarizabilities (2DACorr_Polariz_1, 2DACorr_Polariz_2, 2DACorr_Polariz_3, 2DACorr_Polariz_4, 2DACorr_Polariz_7, 2DACorr_Polariz_9), electronegativities (2DACorr_SigEN_5, 2DACorr_SigEN_6, 2DACorr_PiEN_4, 2DACorr_PiEN_8, 2DACorr_PiEN_11, 2DACorr_LpEN_3, 2DACorr_LpEN_10), atom charges (2DACorr_SigChg_1, 2DACorr_SigChg_2, 2DACorr_SigChg_5, 2DACorr_TotChg_3), atom identity (2DACorr_Ident_9), and structural descriptors (Weight, NAtoms, NViolationsExtRo5). Some of these descriptors represent critical interactions between Src kinase and its inhibitors.

As discussed above, H-bond interactions between Src kinase and its ATP competitive inhibitors are essential for the inhibition. For Src kinase inhibition, it has been found that a H-bond interaction between the nitrogen atom of the inhibitor and the NH group of Met341 at the ATP-binding site is one of the critical interactions for Src kinase inhibition [36]. In this study, two H-bond-related descriptors, HDon_N

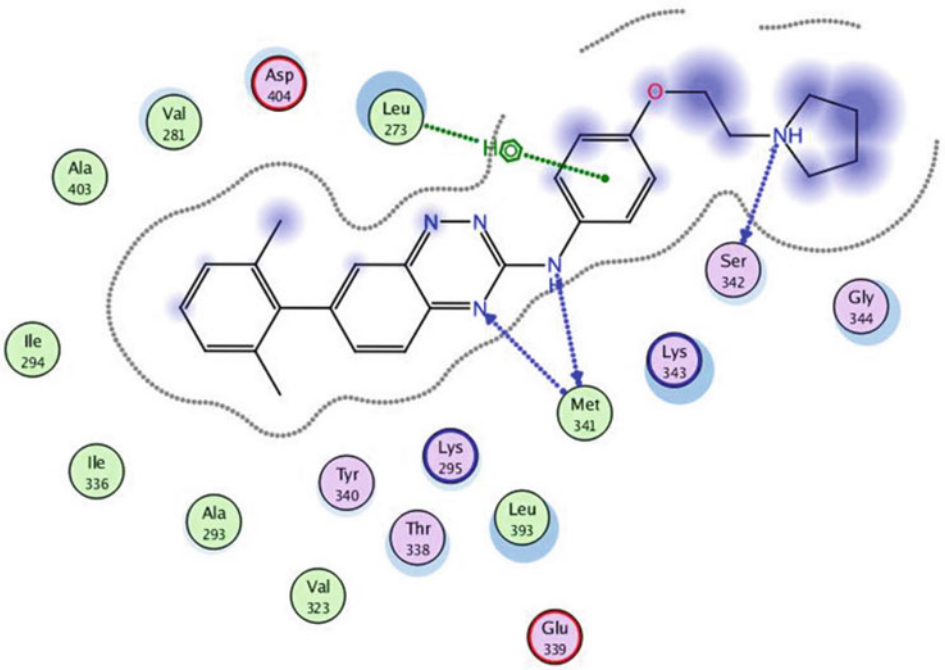

Fig. 6 Representative hydrogen bonds between a inhibitor and Src kinase (PDB:2H8H) generated using MOE. *Blue dash* represents hydrogen bonds between the nitrogen atoms of the inhibitor and residues Met341, Ser342 in the ATP-binding pocket of the Src kinase. *Green dash* represents the arene-H interaction. (Color figure online)

and HAcc_N, were selected and used for classification analysis. HDon_N represents the number of hydrogen bonding donors derived from NH group in the molecule; HAcc_N represents the number of hydrogen bonding acceptors derived from nitrogen atoms in the molecule. The two descriptors represent the H-bond interactions between nitrogen atom of inhibitors and ATP-site of Src kinase (see Fig. 6). This also indicates that the descriptor selection method used in this study works well.

Polarizability is another important feature for Src kinase inhibitors. Some residues, such as Lys295, Glu310, and Asp404, could form strong polar interactions with the inhibitors. All the selected polarizability-related descriptors are highly correlated with the class labels (i.e., among the selected descriptors in this study, polarizability-related descriptors are the most correlated with the class-labels.), as shown in Table 2. As atom polarizabilities are related to the formation of H-bonds, thus, a compound with higher polarizabilities is more likely to form H-bond with H-bond donors or acceptors in the ATP-binding pocket of Src kinase. This result is consistent with a previous study [38] that showed atom polarizabilities correlate to interactions between Src kinase and its inhibitors.

Atom electronegativity is also important for the inhibition of Src kinase [38]. Selected atom electronegativity-related descriptors including σ atom electronegativities, π atom electronegativities, and lone-pair electronegativities, are shown in Table 2. Atom electronegativity of a molecule correlate to polarizability of the molecule, thus it can affect the formation of H-bond between inhibitors and Src kinase.

On the basis of molecular structures, 23 physical chemical descriptors were selected and used for classification analysis. These descriptors represent H-bond interactions, hydrophobic interactions, and polar interactions between inhibitors and Src kinase. The good classification results using different methods indicate that the selected descriptors can distinguish Src kinase inhibitors from decoys.

Conclusions

Computational models were built for the identification of Src kinase ATP-competitive inhibitors, using a SOM and a SVM method, respectively. 23 global and topological molecular descriptors were used for the classification analysis. Comparable classification results were achieved for the two models, and the Src kinase inhibitors were distinguished from the decoys very well. The utilized molecular descriptors represent important molecular properties of Src kinase inhibitors, for example, hydrogen bonding-related descriptors representing the hydrogen bonding interactions between inhibitors and ATP-binding site of the Src kinase, atom charges, atom electronegativities, atom polarizabilities, and hydrophobic properties. Hence, the classification models built in this study can be used for further identification of new potent inhibitors against Src kinase.

Acknowledgments This work was supported by the National Natural Science Foundation of China (20605003 and 20975011). We thank Molecular Networks GmbH, Erlangen, Germany for making the programs ADRIANA.Code and SONNIA available for our scientific work.

References

1. Jr RR (2004) Src protein–tyrosine kinase structure and regulation. Biochem Biophys Res Comm 324: 1155–1164. doi:10.1016/j.bbrc.2004.09.171
2. Robinson DR, Wu Y, Lin S (2000) The protein tyrosine kinase family of the human genome. Oncogene 19: 5548–5557. doi:10.1038/sj.onc.1203957
3. Zhang N, Wua B, Boschelli DH, Golas JM, Boschelli F (2009) 4-Anilino-7-pyridyl-3-quinolinecarbonitriles as Src kinase inhibitors. Bioorg Med Chem Lett 19: 5071–5074. doi:10.1016/j.bmcl.2009.07.043
4. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucl Acids Res 35: D198–D201. doi:10.1093/nar/gkl999
5. Wang Y, Metcalf CA, Shakespeare W, Sundaramoorthi R, Keenan TP, Bohacek RS, Schravendijk MR, Violette SM, Narula SS, Dalgarno DC, Haraldson C, Keats J, Liou S, Mani U, Pradeepan S, Ram M, Adams S, Weigele M, Sawyer TK (2003) Bone-Targeted 2,6,9-Trisubstituted purines: novel inhibitors of Src tyrosine kinase for the treatment of bone diseases. Bioorg Med Chem Lett 13: 3067–3070. doi:10.1016/S0960-894X(03)00648-6
6. Blake RA, Broome MA, Liu X, Wu J, Gishizky M, Sun L, Courtneidge SA (2000) SU6656, a selective Src family kinase inhibitor, used to probe growth factor signaling. Mol Cell Biol 20: 9018–9027. doi:10.1128/MCB.20.23.9018-9027.2000
7. Ple PA, Green TP, Hennequin LF, Curwen J, Fennell M, Allen J, Vander Brempt CL, Costello G (2004) Discovery of a new class of anilinoquinazoline inhibitors with high affinity and specificity for the tyrosine kinase domain of c-Src. J Med Chem 47: 871–887. doi:10.1021/jm030317k
8. Missbach M, Jeschke M, Feyen J, Muller K, Glatt M, Green J, Susa M (1999) A novel inhibitor of the tyrosine kinase Src suppresses phosphorylation of its major cellular substrates and reduces bone resorption in vitro and in rodent models in vivo. Bone 24: 437–449. doi:10.1016/S8756-3282(99)00020-4
9. Kraker AJ, Hartl BG, Amar AM, Barvian MR, Showalter HDH, Moore CW (2000) Biochemical and cellular effects of c-Src kinase-selective pyrido[2,3-d]pyrimidine tyrosine kinase inhibitors. Biochem Pharmacol 60: 885–898. doi:10.1016/S0006-2952(00)00405-6
10. Boschelli DH, Powell D, Golas JM, Boschelli F (2003) Inhibition of Src kinase activity by 4-anilino-5,10-dihydropyrimido[4,5-*b*]-quinolines. Bioorg Med Chem Lett 13: 2977–2980. doi:10.1016/S0960-894X(03)00628-0
11. Saad F, Lipton A (2010) SRC kinase inhibition: targeting bone metastases and tumor growth in prostate and breast cancer. Cancer Treat Rev 36: 177–184. doi:10.1016/j.ctrv.2009.11.005
12. Zhu J, Lu W, Liu L, Gu T, Niu B (2009) Classification of Src kinase inhibitors based on support vector machine. QSAR Comb Sci 28: 719–727. doi:10.1002/qsar.200860105
13. Tintori C, Magnani M, Schenone S, Botta M (2009) Docking, 3D-QSAR studies and in silico ADME prediction on c-Src tyrosine

kinase inhibitors. Eur J Med Chem 44: 990–1000. doi:10.1016/j.ejmech.2008.07.002
14. Vidal M, Gigoux V, Garbay C (2001) SH2 and SH3 domains as targets for anti-proliferative agents. Crit Rev Oncol Hematol 40: 175–186. doi:10.1016/S1040-8428(01)00142-1
15. Altmann E, Missbach M, Green J, Susa M, Wagenknecht H, Widler L (2001) 7-Pyrrolidinyl- and 7-Piperidinyl-5-aryl-pyrrolo[2,3-d]-pyrimidines—potent inhibitors of the tyrosine kinase c-Src. Bioorg Med Chem Lett 11: 853–856. doi:10.1016/S0960-894X(01)00080-4
16. Boschelli DH, Wang YD, Johnson S, Wu B, Ye F, Sosa ACB, Golas JM, Boschelli F (2004) 7-Alkoxy-4-phenylamino-3-quinolinecarbonitriles as dual inhibitors of Src and Abl kinases. J Med Chem 47: 1599–1601. doi:10.1021/jm0499458
17. Guan H, Laird AD, Blake RA, Tanga C, Liang C (2004) Design and synthesis of aminopropyl tetrahydroindole-based indolin-2-ones as selective and potent inhibitors of Src and Yes tyrosine kinase. Bioorg Med Chem Lett 14: 187–190. doi:10.1016/j.bmcl.2003.09.069
18. Boschelli DH, Sosa ACB, Golas JM, Boschelli F (2007) Inhibition of Src kinase activity by 7-ethynyl-4-phenylamino-3-quinolinecarbonitriles: identification of SKS-927. Bioorg Med Chem Lett 17: 1358–1361. doi:10.1016/j.bmcl.2006.11.077
19. Boschelli DH, Wang D, Wang Y, Wu B, Honores EE, Sosa ACB, Chaudhary I, Golas J, Lucas J, Boschelli F (2010) Optimization of 7-alkene-3-quinolinecarbonitriles as Src kinase inhibitors. Bioorg Med Chem Lett 20: 2924–2927. doi:10.1016/j.bmcl.2010.03.025
20. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49: 6789–6801. doi:10.1021/jm0608356
21. ADRIANA.Code. Molecular Networks GmbH, Erlangen. http://www.molecular-networks.com/. Accessed April 2012
22. Gasteiger J (2006) Of molecules and humans. J Med Chem 49: 6429–6434. doi:10.1021/jm0608964
23. Wagener M, Sadowski J, Gasteiger J (1995) Autocorrelation of Molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. J Am Chem Soc 117: 7769–7775. doi:10.1021/ja00134a023
24. Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P, Witten IH (2010) WEKA-experiences with a Java open-source project. J Mach Learn Res 11: 2533–2541
25. Vapnik V, Chapelle O (2000) Bounds on error expectation for support vector machines. Neural Comput 12: 2013–2036. doi:10.1162/089976600300015042
26. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43: 59–69. doi:10.1007/BF00337288
27. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design, 2nd ed. Wiley: Weinheim
28. SONNIA, version 4.2. Molecular Networks GmbH, Erlangen. http://www.molecular-networks.com. Accessed April 2012
29. Boser BE, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) Fifth annual workshop on computational learning theory. ACM, New York, pp 144–152
30. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20: 273–297. doi:10.1007/BF00994018
31. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2: 121–167. doi:10.1023/A:1009715923555
32. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machine. http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed April 2012
33. MOE (The Molecular Operating Environment), version 2009.2010. Chemical Computing Group Inc., Montreal
34. Kryger G, Harel M, Giles K, Toker L, Velan B, Lazar A, Kronman C, Barak D, Ariel N, Shafferman A, Silman I, Sussman JL (2000) Structures of recombinant native and E202Q mutant human acetylcholinesterase c omplexed with the snake-venom toxin fasciculin-II. Acta Crystallogr D 56: 1385–1394. doi:10.1107/S0907444900010659
35. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451
36. Hennequin LF, Allen J, Breed J, Curwen J, Fennell M, Green TP, Brempt CL, Morgentin R, Norman RA, Olivier A, Otterbein L, Plé PA, Warin N, Costello G (2006) *N*-(5-Chloro-1,3-benzodioxol -4-yl)-7-[2- (4-methylpiperazin-1-yl) ethoxy] - 5-(tetrahydro-2H-pyran-4-yloxy)quinazolin-4-amine, a Novel, Highly Selective, Orally Available, Dual-Specific c-Src/Abl Kinase Inhibitor. J Med Chem 49: 6465–6488. doi:10.1021/jm060434q
37. Noronha G, Barrett K, Cao J, Dneprovskaia E, Fine R, Gong X, Gritzen C, Hood J, Kang X, Klebansky B, Li G, Liao W, Lohse D, Mak CC, McPherson A, Palanki MSS, Pathak VP, Renick J, Soll R, Splittgerber U, Wrasidlo W, Zeng B, Zhao N, Zhou Y (2006) Discovery and preliminary structure–activity relationship studies of novel benzotriazine based compounds as Src inhibitors. Bioorg Med Chem Lett 16: 5546–5550. doi:10.1016/j.bmcl.2006.08.035
38. Sun M, Zheng Y, Wei H, Chen J, Cai J, Ji M (2009) Enhanced replacement method-based quantitative structure–activity relationship modeling and support vector machine classification of 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors. QSAR Comb Sci 28: 312–324. doi:10.1002/qsar.200860107