Bioorganic & Medicinal Chemistry Letters 21 (2011) 2238–2243

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Prediction of biological activity of Aurora-A kinase inhibitors by multilinear regression analysis and support vector machine

Aixia Yan *, Yang Chong, Liyu Wang, Xiaoying Hu, Kai Wang

State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, PO Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, PR China

ARTICLE INFO

Article history:
Received 1 September 2010
Revised 19 February 2011
Accepted 28 February 2011
Available online 3 March 2011

Keywords:
Aurora-A kinase inhibitors
Kohonen's self-organizing map (SOM)
Multilinear regression
(MLR)
Support vector machine (SVM)
Quantitative structure–activity
relationships (QSAR)

ABSTRACT

Several QSAR (quantitative structure–activity relationships) models for predicting the inhibitory activity of 117 Aurora-A kinase inhibitors were developed. The whole dataset was split into a training set and a test set based on two different methods, (1) by a random selection; and (2) on the basis of a Kohonen's self-organizing map (SOM). Then the inhibitory activity of 117 Aurora-A kinase inhibitors was predicted using multilinear regression (MLR) analysis and support vector machine (SVM) methods, respectively. For the two MLR models and the two SVM models, for the test sets, the correlation coefficients of over 0.92 were achieved.

The Aurora kinases are a family of three highly homologous serine/threonine protein kinases, including Aurora-A, -B and -C. Since their discovery in 1995 and the first observation of expression in human cancer tissue in 1998, these kinases have been the subject of intense research in both the academic and industrial oncology communities.[1]

Among the three human Aurora kinases, Aurora-A has been the family member most consistently associated with cancer. The Aurora-A gene lays within a region of chromosome 20q13, which is amplified in many epithelial malignant tumors, including breast, gastric, colon, ovarian and pancreatic cancers.[2–4] The Aurora-B gene is located at chromosome 17p13.1, which has not been associated with significant amplification. Despite reports of overexpression in certain cancers, the increased expression may simply reflect hyperproliferation rather than carcinogenesis.[5] Aurora-C has similar functions as Aurora-B. The Aurora-C gene lies within a region of chromosome 19q13. It is highly expressed in the testis but is also present at a low level in other tissues.[6] Recently, Aurora-A has become an attractive target for the treatment of cancer. To date, more than 10 small molecules have entered clinical studies, such as VX-680, MLN8054, MLN8237, PHA-739358, AT-9238, SNS-314, and ENMD-2076.[1]

In previous studies, much attention has been given to the production of novel compounds for Aurora-A kinase inhibitors. However, it is not feasible to carry out the binding bioassays on every compound because of the constraint of time and cost. Alternatively, the binding affinity of a new molecule might be predicted using some quantitative structure–activity relationship (QSAR) models derived from the known inhibitors and their experimental bioassay values; and it can help to reduce the research time of new drugs. This methodology is very helpful in screening a large library of possible drug candidates for selectivity and potency.[7]

For kinase inhibitors, several QSAR models[8–10] have been built to predict the biological activity, which could predict the activities of newly designed compounds before a decision is made whether these compounds should be really synthesized and tested. Dong et al. built QSAR models for the prediction of the inhibitory activity of 148 Akt/protein kinase B (PKB) inhibitors using Support Vector Machine method.[8] 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; In addition, 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) was built by support vector regression (SVR) method.[8] Lather and colleagues have carried out a QSAR study on 44 GSK-3b (glycogen synthase kinase-3) inhibitors.[9] Both the 2D and 3D QSAR models were obtained. The predictive ability of both the models was determined using a randomly chosen test set containing eight molecules. The predictive correlation coefficients (R^2_{pred}) of 0.6 and 0.91 were obtained for the 2D and 3D models, respectively.[9] Recently, Jayashankar and Sundar have

* Corresponding author. Tel.: +86 10 64421335; fax: +86 10 64416428.
E-mail addresses: yanax@mail.buct.edu.cn, aixia_yan@yahoo.com (A. Yan).

0960-894X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2011.02.110

built 3D QSAR models for predicting the binding activity of 32 quinazoline derivatives to Aurora-B kinase.[10] Two models were built by molecular field analysis (MFA) and receptor surface analysis (RSA) methods, whose corresponding R^2 value of 0.954 and 0.949 were achieved, respectively.[10]

Besides simply as predictors, QSAR models can also be employed to bring new insights on the mechanism of activity.[11] The structural optimizations aided by QSAR models that lead to subsequently validated experimental findings are encouraged. The mechanistic analysis on predictive models can identify structural features that are deterministic to inhibitors' biological activities and/or subtype selectivities. Moreover, the QSAR-based virtual screening has been increasingly employed in recent years and brought forth great successes onto various biomedical targets.[11]

The aim of our work is to build QSAR models for the prediction of the inhibition of Aurora-A kinase inhibitors. The procedure of this work for building QSAR models includes four steps: (1) preparation a dataset containing 117 Aurora-A kinase inhibitors having experimental inhibitory activity of Aurora-A kinase; (2) calculation the 2D and 3D molecular descriptors using the ADRIANA.Code 2.2.2,[12,13] and selection the appropriate descriptors for activity prediction; (3) dividing the whole dataset into a training set and a test set by a random selection or by a Kohonen's self-organizing map (SOM);[14] (4) building models using the multilinear regression (MLR) method and the support vector machine (SVM).[15]

The 117 Aurora-A kinase inhibitors and their inhibitory activities were taken from nine literatures.[1,16–23] The bioassay (IC_{50}) values of the 117 inhibitors cover a broad range from 1 to 18,300 nM. The bioassay (IC_{50}) values were converted into the corresponding pIC_{50} ($-\log IC_{50}$) values, which is in the range of 4.74–9. A completed listed of the compounds structures, their corresponding experimental IC_{50} and pIC_{50} values are shown in Table S1 , Supplementary data.

The structure building and energy minimization were carried out using the software MOE (molecular operating environment).[24] The optimization of molecular 3D structure was generated by the CORINA software.[25] A total of 1143 descriptors were calculated using ADRIANA.Code 2.2.2,[12] including 19 global molecular descriptors,[26–28] eight size and shape descriptors,[29–31] 56 2D property autocorrelation descriptors,[32,33] 36 surface property autocorrelation descriptors[33,34] and 1024 radial distribution functions (RDF) codes.[35,36]

A global molecular descriptor represents a chemical structure by a structural, chemical or physicochemical feature or property of the molecule expressed by a single value. 19 global descriptors such as molecular weight, topological polar surface area (TPSA),[26] number of Lipinski's rule of five violations[27] and mean molecular polarizability,[28] were calculated.

A size or shape descriptor also represents a molecule by a single value and it is derived from the 3D structure of a molecule. Eight size and shape descriptors such as molecular diameter,[29] molecular span[30] and molecular radius of gyration,[30,31] were calculated.

The 2D property autocorrelation uses the molecular 2D structure and atom pair properties as a basis to obtain vectorial molecular descriptors.[32,33] The atom pair properties are summed up for certain topological distances which count the number of bonds on the shortest path between two atoms. The 2D molecular autocorrelation vectors[33] were calculated based on the following seven atomic properties: σ charge (SigChg),[37,38] π charge (PiChg),[39] total charges (TotChg), σ electronegativity (SigEN), π electronegativity (PiEN), lone-pair electronegativity (LpEN) and atomic polarizability (Apolariz).[40] For each molecule, all the hydrogen atoms were considered. For each property, the autocorrelation values for eight distances (0–7 bonds) were calculated. Thus, for each molecule, 56 2D property autocorrelations can be obtained.

The surface property autocorrelation[33,34] uses the molecular 3D structure properties as a basis to obtain vectorial molecular descriptors by spatial (3D) autocorrelation of properties of points on the molecular surface. The surface property autocorrelation vectors were calculated based on three properties: molecular electrostatic potential (SurfAcorr_ESP), hydrogen bonding potential (SurfAcorr_HBP) and hydrophobicity potential (SurfAcorr_HPP). For each of the three surface autocorrelation coefficients, a series of 12 vectors were computed, where Ln correspond to the 12 3D distance intervals from 1–2 Å, 2–3 Å, ... to 12–13 Å. Thus for each molecule, 36 autocorrelations of surface properties can be obtained.

Property-weighted radial distribution functions (property-weighted RDF) use the 3D structure of a molecule and atom pair properties as the basis to derive vectorial molecular descriptors.[35,36] For a certain atomic property, 128 RDF codes can be obtained. In addition, for each molecule, all the hydrogen atoms are considered during the computation. Here for each molecule, eight atomic properties (atom identity, charge (SigChg),[37,38] charge (PiChg),[39] total charges (TotChg), electronegativity (SigEN), electronegativity (PiEN), lone-pair electronegativity (LpEN) and atomic polarizability (Apolariz) were considered. Thus, for each molecule, 1024 RDF codes can be obtained.

Based on the compounds in dataset, molecular descriptors for building the QSAR models were selected. Molecular descriptors that were not significantly correlated with the activity (*r* <*0.1, r* is the correlation coefficient) were not used. If the pairwise correlation coefficient between any two descriptors was more than 0.85, the descriptor, which had the lower correlation to the activity, was removed. The remaining descriptors were chosen using stepwise linear regression variable selection method. Stepwise variable entry and removal examines the variables in the block at each step for entry or removal. According to the criteria, 13 descriptors were selected, which include RDF_Ident_39, RDF_Totchg_67, RDF_LpEN_38, RDF_LpEN_45, RDF_PiEN_26, RDF_PiEN_60, RDF_PiEN_81, RDF_PiEN_82, RDF_PiEN_84, RDF_PiChg_63, RDF_Polariz_12, 2DACorr_Polariz_2 and SurfACorr_ESP_5. The intercorrelations between the 13 descriptors and the activity are shown in Table 1. The selected descriptors were used in the following study.

The whole dataset was split into a training set and a test set based on two different methods, (1) by a random selection; and (2) on the basis of a Kohonen's self-organizing map (SOM).[14,41]

The dataset was randomly divided into a training set (87 compounds) and a test set (30 compounds) (as shown in Table S2 , Supplementary data). The training set was used to build two models: Model 1A by MLR and Model 1B by SVM, respectively. The test set was used to test the corresponding models.

The SONNIA software[41] was used for separating the dataset into a training set and a test set based on a Kohonen's self-organizing map (SOM). A two-dimensional Kohonen map with neurons was generated to classify the dataset. Data with similar input were mapped into the same neuron or neighbor neurons in the neural network. According to the Kohonen map, the dataset can be divided into a training set and a test set.

In this work, all the selected descriptors were scaled into a [0.1, 0.9] range using the formula (1), and were treated as input vectors in a Kohonen's self-organizing map (SOM) training.

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}} \times 0.8 + 0.1 \tag{1}$$

Where x_i was the original value, and x_i^* is the scaled value. x_min and x_max are the corresponding minimum and maximum values of the descriptor variable, respectively.

A planar 10 × 7 rectangular Kohonen map (shown in Fig. 1) was obtained with thirteen descriptors as input vectors. The initial

Table 1
The intercorrelations between the 13 descriptors and the activity for 117 compounds[a]

	pIC_{50}	RDF_TotChg_67	RDF_Ident_39	RDF_LpEN_38	RDF_LpEN_45	RDF_PiEN_26	RDF_PiEN_60	RDF_PiEN_81	RDF_PiEN_82	RDF_PiEN_84	RDF_PiChg63	RDF_Polariz_12	2DACorr_Polariz_2
RDF_TotChg_67	−0.32	1											
RDF_Ident_39	0.14	−0.34	1										
RDF_LpEN_38	−0.51	0.20	0.18	1									
RDF_LpEN_45	−0.17	0.32	−0.16	0.51	1								
RDF_PiEN_26	0.12	0.12	−0.20	0.05	0.60	1							
RDF_PiEN_60	0.43	−0.08	−0.20	−0.67	−0.31	0.10	1						
RDF_PiEN_81	−0.38	0.14	0.03	0.38	0.25	0.16	−0.13	1					
RDF_PiEN_82	−0.24	−0.02	0.08	0.49	0.26	0.08	−0.16	0.80	1				
RDF_PiEN_84	0.55	−0.38	0.22	−0.27	−0.48	−0.34	0.13	−0.23	−0.14	1			
RDF_PiChg_63	−0.27	0.09	−0.29	−0.09	−0.01	−0.02	0.10	−0.17	−0.20	−0.22	1		
RDF_Polariz_12	−0.55	−0.08	0.10	0.61	0.32	0.21	−0.35	0.38	0.48	−0.42	−0.03	1	
2DACorr_Polariz_2	0.56	−0.18	−0.02	−0.45	−0.54	−0.29	0.37	−0.11	−0.13	0.59	−0.12	−0.60	1
SurfACorr_ESP_5	0.10	0.33	−0.20	0.07	0.40	0.33	−0.20	−0.30	−0.44	−0.32	0.03	−0.10	−0.18

[a] RDF_Totchg_67: radial distribution functions weighted by the total atom charges (sum of σ and π charges), where r is in the range of 6.6–6.7 Å; RDF_Ident_39: radial distribution functions weighted by atom idendities, where r is in the range of 3.8–3.9 Å; RDF_LpEN_38, 45: radial distribution functions weighted by lone pair electronegativities, where r are in the range of 3.7–3.8 Å and 4.4–4.5 Å, respectively; RDF_PiEN_26, 60, 81, 82, 84: radial distribution functions weighted by π atom electronegativities, where r are in the range of 2.5–2.6 Å, 5.9–6.0 Å, 8.0–8.1 Å, 8.1–8.2 Å and 8.3–8.4 Å, respectively; RDF_PiChg_63: radial distribution functions weighted by π atom charge, where r is in the range of 6.2–6.3 Å; RDF_Polariz_12: radial distribution functions weighted by effective atom polarizabilities, where r is in the range of 1.1–1.2 Å; 2DACorr_Polariz_2: the second component of 2D autocorrelation coefficients for effective polarizability, where the distance *d* = 1; SurfACorr_ESP_5: the fifth components of autocorrelation coefficients for surface electrostatic potential, where *r* is in the range of 4–5 Å.

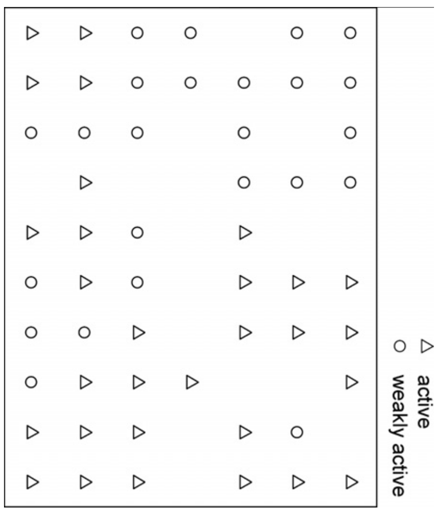

Figure 1. A rectangular Kohonen map for 117 compounds obtained by using the 13 selected input descriptors. 'active' means compounds with a pIC_{50} in the range of 6.87–9; 'weakly active' means compounds with a pIC_{50} in the range of 4.74–6.86.

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 1600 epochs in an unsupervised manner.

As shown in Figure 1, one can see that compounds with a different range of pIC_{50} are projected into different areas. In the Kohonen map, 54 of a total of 70 neurons are occupied. Afterwards, one of molecules in every neuron was taken for the training set (two molecules were taken for the training set if a neuron contains four compounds); the other molecules represented the test set. Thus, the 117 compounds were divided into a training set containing 77 compounds and a test set containing 40 compounds (as shown in Table S3 , Supplementary data) after the Kohonen's self-organizing map classification. The training set was used to build Model 2A by MLR and Model 2B by SVM, respectively; and the test set was used to validate the models.

A multilinear regression (MLR) analysis was performed using the selected 13 descriptors as input variables. The compounds of the training set were used to build a model, and the compounds of the test set were used for the prediction of pIC_{50}. In this work, two models were obtained using MLR method: Model 1A and Model 2A. Model 1A was achieved using the training set and test set divided by a random selection, and Model 2A obtained using training and test sets divided by a Kohonen's self-organizing map (SOM), the pIC_{50} was represented by the following equation:

$$pIC_{50} = \sum(C_iD_i) + D_c \qquad (2)$$

In the equation, D_c is a constant, D_i is a descriptor and C_i is its corresponding regression coefficient in MLR models. The corresponding regression coefficients are shown in Table 2.

For the training set of Model 1A, $r = 0.90$, $s = 0.42$, MAE = 0.36, $F = 24.72$ and $n = 87$ and for the test set, $r = 0.93$, $s = 0.40$, MAE = 0.35 and $n = 30$ (r is the correlation coefficient, s is the standard deviation, MAE is the mean absolute error, which equals to the mean value of the absolute errors and n is the number of compound in the training and test set); and for the training set of Model 2A, $r = 0.91$, $s = 0.42$, MAE = 0.35, $F = 22.65$ and $n = 77$ and for the test set $r = 0.92$, $s = 0.44$, MAE = 0.40 and $n = 40$. The results are shown in Table 3 and Figure 2. In addition, the prediction results of pIC_{50} are listed in Table S4 , Supplementary data.

The Libsvm program was used to build SVM models.[42] This software is based on the function of classification. After some improvement, it can also be applied to the regression problem well. More introductions and implementations about Libsvm can be found in their website.[42] The Libsvm regression was realized by the ε-support vector regression (ε-SVR) with a radial basis function (RBF) kernel function.

Table 2
Selected descriptors and the corresponding regression coefficients in Model 1A and Model 2A

Descriptor D_i	Coefficient C_i	
	Model 1A	Model 2A
RDF_TotChg_67	−2.217	−0.855
RDF_Ident_39	13.003	13.873
RDF_LpEN_38	−1.680	−2.205
RDF_LpEN_45	1.489	1.353
RDF_PiEN_26	6.748	8.361
RDF_PiEN_60	6.773	7.447
RDF_PiEN_81	−13.164	−13.920
RDF_PiEN_82	14.765	17.362
RDF_PiEN_84	9.170	9.977
RDF_PiChg_63	−1.687	−1.589
RDF_Polariz_12	−69.584	−56.283
2DACorr_Polariz_2	0.00041	0.00045
SurfACorr_ESP_5	0.072	0.084
D_c	4.132	3.239

According to the program guide, two necessary steps had to be taken in advance: the scaling of input data and searching for best parameters. The input data (the selected descriptors) was compressed into [0.1, 0.9] through the formula (1).

There are three parameters to adjust the efficiency of Libsvm program: C, γ and ε. An autosearching program (gridregression.py) named 'grid regression' was first adopted. It could search for best parameters C, γ and ε through a leave-k-out cross validation method.[42] Meanwhile, overfitting of training set could be prevented. Here a leave-20%-out cross validation was carried out. Afterward, manual searches were performed around the leave-20%-out cross validation results to select the best parameters.[43]

Model 1B and Model 2B were built using the support vector machine method with the Libsvm program.[42] The 13 selected ADRIANA.Code descriptors were used to build the models. Model 1B was built using the training and test sets split by a random selection; Model 2B was built using the training and test sets split by a Kohonen's self-organizing map (SOM).

For Model 1B, 87 compounds in training set were used to train an SVM model. The optimum parameters were: $C = 512$, $\gamma = 0.0009765625$, $\varepsilon = 1$. 30 compounds in the test set were used for prediction of pIC_{50}. For Model 1B, for the training set, $r = 0.90$, $s = 0.40$, MAE = 0.38, $n = 87$, and for the test set $r = 0.92$, $s = 0.41$, MAE = 0.40, and $n = 30$. The results are shown in Table 3 and Figure 3a.

For Model 2B, 77 compounds in training set were used to train a Support Vector Machine (SVM) model. The option parameters were set as: $C = 1024$, $\gamma = 0.0009765625$, $\varepsilon = 1$, and 40 compounds in the test set were used for prediction of pIC_{50}. For the training set, $r = 0.91$, $s = 0.40$, MAE = 0.37, and $n = 77$, for the test set, $r = 0.93$, $s = 0.39$, MAE = 0.40 and $n = 40$. The results are shown in Table 3 and Figure 3b. The predictions of inhibitory activity of 117 compounds are shown in Table S4 , Supplementary data.

Table 3
The prediction performances of four models: Multilinear Regression (MLR) Models (Model 1A and Model 2A) and Support Vector Machine (SVM) Models (Model 1B and Model 2B)

Model	Training set				Test set			
	n	r	sd	MAE	n	r	sd	MAE
Model 1A	87	0.90	0.42	0.36	30	0.93	0.40	0.35
Model 2A	77	0.91	0.42	0.35	40	0.92	0.44	0.40
Model 1B	87	0.90	0.40	0.38	30	0.92	0.41	0.40
Model 2B	77	0.91	0.40	0.37	40	0.93	0.39	0.40

n: number of compounds. r: correlation coefficient. sd: standard deviation. MAE: mean absolute error.

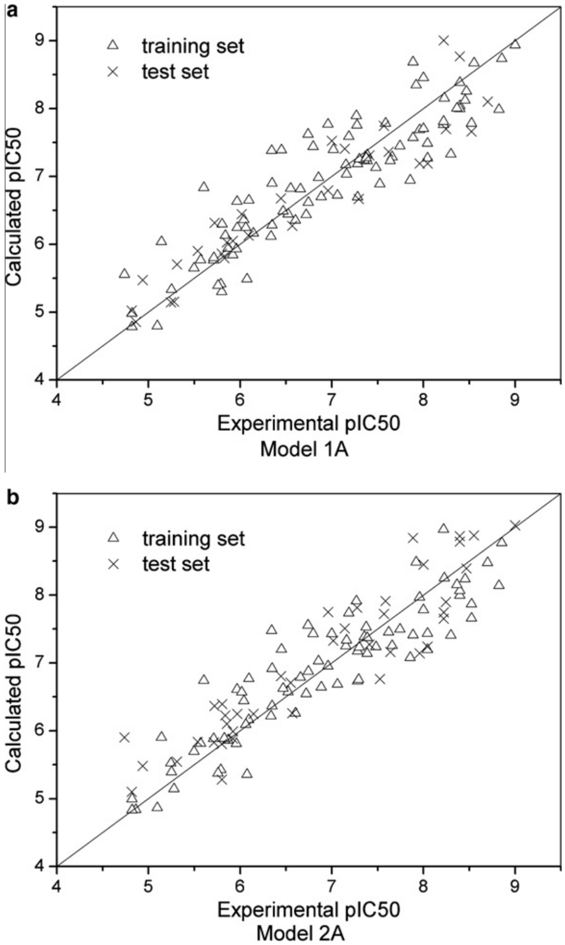

Figure 2. Calculated versus experimental pIC_{50} for 117 Aurora-A kinase inhibitors based on 13 selected descriptors by Multilinear Regression analysis. (a) Model 1A based on a training set and a test set using a random selection method; (b) Model 2A using a training set and a test set split based on a Kohonen's self-organizing map.

According to the MLR and SVM prediction figures (Figs. 2 and 3) and the prediction results shown in Table 3, it is obvious that all the models had a good prediction of pIC_{50}. Though the whole dataset was split into training set and test set based on different methods, the models obtained lead to a powerful predictability. For the training and test sets, the correlation coefficients of over 0.90 for MLR and SVM were achieved.

Y-Randomization is a technique that ensures the robustness of a QSAR model. We use this method to validate the MLR models we developed. The dependent variable vector (biological activity) is randomly shuffled and a new QSAR model is developed, using the given modeling algorithm. The procedure is repeated several times and the new QSAR models are expected to have low r^2 values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.[44]

The MLR models (Model 1A and Model 2A) were further validated by applying the Y-randomization test. In particular, 50 random shuffles of the Y-vector gave r^2 values in the ranges of 0.264–0.646 (the training set was selected by random) and 0.320–0.648 (the training set was selected by Kohonen map) as shown in Table S5, Figures S1 and S2 , Supplementary data. The low r^2 values obtained show that the good results in our original models are not due to a chance correlation or structural dependency of the training set.

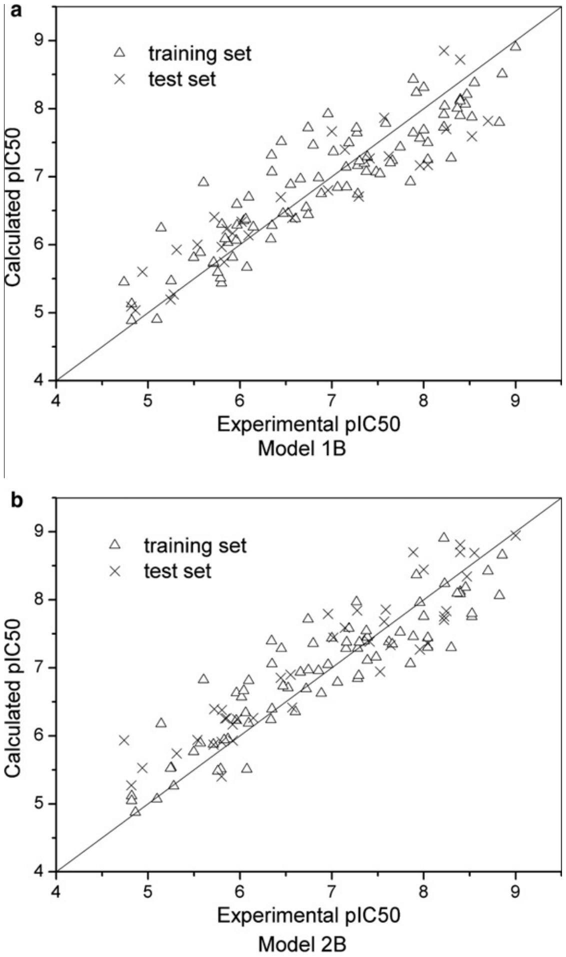

Figure 3. Calculated versus experimental pIC_{50} for 117 Aurora-A kinase inhibitors based on 13 selected descriptors by support vector machine. (a) Model 1B based on a training set and a test set using a random selection method; (b) Model 2B using a training set and a test set split based on a Kohonen's self-organizing map.

The domain of applicability of a QSAR model must be defined if the model is to be used for screening new compounds. Predictions for only those compounds that fall into this domain may be considered reliable.[45]

Extent of extrapolation is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i for each chemical, where the QSAR model is used to predict its activity:[46]

$$h_i = x_i(X^TX)^{-1}x_i^T \qquad (3)$$

In Eq. (3), x_i is the row vector containing the k model parameters of the query compound and X is the $n \times k$ matrix containing the k model parameters for each one of the n compounds in the training set. A leverage value greater than 3 k/n is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

However, no matter how robust, significant and validated a QSAR model may be, it cannot be expected to predict reliably the modeled activity for the entire universe of chemicals.[45] The domains of applicability of the models we developed were defined using the extent of extrapolation method. According to this method, we consider the predictions of the compounds as reliable, whose leverages lie within the domain of applicability. In Table S6 , Supplementary data, all leverages for the training and test sets are presented. The domains (warning leverage limits) for the models were 0.45 (the training set was selected by random) and 0.51 (the training set was selected by Kohonen map). As it can be concluded from the leverage values in Table S6, the predictions of the QSAR models for all the compounds (both the training and test sets) are considered reliable.

This research used broad molecular descriptors (1143 descriptors) based on both 2D and 3D molecular structures for predicting the activity of Aurora-A kinase inhibitors. The models built in this work do not need complicated molecules alignment based on their specific optimized 3D conformations. In our work, it is found that the RDF codes are important for predicting the inhibition of Aurora-A kinase inhibitors. Five RDF codes based on π atom electronegativities were selected (including RDF_PiEN_26, RDF_PiEN_60, RDF_PiEN_81, RDF_PiEN_82 and RDF_PiEN_84), which indicate that the representation of the inhibitors structures is much attribute to the π atom electronegativities. RDF_Polariz_12 and 2DACorr_Polariz_2 based on effective atom polarizabilities were both having the highest correlation coefficients with the inhibitory activity, which means that the molecular polarizability plays an important role in its binding to the Aurora-A kinase. In addition, RDF_Totchg_67 weighted by total charge (sum of σ and π charges), RDF_PiChg_63 weighted by π atom charge, RDF_LpEN_38 and RDF_LpEN_45 weighted by lone pair electronegativities were also selected, which means the molecular properties based on atom charge and lone pair electronegativities are also highly correlated with the inhibitory activity. For surface descriptors, it is generally accepted that receptor and substrate molecules recognize each other at their molecular surfaces.[33] Therefore, the binding strength of a receptor-drug complex depends on the shape of the substrate surface and on the distribution of certain properties on this surface. Here one surface descriptor SurfACorr_ESP_5 was selected, which suggests that the surface electrostatic potential of a compound is also important in its binding to the Aurora-A kinase.

Among the selected thirteen descriptors, eleven descriptors are the RDF codes, which indicate that the RDF codes are powerful descriptors for representing the 3D structure and characteristics of a molecule in detail. This kind of structure representation method has been found effective for the prediction of some molecular properties.[36,47] And the good models obtained in this study can be used for predicting the inhibitory activity of Aurora-A kinase inhibitors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (20605003 and 20975011). The authors thank Proffesor J. Gasteiger and the Molecular Networks GmbH, Erlangen, Germany for providing the programs ADRIANA.Code, CORINA and SONNIA for our scientific work.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2011.02.110.

References and notes

1. Pollard, J. R.; Mortimore, M. *J. Med. Chem.* **2009**, *52*, 2629.
2. Mountzios, G.; Terpos, E. *Cancer Treat. Rev.* **2008**, *34*, 175.
3. Bischoff, J. R.; Anderson, L.; Zhu, Y. F.; Mossie, K.; Ng, L.; Souza, B.; Schryver, B.; Flanagan, P.; Clairvoyant, F.; Ginther, C. *EMBO J.* **1998**, *17*, 3052.
4. Fancelli, D.; Moll, J.; Varasi, M.; Bravo, R.; Artico, R.; Berta, D.; Bindi, S.; Cameron, A.; Candiani, I.; Cappella, P.; Carpinelli, P.; Croci, W.; Forte, B.; Giorgini, M. L.; Klapwijk, J.; Marsiglio, A.; Pesenti, E.; Rocchetti, M.; Roletto, F.; Severino, D.; Soncini, C.; Storici, P.; Tonani, R.; Zugnoni, P.; Vianello, P. *J. Med. Chem.* **2006**, *49*, 7247.
5. Lok, W.; Klein, R. Q.; Saif, M. W. *Anti-Cancer Drugs* **2010**, *21*, 339.

6. Wang, S. D.; Midgley, C. A.; Scaerou, F.; Grabarek, J. B.; Griffiths, G.; Jackson, W.; Kontopidis, G.; McClue, S. J.; McInnes, C.; Meades, C.; Mezna, M.; Plater, A.; Stuart, I.; Thomas, M. P.; Wood, G.; Clarke, R. G.; Blake, D. G.; Zheleva, D. I.; Lane, D. P.; Jackson, R. C.; Glover, D. M.; Fischer, P. M. *J. Med. Chem.* **2010**, *53*, 4367.
7. Khadikar, P. V.; Sharma, V.; Karmarkar, S.; Supuran, C. T. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 923.
8. Dong, X. W.; Jiang, C. Y.; Hu, H. Y.; Yan, J. Y.; Chen, J.; Hu, Y. Z. *Eur. J. Med. Chem.* **2009**, *44*, 4090.
9. Lather, V.; Kristam, R.; Saini, J. S.; Kristam, R.; Karthikeyan, N. A.; Balaji, V. N. Q. S. A. R. *Comb. Sci.* **2008**, *27*, 718.
10. Jayashankar, L.; Sundar, B. Syama *J. Pharm. Sci. Res.* **2010**, *2*, 272.
11. Jorgensen, W. L. *J. Chem. Inf. Model.* **2006**, *46*, 937.
12. ADRIANA.Code, version 1.0, Molecular Networks GmbH, Erlangen, Germany. <http://www.molecular-networks.com> (accessed February 2011).
13. Gasteiger, J. *J. Med. Chem.* **2006**, *49*, 6429.
14. Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59.
15. Vapnik, V. N. *The Natural of Statistical Learning Theory*; Springer, 1995.
16. Coumar, M. S.; Leou, J. S.; Shukla, P.; Wu, J. S.; Dixit, A. K.; Lin, W. H.; Chang, C. Y.; Lien, T.; Tan, W. U. K.; Chen, C. H.; Hsu, J. T. A.; Chao, Y. S.; Wu, S. Y.; Hsieh, H. P. *J. Med. Chem.* **2009**, *52*, 1050.
17. Howard, S.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Curry, J.; Devine, L. A.; Early, T. R.; Fazal, L.; Gill, A. L.; Heathcote, M.; Maman, S.; Matthews, J. E.; McMenamin, R. L.; Navarro, E. F.; O'Brien, M. A.; O'Reilly, M.; Rees, D. C.; Reule, M.; Tisi, D.; Williams, G.; Vinkovic, M.; Wyatt, P. G. *J. Med. Chem.* **2009**, *52*, 379.
18. Oslob, J. D.; Romanowski, M. J.; Allen, D. A.; Baskaran, S.; Bui, M.; Elling, R. A.; Flanagan, W. M.; Fung, A. D.; Hanan, E. J.; Harris, S.; Heumann, S. A.; Hoch, U.; Jacobs, J. W.; Lam, J.; Lawrence, C. E.; McDowell, R. S.; Nannini, M. A.; Shen, W.; Silverman, J. A.; Sopko, M. M.; Tangonan, B. T.; Teague, J.; Yoburn, J. C.; Yu, C. H.; Zhong, M.; Zimmerman, K. M.; O'Brien, T.; Lew, W. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4880.
19. Coumar, M. S.; Wu, J. S.; Leou, J. S.; Tan, U. K.; Chang, C. Y.; Chang, T. Y.; Lin, W. H.; Hsu, J. T.; Chao, Y. S.; Wua, S. Y.; Hsieha, H. P. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1623.
20. Talele, T. T.; McLaughlin, M. L. *J. Mol. Graphics Modell.* **2008**, *26*, 1213.
21. Sardon, T.; Cottin, T.; Xu, J.; Giannis, A.; Vernos, I. *Chem. Bio. Chem.* **2009**, *10*, 464.
22. Rawson, T. E.; Ruth, M.; Blackwood, E.; Burdick, D.; Corson, L.; Dotson, J.; Drummond, J.; Fields, C.; Georges, G. J.; Goller, B.; Halladay, J.; Hunsaker, T.; Kleinheinz, T.; Krell, H. W.; Liang, J. J.; Limberg, A.; McNutt, A.; Moffat, J.; Phillips, G.; Ran, Y. Q.; Safina, B.; Ultsch, M.; Walker, L.; Wiesmann, C.; Zhang, B.; Zhou, A.; Zhu, B. Y.; Ruger, P.; Cochran, A. G. *J. Med. Chem.* **2008**, *51*, 4465.
23. Martin, I. A.; Burdick, D.; Corson, L.; Dotson, J.; Drummond, J.; Fields, C.; Huang, O. W.; Hunsaker, T.; Kleinheinz, T.; Krueger, E.; Liang, J.; Moffat, J.; Phillips, G.; Pulk, R.; Rawson, T. E.; Ultsch, M.; Walker, L.; Wiesmann, C.; Zhang, B.; Zhu, B. Y.; Cochran, A. G. *J. Med. Chem.* **2009**, *52*, 3300.
24. MOE version 2008.10 <http://www.chemcomp.com>.
25. Sadowski, J.; Gasteiger, J. *Chem. Rev.* **1993**, *93*, 2567. CORINA can be obtained from Molecular Networks GmbH, Erlangen, Germany. <http://www.molecular-networks.com> (accessed February 2011).
26. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
27. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3.
28. Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533.
29. Petitjean, M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331.
30. Volkenstein, M. V. *Configurational Statistics of Polymeric Chains*; Wiley-Interscience: New York, 1963.
31. Tanford, C. *Physical Chemistry of Macromolecules*; Wiley: New York, 1961.
32. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.
33. Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Am. Chem. Soc.* **1995**, *117*, 7769.
34. Teckentrup, A.; Briem, H.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 626.
35. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, *19*, 151.
36. Yan, A. X.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429.
37. Gasteiger, J.; Marsili, M. *Tetrahedron Lett.* **1978**, *34*, 3181.
38. Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
39. Kleinoeder, T. Ph.D. thesis, University of Erlangen-Nuernberg, 2005.
40. Gasteiger, J.; Hutchings, M. G. *J. Am. Chem. Soc.* **1984**, *106*, 6489.
41. SONNIA can be obtained from Molecular Networks GmbH, Erlangen, Germany. <http://www.molecular-networks.com> (accessed Feb 2011).
42. Chang, C. C.; Lin, C. J. LIBSVM: A Library for Support Vector Machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (accessed Feb 2011).
43. Yan, A. X.; Wang, Z.; Cai, Z. Y. *Int. J. Mol. Sci.* **2008**, *9*, 1961.
44. Melagraki, G.; Afantitis, A.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *Bioorg. Med. Chem.* **2007**, *15*, 7237.
45. Tropsha, A.; Gramatica, P.; Gombar, V. K. Q. S. A. R. *Comb. Sci.* **2003**, *22*, 69.
46. Atkinson, A. *Plots, Transformations and Regression*; Clarendon Press: Oxford, UK, 1985.
47. Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J. *Neuroscience* **2010**, *1*, 288.