Monatsh Chem (2011) 142:949–959
DOI 10.1007/s00706-011-0493-7

ORIGINAL PAPER

QSAR study on the interactions between antibiotic compounds and DNA by a hybrid genetic-based support vector machine

Xi Bin Zhou · Wen Jing Han · Jing Chen ·
Xiao Quan Lu

Received: 8 October 2010 / Accepted: 30 March 2011 / Published online: 19 May 2011

Abstract Studies on the interactions of antibiotic compounds with DNA can provide useful suggestions and guidance for the design of new and more efficient DNA-binding drugs. A quantitative structure–activity relationship (QSAR) study of the binding modes and binding affinities of the interactions between 30 antibiotic compounds and DNA was performed. A large number of descriptors that encode hydrophobic, topological, geometrical, and electronic properties were calculated to represent the structures of the antibiotic compounds. Aiming at a system with small, multidimensional samples, we utilized the genetic algorithm-support vector machine (GA-SVM) method to develop the QSAR, which can select an optimized feature subset and optimize SVM parameters simultaneously. A binary QSAR model for predicting binding mode and conventional QSAR models for predicting binding affinity were built based on the GA-SVM approach. The descriptors selected using GA-SVM represented the overall descriptor space and can account well for the binding nature of the considered dataset. The descriptors selected using the GA-SVM method were then used for developing conventional QSAR models by the artificial neural network (ANN) approach. A comparison between the conventional QSAR models using GA-SVM with those using ANN revealed that the former were much better. GA-SVM models can be useful for predicting binding modes and binding activities of the interactions of new antibiotic compounds with DNA.

X. B. Zhou · W. J. Han · J. Chen · X. Q. Lu (✉)
Key Laboratory of Bioelectrochemistry
and Environmental Analysis of Gansu Province,
College of Chemistry and Chemical Engineering,
Northwest Normal University, Lanzhou 730070, China
e-mail: luxq@nwnu.edu.cn

Keywords Antibiotic compound · DNA · Genetic algorithm-support vector machine · Binary QSAR · Regression

Introduction

Molecular recognition of proteins and nucleic acids by low-molecular-weight compounds is an area of fundamental interest [1]. Within this general area, drug–DNA interactions are of particular importance, not only because they provide a molecular basis for antitumor, antiviral, and antibiotic drugs to elucidate their structure–activity relationships and thereby improve understanding of their bioactivity mechanisms, but also because they can guide rational design of more efficient DNA-binding drugs. Various small molecules, which exert their activities through site-specific noncovalent binding to DNA, are attractive as versatile platforms for the development of DNA ligands that may lead to more efficient DNA-binding drugs through structural modification of compounds.

It was found that there are two principal modes by which compounds can bind noncovalently to DNA: intercalation and groove binding [2, 3]. Intercalation was first proposed by Lerman [4]. Intercalators bind to DNA by insertion of a planar, aromatic substituent between base pairs, simultaneously lengthening and unwinding the helix. Intercalators vary in the extent to which they unwind DNA, but all lengthen DNA to about the same extent [5]. Groove binding was first proposed by Wartell et al. [6]. In groove binding, the crescent-shaped ligand fits into the minor groove with little steric hindrance and little perturbation of the DNA structure.

In order to improve clinical efficacy of the DNA binding-drugs and also to design new drugs it is necessary to generate considerable interests in the interactions between antibiotic compounds and DNA, there have been extensive studies using various physical and chemical techniques. These methods include UV–Vis spectroscopy [7–11], cyclic voltammetry [12, 13], electrospray ionization mass spectrometry [14], nuclear magnetic resonance [15, 16], circular dichroism [17–19], fluorescence spectroscopy [20–22], and piezoelectric quartz crystal impedance (PQCI) [23]. In addition, a variety of computational and simulation methods have been utilized to investigate the interactions of antibiotic compounds with DNA. Lu et al. [24] predicted the binding affinity of the interactions of antibiotic compounds with DNA using multiple linear regression (MLR) and artificial neutral network (ANN) based on 12 physicochemical descriptors. Chen et al. [25] simulated the docking interactions of antibiotic compounds with proteins and constructed a prediction model for the binding mode using the best prediction set support vector machine (BPSSVM) method based on 12 physicochemical descriptors. In this paper, we built binary and conventional QSAR models using the genetic algorithm-support vector machine (GA-SVM) method to study the interactions between antibiotic compounds and DNA based on 1,777 calculated molecular descriptors, as shown in Fig. 1. We used the GA-SVM method for descriptor space reduction and optimization of SVM parameters. The binary QSAR model predicted the binding mode of the interactions between antibiotic compounds and DNA. The conventional QSAR models predicted the binding affinity of the interactions between antibiotic compounds and DNA. The descriptors selected using the GA-SVM method were then used for developing conventional QSAR models by using the artificial neural network (ANN) approach. The performance of the ANN models was compared with that of the GA-SVM models.

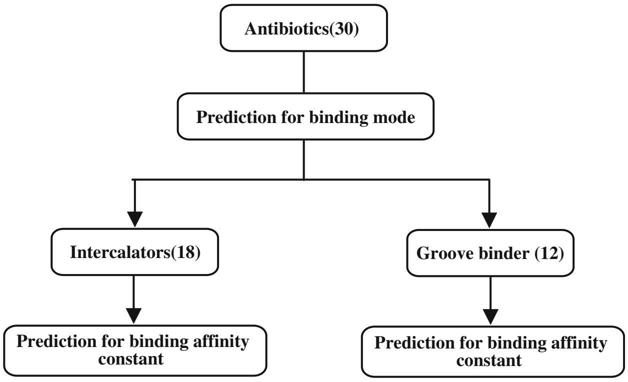

Antibiotics(30)
Prediction for binding mode
Intercalators(18)
Groove binder (12)
Prediction for binding affinity constant
Prediction for binding affinity constant

Fig. 1 Study process for interaction of antibiotic compounds and DNA

Results and discussion

Results of the binary QSAR model

In this study, we have, for the first time, applied the GA-SVM approach to develop an efficient model for prediction of binding mode. In the binary QSAR model, binding mode was expressed in binary format (1, intercalation; 0, groove binding) [26]. The dataset was randomly divided into a training set of 22 compounds and an external test set of 8 compounds (4 intercalators and 4 groove binders). The training set was used to generate the binary QSAR model, and the external test set was employed to evaluate the predictive ability of the binary QSAR model. The prediction accuracy for the training set was 90.91% by fivefold cross-validation. For the external test set, six compounds were predicted correctly. The prediction accuracy for the external test set was 75%. The predicted results and prediction errors for the binary QSAR model are listed in Table 1. The binary QSAR model clearly demonstrated its ability to forecast the binding mode with high accuracy.

The predictive ability of the binary QSAR model built by GA-SVM is affected by two SVM parameters: the kernel function parameter σ and the penalty parameter C. A good performance model depends on the two SVM parameters. The kernel function parameter σ greatly affects the number of support vectors, which has a close relation with the performance of the SVM and its training time. Too many support vectors could result in overfitting and prolong the training time. In addition, σ controls the amplitude of the Gaussian radial basis kernel function (RBF) and, therefore, controls the generalization ability of the SVM. The penalty parameter C controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, insufficient stress would be placed on fitting the training data. On the other hand, if C is too large, the SVM model would overfit the training dataset. The optimization of the SVM parameters was performed by systematically changing their values in the training step. Along with the implementation of the GA-SVM process, the number of features selected gradually decreased while accuracy was improved. When the number of features reached a minimum and accuracy reached a maximum, or the fitness value did not improve during the last 200 generations, the most suitable values of the SVM parameters were obtained. The best choices for C and σ were 1.2609 and 0.5622.

Five features were selected from the 1,777 to build the binary QSAR model. The descriptors selected by the GA-SVM were I_{B}, 6thMS-SAS, Mor01i, MB-ATS_5 (W), and $\lambda^{\mathrm{q}}_{\mathrm{H5}}$. These descriptors encode two-dimensional (2D) or three-dimensional (3D) structural information weighted by atomic physicochemical properties. These atomic

Table 1 Experimental binding mode versus predicted binding mode and corresponding errors in the training set and test set for the binary QSAR model

Number	Antibiotic	Binding mode	Predicted	Error
1	DAM	0	0	0
2	2,7-DAM	0	0	0
3	Adriamycin	0	0	0
4	WP776	0	0	0
5	WP756	0	0	0
6	WP758	0	0	0
7	Mitoxantrone	0	1	−1
8	MDPTQ	0	1	−1
9	NMHE	0	0	0
10	AMAC	0	0	0
11	MMQ1*	0	1	−1
12	Proflavine*	0	0	0
13	Propidium*	0	0	0
14	ADM	0	0	0
15	DADM	0	0	0
16	MHE	0	0	0
17	APTQ	0	0	0
18	MMQ2*	0	1	−1
19	Distamycin	1	1	0
20	Netropsin	1	1	0
21	13SAB89	1	1	0
22	Berenil	1	1	0
23	DB244*	1	1	0
24	DB351*	1	1	0
25	DB75	1	1	0
26	DB818*	1	1	0
27	DB226*	1	1	0
28	DB293	1	1	0
29	DAPI	1	1	0
30	DB921	1	1	0

* Molecule included in the test set of the binary QSAR model

physicochemical properties include molecular volume, solvent accessible surface, atomic charge, and *E*-state indices. I_{B} calculates the principal moments of inertia about the principal axes of a molecule. The principal moments of inertia are the physical quantity related to the rotational dynamics of a molecule. The 6thMS-SAS is a molecular surface derived descriptor. Mor01i is the 3D-MoRSE signal 1.0 weighted by the *E*-state descriptor, combining the 3D arrangement of the atoms of a molecule and atomic electrotopological state information to characterize the molecule. The 3D-MoRSE descriptors are calculated by summing atomic weights viewed by a different angular scattering function. The values of these functions are calculated at 32 evenly distributed values of scattering angle(s) in the range 1–32 $\mathrm{\AA}^{-1}$ from the 3D atomic coordinates of a molecule. The entire 32 3D-MoRSE values span a 32-dimensional space where each structure corresponds to a point in this space. $\lambda^{\mathrm{q}}_{\mathrm{H5}}$ (the Burden-CAS-University of Texas eigenvalues (BCUT) 5th highest of atomic charges) belongs to the BCUT descriptors which are extensions of Burden descriptors [27, 28]. The Burden descriptors are based on a combination of the atomic number for each atom and a description of the nominal bond type for adjacent and nonadjacent atoms. The BCUT descriptors expand the number and types of atomic features that can be considered, and also provide a greater variety of proximity measures and weighting schemes. $\lambda^{\mathrm{q}}_{\mathrm{H5}}$ incorporates atomic charge and connectivity information. The BCUT descriptors can capture sufficient structural features of molecules to yield useful measurement of molecular diversity.

Results of conventional QSAR model using GA-SVM for predicting binding affinity of intercalation

Conventional QSAR models were constructed for the 13 antibiotic compounds based on the measured binding affinities ($\log_{10}k'$) using the GA-SVM regression method. Fivefold cross-validation was used to build the model. The model was then validated using an external test set (five randomly chosen antibiotic compounds) to examine the

Table 2 Experimental $\log_{10}k'$ versus predicted $\log_{10}k'$ and corresponding residuals in the training set and test set for the intercalation regression model

Number	Exp ($\log_{10}k'$)	GA-SVM		ANN	
		Pre ($\log_{10}k'$)	Residual	Pre ($\log_{10}k'$)	Residual
1	4.9085	4.9716	−0.0631	4.7287	0.1798
2	4.5955	4.8506	−0.2551	4.3448	0.2507
3	8.4	8.6703	−0.2703	8.1012	0.2988
4	6.6	6.6908	−0.0968	6.7044	−0.1044
5	5.9	5.8039	0.0961	6.0009	−0.1009
6	7.2	6.9758	0.2242	7.2744	−0.0744
7	9.9494	9.8243	0.1251	7.8396	2.1098
8	6.3711	6.4523	−0.0812	7.0311	−0.6600
9	7.1139	6.7840	0.3299	6.3398	0.7741
10	4.8921	5.0617	−0.1696	4.7749	0.1172
11	5.9345	6.2710	−0.3365	5.8757	0.0588
12	5.4314	5.4428	−0.0114	5.0789	0.3525
13	4.7160	4.7888	−0.0728	4.9084	−0.1924
14*	5.2529	5.3059	−0.0530	4.6627	0.5902
15*	4.9370	5.7996	−0.8626	5.6468	−0.7098
16*	6.3617	6.0942	0.2675	6.4851	−0.1234
17*	6.2787	5.8093	0.4694	6.3689	−0.0902
18*	6.041	5.5084	0.5326	6.7626	−0.7216

* Molecule included in the test set of the intercalation regression model

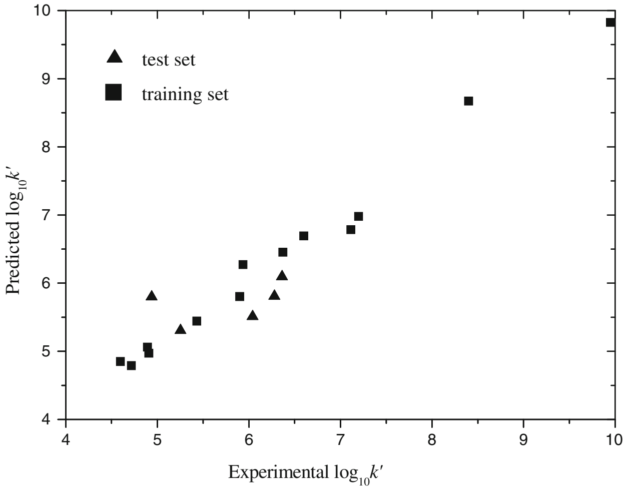


Fig. 2 Predicted versus experimental $\log_{10}k'$ for the training and test sets derived from the intercalation regression model built by GA-SVM

predictability and robustness of the QSAR model. The predicted binding affinity constant values and predicted errors for the training set and the external test set are listed in Table 2 and plotted against the experimental values in Fig. 2. It is seen that the predicted results are close to their experimental values with reasonable deviations. The cross-validated coefficient q^2, the corresponding correlation coefficient R, and the root-mean-square error (RMSE) for the training set were 0.9835, 0.9926, and 0.1938. The external prediction capacity of the model was judged by the externally validated coefficient of determination R^2_{pred}, which was 0.5694 for the external test set. The statistical parameter values of q^2, R, RMSE, and R^2_{pred} are presented in Table 3.

The performance of the conventional QSAR model built by using the GA-SVM method depended on the combination of features and three parameters, i.e., the kernel function parameter σ, penalty parameter C, and ε-insensitive loss function ε. The ε-insensitive loss function ε prevents the entire training set from meeting the boundary conditions and allows the possibility of scarcity in the dual formulation solutions. The initial number of molecular descriptors for each compound was 1,777. The number of descriptors decreased rapidly from 1,559 to 858 after the first generation. The QSAR model using these 858 descriptors gave RMSE of 0.4996. After 20,000 generations, the number of descriptors was reduced to eight and the RMSE was 0.03485. When the number of features reached five and the values of σ, C, and ε remained at 1.2410, 246.564, and 0.00829, the QSAR model gave the lowest RMSE. These results suggest that GA-SVM is useful for removing redundant descriptors and helpful for optimization of SVM parameters.

Table 3 Statistics for four regression models

Model	Methods	q^2	R	RMSE	R^2_{pred}
Intercalation	GA-SVM	0.9835	0.9937	0.1938	0.5694
	ANN	0.8020	0.8133	0.6723	0.5454
Groove binding	GA-SVM	0.6240	0.7585	0.4083	0.5137
	ANN	0.7429	0.9037	0.3376	0.5048

q^2, cross-validated correlation coefficient; R, correlation coefficient; $\mathrm{R}^2_{\mathrm{pred}}$, predicted correlation coefficient; *RMSE*, root-mean-square error

The five most significant descriptors were selected by GA-SVM as independent variables of the best model, being MB-ATS$_6$(VDW), MB-ATS$_6$(Q), C$_7$(Q), Ã$_1$(SAS), and $g_{\mathrm{SE}}(11.0)$. These descriptors encoded different aspects of the molecular structure. Autocorrelation descriptors calculated for 3D spatial molecular geometry are based on interatomic distances collected in the geometry matrix and the property function defined by the set of atomic properties. MB-ATS$_6$(VDW) and MB-ATS$_6$(Q), i.e., the van der Waals radius and the atomic charge weighted Moreau–Broto topological autocorrelation descriptor, describe how the van der Waals radius and atomic charge are distributed along the topological structure of a compound. C$_7$(Q) is the charge-weighted Geary topological autocorrelation descriptor. The Moreau–Broto and Geary topological autocorrelation descriptors are distance-type functions varying from zero to infinity. Ã$_1$(SAS) belongs to the molecular volume and molecular surface derived descriptors. $g_{\mathrm{SE}}(11.0)$ is an electronegativity weighted radial distribution function descriptor. The radial distribution function meets all the requirements for a 3D structure descriptor: it is independent of the atom number (i.e., the size of a molecule), it is unique regarding the 3D arrangement of the atoms, and it is invariant against translation and rotation of the entire molecule. Additionally, the radial distribution function (RDF) descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain 3D structure space (e.g., to describe the steric hindrance or the structure/activity properties of a molecule) [29]. The RDF descriptors are based on the distance distribution in the molecule. The radial distribution function of an ensemble of atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of given radius. For this descriptor, the weighting is the electronegativity, which shows that the electronegativity of the molecule plays a main role in this descriptor. An interesting result here was that the descriptors selected were mostly weighted by atomic charge.

Result of conventional QSAR model using GA-SVM for predicting binding affinity of groove binding

The 12 groove binders were randomly divided into a training set and an external test set; 8 compounds were included in the training set and were used to develop a

Table 4 Experimental $\log_{10}k'$ versus predicted $\log_{10}k'$ and corresponding residuals in the training set and test set for the groove binding regression model

Number	Exp ($\log_{10}k'$)	GA-SVM		ANN	
		Pre ($\log_{10}k'$)	Residual	Pre ($\log_{10}k'$)	Residual
19	7.6628	7.2935	0.3693	7.7245	−0.0617
20	6.3874	6.8644	−0.4770	6.3709	0.0165
21	7.6532	7.5528	0.1004	7.6636	−0.0104
22	6.3118	6.1759	0.1359	6.1543	0.1575
23	7.3424	6.9964	0.3460	7.7485	−0.4061
24	6.8751	7.5361	−0.6610	6.9564	−0.0813
25	7.1461	7.4758	−0.3297	6.9618	0.1843
26	8.4472	7.9219	0.5253	7.6242	0.8230
27*	6.0792	6.8269	−0.7477	6.7189	−0.6397
28*	7.0414	7.7751	−0.7337	7.9259	−0.8845
29*	8.477	7.7667	0.7103	7.9012	0.5758
30*	8.431	7.7104	0.7206	7.6331	0.7979

* Molecule included in the test set of the groove binding regression model

regression model, while the remaining 4 compounds were used in the external test set for assessing the prediction ability of the model. Threefold cross-validation was applied to build the model for the training set. The model was obtained with values $\sigma = 0.2462$, $C = 237.6482$, and $\varepsilon = 0.3333$. Then, the constructed model was used to predict the external test set. The prediction results are given in Table 4. The predicted values of binding affinity for the compounds in the training and test sets are plotted against the experimental values of the compounds in Fig. 3. The statistical parameter values are presented in Table 3.

In the model, the 4 most significant descriptors were selected from the 1,777 using the GA-SVM, being the Gutman molecular topological index (S_G), spectral moment of the edge adjacent matrix of order 10 ($\grave{I}_{10}$), electronegativity-weighted H-GETAWAY descriptor ($ATS_{7,SE}$), and core–core repulsion descriptor (E_n). Gutman molecular topological indices encode the interatomic connection information. The order 10 spectral moment ($\grave{I}_{10}$) corresponds to the sum of all self-returning walks of length 10 in the line graph of the molecular graph, which can be expressed as a linear combination of the embedding frequencies of the molecular graph. $ATS_{7,SE}$ encodes geometrical information given by the influence matrix, topological information given by the molecular graph, and chemical information from selected atomic properties, weighted by electronegativity [30, 31].

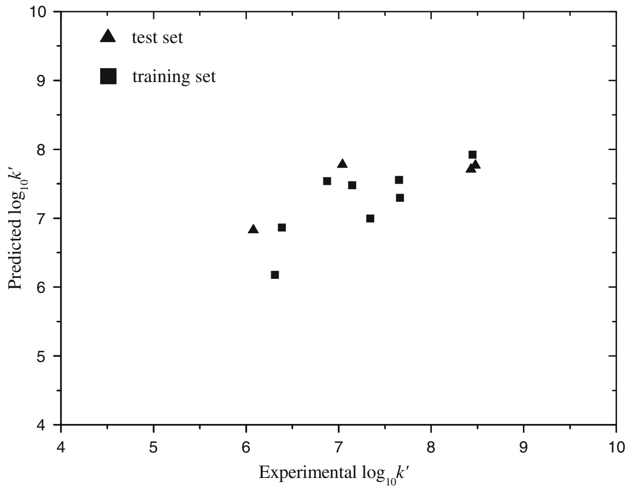


Fig. 3 Predicted versus experimental $\log_{10}k'$ for the training and test sets derived from the groove binding regression model built by GA-SVM

Result of conventional QSAR models using ANN for predicting binding affinity of intercalation and groove binding

Using the descriptors selected by the GA-SVM, we built two models using the ANN method. A three-layer network with a sigmoid transfer function was designed for ANN. Before training the networks, the input and output values were normalized between −1 and 1. The network was then trained using the training set by the backpropagation strategy for optimization of the weights and bias values. The proper number of nodes in the hidden layer was determined by training the network with different numbers of nodes in the hidden layer. The RMSE value measures how good the outputs were in comparison with the target values. All of the above-mentioned steps were carried out using backpropagation with the Levenberg–Marquardt update function. The predicted results and predicted errors for the two models based on 18 intercalators and 12 groove binders are listed in Table 3. The plots of the experimental versus predicted binding affinity ($\log_{10}k'$) are shown in Figs. 4 and 5. It is observed from the plots that the conventional QSAR models developed using GA-SVM provided better prediction than those using ANN. The calculated R^2_{pred} values of the two ANN models obtained using the same descriptors and the training and test sets were lower than the models obtained using GA-SVM, as presented in Table 3.

Experimental

Experimental data

Binding affinity constants (k') and binding modes of the 30 antibiotic compounds used in this study were taken from Refs. [32–36]. The structures of all antibiotic compounds are shown in Fig. 6. Binding affinity constants (k') were converted to a logarithmic scale ($\log_{10}k'$) for modeling purposes.

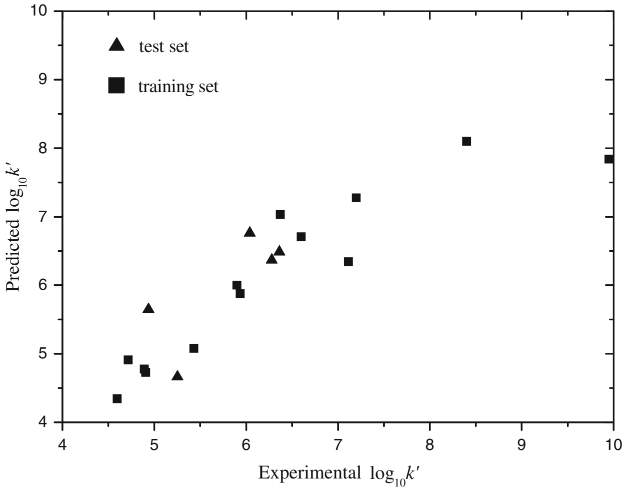


Fig. 4 Predicted versus experimental $\log_{10}k'$ for the training and test sets derived from the intercalation regression model built by ANN

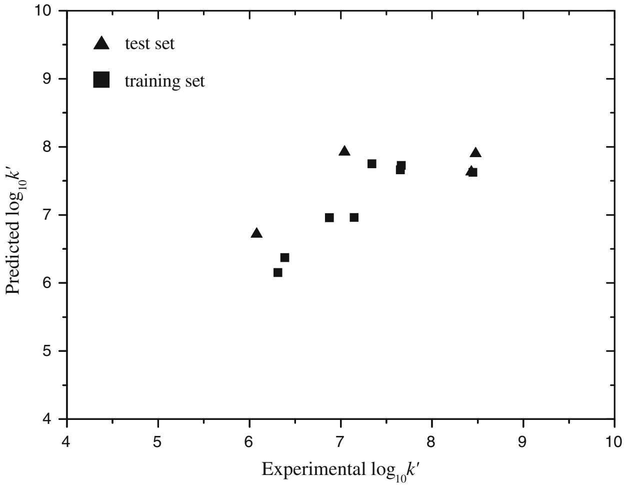


Fig. 5 Predicted versus experimental $\log_{10}k'$ for the training and test sets derived from the groove binding regression model built by ANN

Molecular descriptors

All molecules were drawn using Hyperchem [37] and pre-optimized using the MM+ molecular mechanics force field to generate their initial conformations. More precise optimization was done using the semiempirical PM3 method in Hyperchem. All calculations were carried out at the restricted Hartree–Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root-mean-square gradient reached 0.04 kJ mol^{-1} [38]. The resulting geometries were transferred to the Molecular Descriptor Lab (MODEL) [39] to calculate about 1,777 molecular descriptors of constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.), charge (minimum and maximum partial charges, dipole moment, polarity parameter, etc.), physical chemistry properties (hydrophobicity, molecular polarizability), topological indices (Wiener index, molecular path count, Geary topological autocorrelation descriptors, BCUT descriptors, Kier–Hall shape indices, etc.), geometrical (principal moments of inertia, radial distribution function, weighted holistic invariant molecular descriptors (WHIM) descriptors, 3D-MoRSE descriptors, geometry, topology and atom-weights assembly (GETAWAY) descriptors, charged partial surface area, etc.), and quantum chemistry (total energy, total softness, highest occupied molecular orbital (HOMO) and lower unoccupied molecular orbital (LUMO) energies, etc.). The MODEL web server is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi.

Support vector machine

The SVM developed by Vapnik [40] implemented the principle of structural risk minimization by constructing an optimal separating hyperplane. SVM is typically used to solve sample classification problems. However, with the introduction of the ε-insensitive loss function, SVM has been extended to solve nonlinear regression problems. The basic idea of SVM regression is to map the data X into a high-dimensional feature space F via a nonlinear mapping φ and then to solve a linear regression problem in this space. Given a set of data $\{(x_i, y_i)\}$ $i = 1, 2, \ldots, N$ (where x_i is the input vector, y_i is the actual value, and N is the total number of data patterns), the SVM regression function is

$$f(x) = \omega^{\mathrm{T}}\varphi(x_i) + b. \tag{1}$$

The coefficients ω and b are estimated by minimizing the following regularized risk function:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(y_i,f_i) + \frac{1}{2}\|\omega\|^2, \tag{2}$$

where

$$L_\varepsilon(y) = \begin{cases} 0 & \text{if} \quad |f(x) - y| \le \varepsilon \\ |f(x) - y| - \varepsilon & \text{otherwise} \end{cases} \tag{3}$$

and ε are prescribed parameters. Equation 3 is called the ε-insensitive loss function. This function defines a ε tube such that, if the predicted value is within the tube, the loss is zero, while if the predicted point is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius ε of the tube. By substituting the ε-insensitive loss function into the Eq. 2, the optimization object becomes

$$R(\omega, \xi, \xi_i^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{4}$$

with the constraints

$$\begin{cases} f(x) - y_i \leq \varepsilon + \xi_i^* \\ y_i - f(x) \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{5}$$

This constrained optimization problem is solved using the following primal Lagrangian form:

$$\begin{aligned} L = & \left(\omega, b, \xi_i, \xi_i^*, a_i, a_i^*, \beta_i\beta_i^*\right) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \\ & -\sum_{i=1}^{N} a_i(\xi_i + \varepsilon - y_i + f(x)) - \sum_{i=1}^{N} a_i^*\left(\xi_i^* + \varepsilon + y_i - f(x)\right) \\ & -\sum_{i=1}^{N}\left(\xi_i\beta_i + \xi_i^*\beta_i^*\right) \end{aligned} \tag{6}$$

Fig. 6 Structures of all antibiotic compounds

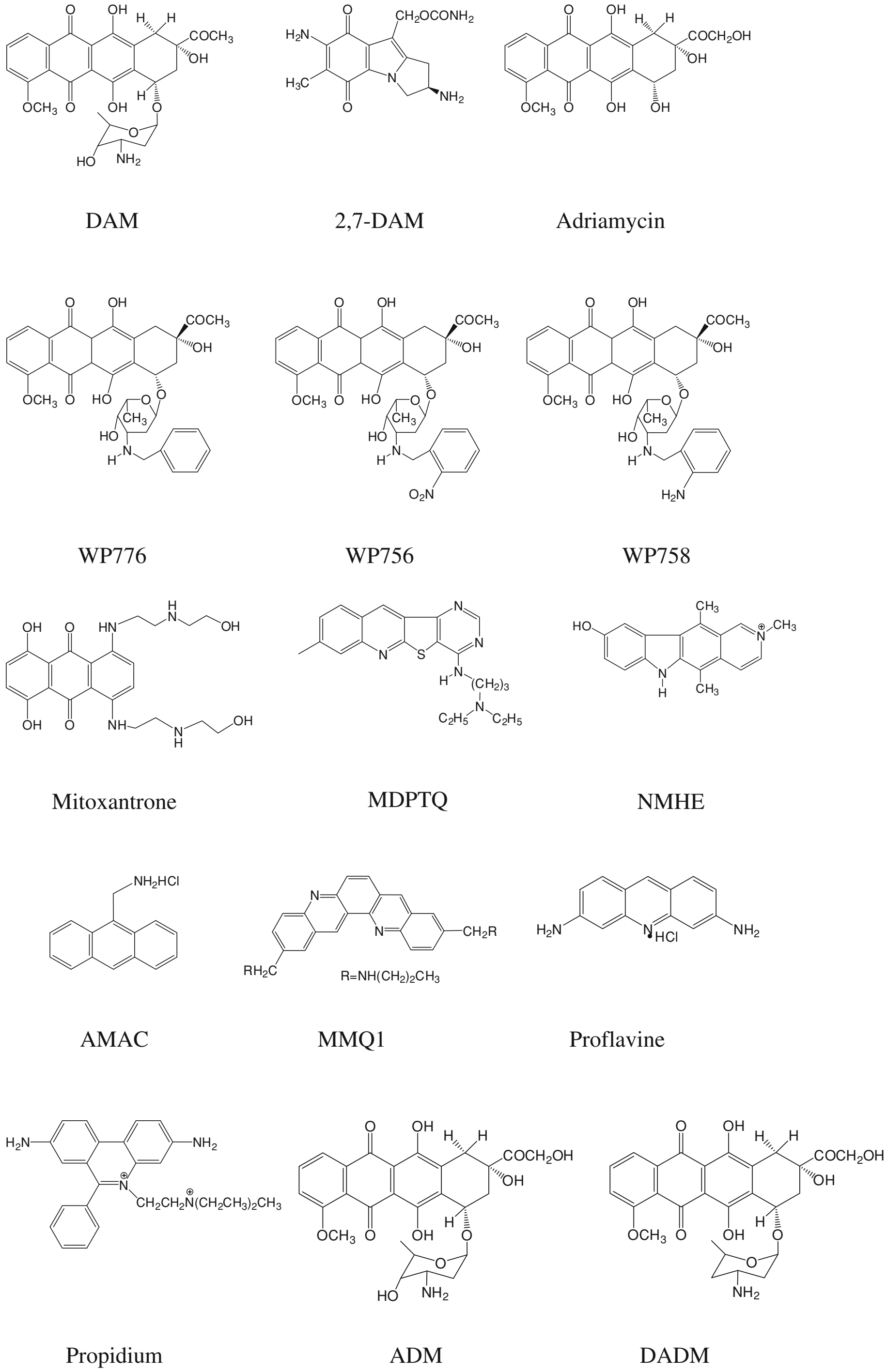
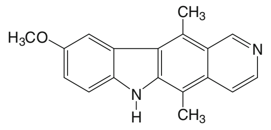
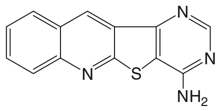


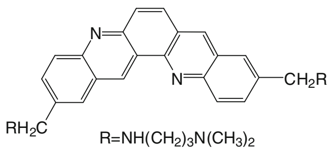
Fig. 6 continued



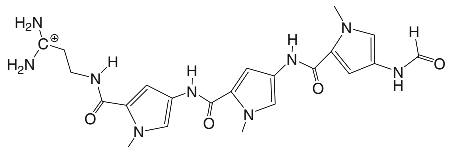
MHE



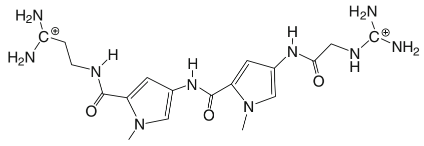
APTQ



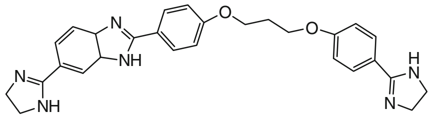
MMQ2



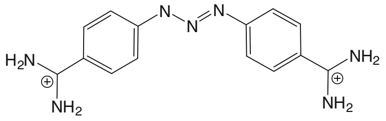
Distamycin



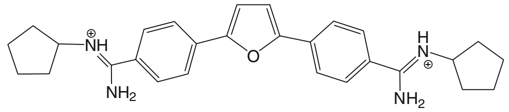
Netropsin



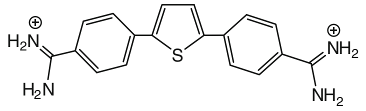
13SAB89



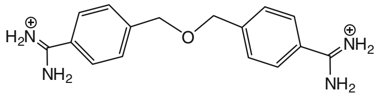
Berenil



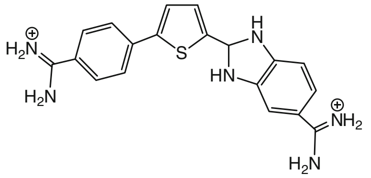
DB244



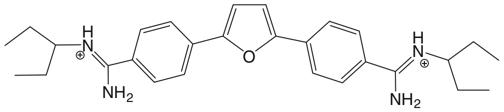
DB351



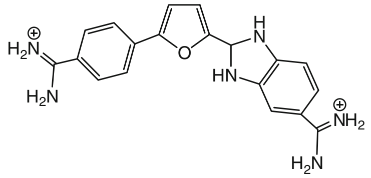
DB75



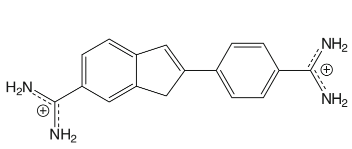
DB818



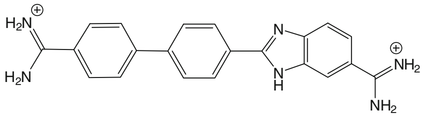
DB226



DB293



DAPI



DB921

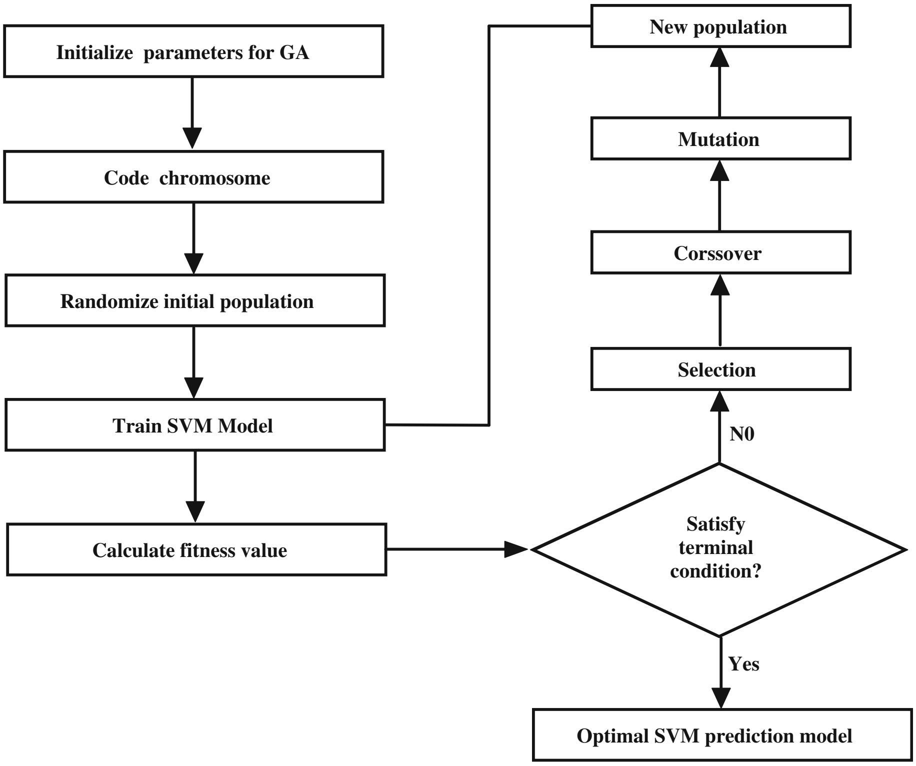

Initialize parameters for GA
Code chromosome
Randomize initial population
Train SVM Model
Calculate fitness value
Satisfy terminal condition?
N0
Selection
Corssover
Mutation
New population
Yes
Optimal SVM prediction model

Fig. 7 Optimization process of GA-SVM

Equation 6 is minimized with respect to primal variables ω, b, ξ_i, and ξ_i^*, and maximized with respect to nonnegative Lagrange multipliers $a_{i,}$ a_i^*, β_i, and β_i^*. Finally, Karush–Kuhn–Tucker conditions are applied to the regression, and Eq. 4 thus yields the dual Lagrangian

$$\theta(a_i, a_i^*) = \frac{1}{2}\sum_{i=1}^{N}(a_i - a_i^*)\left(a_j - a_j^*\right)k(x_i x_j) + \sum_{i=1}^{N}(a_i - a_i^*)y_i - \sum_{i=1}^{N}(a_i + a_i^*)\varepsilon \tag{7}$$

subject to

$$\sum_{i=1}^{N}(a_i - a_i^*) = 0,\ 0 \le a_i^j \le C,\ j = 1,\ 2;\ i = 1,\ 2,\ldots,N \tag{8}$$

Thus, the regression function (1) adopts the following form:

$$f(x) = \sum_{i=1}^{N}(a_i - a_i^*)k(x, x_i) + b, \tag{9}$$

where k is the kernel function mapping the data to a high-dimensional feature space. In this work, RBF $\exp\left(-\frac{1}{2}\left(\frac{|x_i - x_j|}{\sigma}\right)^2\right)$ is used in the SVM due to its good general performance [41].

The software used to implement SVM was LibSVM developed by Chang and Lin [42], which can be downloaded freely from http://www.csie.ntu.edu.tw/~cjlin/libsvm for academic purposes.

Genetic algorithm

Genetic algorithm is a stochastic global search technique based on the theory of natural selection and evolution. In GA, each of the candidate solutions to a problem is represented by a string called the chromosome. The fitness of each chromosome is then evaluated using a performance function after the chromosome has been decoded. Upon completion of the evaluation, a pair of chromosomes is randomly selected to undergo genetic operations that mimic natural phenomena observed in nature (such as crossover and mutation). This evolution process continues until the stopping criteria are reached. To date, GA has been increasingly applied in conjunction with other chemometric techniques such as ANN, MLR, partial least square (PLS), and SVM [43–50].

Genetic algorithm-support vector machine

GA-SVM focuses on improving the performance of the SVM model by integration of GA and SVM. The GA-SVM method for improving SVM performance has two aspects: feature subset selection and SVM parameter optimization.

The whole procedure for GA-SVM is illustrated in Fig. 7 and can be summarized as follows:

Step 1: initialize the parameters for GA-SVM such as the number of evolutionary generations, population size, crossover probability, mutation probability, the type of kernel function, and the range of the SVM parameters.
Step 2: encode chromosomes.
Step 3: randomly generate an initial population of chromosomes.
Step 4: run SVM model.
Step 5: calculate the fitness values of each chromosome in the population using the fitness function.
Step 6: if the terminal condition is reached, stop the process with the output of results; otherwise, go to the next step.
Step 7: select a given percentage of the fittest chromosomes from the current generation based on their fitness value. The selected chromosomes will then be used as parent chromosomes to produce new chromosomes by performing crossover and mutation operations.
Step 8: go back to the step 4.

Genetic algorithm-support vector machine for binary QSAR model

The procedure of GA-SVM started with the random selection of chromosomes which represent the feature set and the parameters of the SVM to send to the SVM model. The selected chromosomes were evolved by GA operators and assessed by the fitness function. The procedure stopped once the terminal condition was reached, then the obtained SVM model with the optimal feature set and the parameters was used for predicting the external test set.

In our implementation of GA, the feature set and the SVM parameters kernel function parameter σ and penalty parameter C were encoded as chromosomes. The chromosomes were represented using the binary system for the feature set and decimal coding system for parameters of the SVM. The feature set was encoded as binary strings. Each binary gene of the chromosome represents whether the corresponding feature was selected or not. A value of 1 means the corresponding feature was selected, whereas 0 means it was not selected. A chromosome consisted of 1,777 binary genes and two decimal genes, C and σ. The fitness function of the binary model must consider two aspects; one was to maximize the classification accuracy of fivefold cross-validation, and the other was to minimize the number of selected features. The performances of these two aspects can be evaluated by Eq. 10.

$$fitness(1) = accuracy + N \times w, \tag{10}$$

where N is the number of features selected and w is the weight of the feature number. The accuracy of the fivefold cross-validation was used as the SVM accuracy. In the fivefold cross-validation method, the training set was partitioned into five sets of size $n/5$ each. Among them, four sets were used for training and the remaining one was used for testing. The procedure was repeated five times, and the average prediction accuracy was computed. According to the fitness function, the appropriate chromosomes were selected to yield offspring. The termination criterion for the running of GA-SVM was minimum number of features and maximum achieved accuracy.

Genetic algorithm-support vector machine for conventional QSAR model

SVM has been employed to solve regression problems through the introduction of the ε-insensitive loss function ε. Therefore, there were four parts to the chromosome of the conventional QSAR model: C, σ, ε, and the feature set. The chromosome consisted of 1,777 binary genes for the feature set and three decimal genes, C, σ, and ε for the optimization of parameters. The fitness function of the conventional model has two aspects; one was to minimize the RMSE of the fivefold cross-validation, and the other was to minimize the number of selected features. The fitness function for the conventional QSAR model was presented as follows:

$$\text{fitness}(2) = \text{RMSE} + N \times w, \tag{11}$$

where N is the number of feature selected and w is the weight of feature number. RMSE is the root-mean-square error of the QSAR model, which measures how good the outputs are in comparison with the target values.

Model validation

A model should be validated both internally and externally to ensure that the built model is robust, reliable, stable, and predictive. The internal performance is characterized by the goodness of fit and robustness of the model. In the current work, several statistical terms such as the correlation coefficient (r—measure of the goodness of fit), cross-validated correlation coefficient (q^2—measure of robustness), and RMSE were used to assess the internal predictive ability of the model. The external performance is characterized by the predictive power of the model. The external prediction capacity of the model is judged by the externally validated coefficient of determination ($R^2_{\text{pred}} > 0.5$), which is defined as follows:

$$R^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_{\text{pred(test)}} - Y_{\text{test}}\right)^2}{\sum_{i=1}^{n}\left(Y_{\text{test}} - \bar{Y}_{\text{training}}\right)^2}, \tag{12}$$

where $Y_{\text{pred(test)}}$ and Y_{test} are, respectively, the predicted and experimental value of the test set compounds, and $\bar{Y}_{\text{training}}$

is the experimental mean value of the training set compounds [51, 52].

Acknowledgments This work was supported by the National Natural Science Foundation of China (Nos. 20945003, 21005063), and the Natural Science Foundation of Gansu (No. 096RJZA121).

References

1. Haq I (2002) Arch Biochem Biophys 403:1
2. Wan KX, Shibue T, Gross ML (2000) J Am Chem Soc 122:300
3. Chaires JB (2006) Arch Biochem Biophys 453:26
4. Lerman LS (1961) J Mol Biol 3:18
5. Waring M (1970) J Mol Biol 54:247
6. Wartell RM, Larson JE, Wells RD (1974) J Biol Chem 249:6719
7. Kumar GS, He QY, Behr-Ventura D, Tomasz M (1995) Biochemistry 34:2662
8. Qu XG, Chaires JB (2001) J Am Chem Soc 123:1
9. Gopal M, Shenoy S (2003) J Photochem Photobiol B 72:69
10. Ibrahim MS (2001) Anal Chim Acta 443:63
11. Leng FF, Priebe W, Chaires JB (1998) Biochemistry 37:1743
12. Wang SF, Peng TZ, Yang CF (2003) J Biochem Biophys Methods 55:191
13. Wang J, Ozsoz M, Cai XH, Rivas G, Shiraishi H, Grant DH, Chicharro M, Fernandes J, Palecek E (1998) Bioelectrochem Bioenerg 45:33
14. Agrawal P, Barthwal SK, Barthwal R (2009) Eur J Med Chem 44:1437
15. Evstigneev MP, Mykhina VY, Davies DB (2005) Biophys Chem 118:118
16. Barthwal R, Sharma U, Srivastava N, Jain M, Awasthi P, Kaur M, Barthwal SK, Govil G (2006) Eur J Med Chem 41:27
17. Mallena S, Lee MPH, Bailly C, Neidle S, Kumar A, Boykin DW, Wilson WD (2004) J Am Chem Soc 126:13659
18. Haj HTB, Salerno M, Priebe W, Kozlowski H, Garnier-Suillerot A (2003) Chem Biol Interact 145:349
19. Miao Y, Lee MPH, Parkinson GN, Batista-Parra A, Ismail MA, Neidle S, Boykin DW, Wilson DW (2005) Biochemistry 44:14701
20. Teulade-Fichou MP, Carrasco C, Guittat L, Bailly C, Alberti P, Mergny GL, David A, Lehn JM, Wilson WD (2003) J Am Chem Soc 125:4732
21. Li VS, Choi D, Wang Z, Jimenez LS, Tang MS, Kohn H (1996) J Am Chem Soc 118:2326
22. Guan Y, Shi R, Li XM, Zhao MP, Li YZ (2007) J Phys Chem B 111:7336
23. Tian L, Wei WZ, Mao Y (2004) Clin Biochem 37:120
24. Lu XQ, Wang L, Liu HD, Chen J (2007) Talanta 73:444
25. Chen J, Lu XQ (2009) Talanta 79:129
26. Ismail I, Francois P, Elodie D, Olivier B, Andre M (2007) Bioorg Med Chem 15:4256
27. Menard PR, Lewis RA, Mason JS (1998) J Chem Inf Comput Sci 38:497
28. Menard PR, Mason JS, Morize I, Bauerschmidt S (1998) J Chem Inf Comput Sci 38:1204
29. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
30. Consonni C, Todeschini R, Pavan M (2002) J Chem Inf Comput Sci 42:682
31. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) J Chem Inf Comput Sci 42:693
32. Kumar CV, Asuncion EH (1993) J Am Chem Soc 115:8541
33. Mazur S, Tanious FA, Ding D, Kumar A, Boykin DW, Simpson IJ, Neidle S, Wilson WD (2000) J Mol Biol 300:321
34. Rahimian M, Kumar A, Say M, Bakunov SA, Boykin DW, Tidwell RR, Wilson WD (2009) Biochemistry 48:1573
35. Banerjee D, Pal SK (2008) J Phys Chem B 112:1016
36. Athri P, Wilson WD (2009) J Am Chem Soc 131:7618
37. HyperChem Release 7, HyperCube Inc. (2002), http://www.hyper.com
38. Long W, Liu P, Li X, Xu Y, Yu J, Ma S, Yu L, Zou Z (2009) J Chemom 23:304
39. Li ZR, Han LY, Xue Y, Yap CW, Li H, Jiang L, Chen YZ (2007) Biotechnol Bioeng 97:389
40. Cortes C, Vapnik V (1995) Mach Learn 20:273
41. Wang WJ, Xu ZB, Lu WZ, Zhang XY (2003) Neurocomputing 55:643
42. Chang CC, Lin CJ (2001) http://www.csie.ntu.edu.tw/cjlin/libsvm
43. Fang SF, Wang MP, Qi WH, Zheng F (2008) Comput Mater Sci 44:647
44. Pai PF, Hong WC (2005) Electric Power Syst Res 74:417
45. Pourbasheer E, Riahi S, Ganjali MR, Norouzi P (2009) Eur J Med Chem 44:5023
46. Dolatabadi M, Nekoei M, Banaei A (2010) Monatsh Chem 141:577
47. Goodarzi M, Freitas MP, Wu CH, Duchowicz PR (2010) Chemometr Intell Lab Syst 101:102
48. Li ZC, Zhou XB, Lin YR, Zou XY (2008) Amino Acids 35:581
49. Li TH, Mei H, Cong PS (1999) Chemometr Intell Lab Syst 45:177
50. Habibi-Yangjeh A (2009) Monatsh Chem 140:523
51. Corwin H, Rajeshwar PV (2009) Mol Pharmaceutics 6:3849
52. Eduardo BM, Marcia MCF (2009) Eur J Med Chem 44:3577