European Journal of Medicinal Chemistry 44 (2009) 5023–5028

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

# Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity

Eslam Pourbasheer [a], Siavash Riahi [a,b,*], Mohammad Reza Ganjali [a,c], Parviz Norouzi [a,c]

[a] Center of Excellence in Electrochemistry, Faculty of Chemistry, University of Tehran, P.O. Box 14155-6455, Tehran, Iran
[b] Institute of Petroleum Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran
[c] Faculty of Pharmacy, Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

*Article history:*
Received 7 March 2009
Received in revised form
11 July 2009
Accepted 4 September 2009
Available online 12 September 2009

*Keywords:*
BK-channels activity
QSAR
Support vector machine
Genetic algorithms
Chemometrics

ABSTRACT

The support vector machine (SVM), which is a novel algorithm from the machine learning community, was used to develop quantitative structure–activity relationship (QSAR) for BK-channel activators. The data set was divided into 57 molecules of training and 14 molecules of test sets. A large number of descriptors were calculated and genetic algorithm (GA) was used to select variables that resulted in the best-fitted for models. A comparison between the obtained results using SVM with those of multi-parameter linear regression (MLR) revealed that SVM model was much better than MLR model. The improvements are due to the fact that the activity of the compounds demonstrates non-linear correlations with the selected descriptors. Also distances between Oxygen and Chlorine atoms, the mass, the van der Waals volume, the electronegativity, and the polarizability of the molecules are the main independent factors contributing to the BK-channels activity of the studied compounds.



## 1. Introduction

The large-conductance calcium-activated potassium (BK) channels are expressed in excitable as well as in non-excitable cells. They control several cell functions in the nervous system, BK channels contribute to the shaping of action potential and modulate the neuronal excitability and the release of neurotransmitters; also, BK channels play a fundamental role in the regulation of the tone of smooth muscle cells [1,2]. BK-activators can guarantee an innovative pharmacological tool for the clinical management of many pathological states, due to a cell hyperexcitability, such as asthma, urge incontinence and bladder spasm, gastric hypermotility, neurological and psichiatric disorders [1,2]. Also, Potassium channel activators have been indicated as emerging drugs for the therapy of several cardiovascular, respiratory or CNS diseases [3].

Although there are several experimental methods available for screening the biological activity of chemicals (e.g. in vivo and in vitro assay tests), and these all have also been carried out using receptors and other biological materials of human, rat, mouse, and calf origin at least [4], they are costly, time-consuming. This has meant that the development of computational methods as an alternative tool for predicting properties of chemicals has been a subject of intensive study. Among the computational methods, the quantitative structure–activity relationships (QSAR), has found diverse applications for predicting compounds' properties, including biological activity prediction [5–7], physical property prediction [8–10], and toxicity prediction [11–13]. In QSAR studies, there are some techniques which can be applied for the construction of model, such as multiple linear regression (MLR) and artificial neural networks (ANN) that were used for inspection of linear and nonlinear relation between interested activity and molecular descriptors, respectively. The flexibility of neural networks enables them to discover more complex nonlinear relationships in experimental data [14]. Neural networks have some problems inherent to its architecture, such as overtraining, overfitting, network optimization, and reproducibility of results, due to random initialization of the networks and variation of stopping criterias [15]. Owing to these reasons there is a tendency to use more accurate and informative techniques in QSAR analysis. The support vector machine (SVM) is a new algorithm developed from the machine learning community [16]. SVM approach automatically controls the flexibility of the resulting classifier on the training data. Accordingly, by the design of the algorithm, the deteriorating effect of the input dimensionality on the generalization ability is largely suppressed. Due to its remarkable

* Corresponding author. Center of Excellence in Electrochemistry, Faculty of Chemistry, University of Tehran, P.O. Box 14155-6455, Tehran, Iran. Tel.: +98 21 61114714; fax: +98 21 88632976.
*E-mail address:* riahisv@khayam.ut.ac.ir (S. Riahi).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.09.006

generalization performance, the SVM has attracted attention and gained extensive application, such as pattern recognition problems [17] drug design [18] QSAR [19] and quantitative structure-property relationship (QSPR) analysis [20]. In most of these cases, performance of SVM modeling either matches or is significantly better than that of traditional machine learning approaches. The application of these techniques usually requires variable selection for building well-fitted models. Nowadays, GA is well-known as an interesting and more widely used variable selection method [21–23]. GA is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation.

Recently, separate multi-parameter linear QSAR models have been proposed for BK-Channel Activators [24]. In the present work, linear (GA-MLR) and non-linear (GA-SVM) methods were employed to generate QSAR models and the obtained results with GA-MLR method were compared with the GA-SVM method and also the experimental values. According to the literature survey, this is the first research on the QSAR of the BK channel activity using GA-SVM technique.

## 2. Data and methodology

The data set of $pIC_{50}$ values for 71 BK-Channel Activators used for the QSAR analyses was selected from the literature [24]. The $IC_{50}$ expresses the parameter of potency, representing the molar concentration of the tested compounds, which evokes half-reduction of the contractile tone induced by KCl 20 Mm [25]. The data set was randomly split into training, and testing sets (57 and 14 compounds, respectively) (Table 1). The z-matrices (molecular models) were constructed with HyperChem 7.0 and molecular structures were optimized using AM1 algorithm [26]. In order to calculate the theoretical descriptors, Dragon package version 2.1 was used [27]. For this propose the output of the HyperChem software for each compound fed into the Dragon program and the descriptors were calculated. As a result, a total of 1481 theoretical descriptors were calculated for each compound in data sets (71 compounds).

The theoretical descriptors were reduced by the following procedure:

1) Descriptors that are constant have been eliminated (324 descriptors). 2) in addition, to decrease the redundancy existing in the descriptors, the correlation of descriptors with each other and with $pIC_{50}$ of the molecules are examined, and collinear descriptors ($R > 0.9$) are detected. Among the collinear descriptors, one that has the highest correlation with activity is retained, and the others are removed from the data matrix (688 descriptors).

We used MLR as a linear technique and SVM as a non-linear feature mapping technique for the construction of QSAR models in this work. Since SVM method is not be able to select the most significant descriptors from the pool of calculated molecular descriptors, it would be necessary to use variable selection method. In the present work genetic algorithm (GA) variable subset selection method [28] was used for the selection of the most relevant descriptors from the pool of remaining 468 descriptors. These descriptors would be used as inputs of MLR and SVM.

### 2.1. Genetic algorithm (GA)

To select the most relevant descriptors, evolution of population was simulated [29–32]. Each individual of the population defined by a chromosome of binary values represented a subset of descriptors. The number of genes at each chromosome was equal to the number of descriptors. The population of the first generation was selected randomly. A gene took a value of 1 if its corresponding descriptor was included in the subset; otherwise, it took a value of zero. The number of genes with a value of 1 was kept relatively low to have a small subset of descriptors, that is, the probability of generating 0 for a gene was set greater (at least 60%) than the value of 1. The operators used here were crossover and mutation. The probability of the application of these operators was varied linearly with generation renewal (0–0.1% for mutation and 60–90% for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness [33] The GA program was written in Matlab 6.5 [34].

### 2.2. Support vector machine

Support vector machine, developed by Vapnik and Cortes [35] as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. A detailed description of SVM theory can be found in several excellent books and tutorials [36–39]. Here, only a brief description is given. In support vector machine, the input data is first mapped into high dimensional feature space by the use of kernel function and then linear regression is performed in the feature space. The non-linear feature mapping will allow the treatment of non-linear problems in a linear space. After training on set data SVM can be used to predict the objects whose values are unknown. The prediction or approximation function used by SVM is:

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

where $\alpha_i$ is some real value, $x_i$ is a feature vector corresponding to a training object. The components of vector $\alpha$ and the constant $b$ represent the hypothesis and are optimized during training. $K(x,x_i)$ is a kernel function. Training points with nonzero weight $\alpha_i$ are called support vectors. The elegance of using a kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly, and it may be useful to think of the kernel, $K(x,x_i)$ as comparing patterns or evaluating the proximity of objects in their feature space. Thus, a test point is evaluated by comparing it with all training points. In the function estimation problems, the Gaussian radial basis function kernel is most commonly used because of its effectiveness and speed in the training process. The form of the Gaussian function in $R$ is:

$$K(u, v) = \exp\left( -\gamma^{*} |u - v|^2 \right) \tag{2}$$

where $\gamma$ is the parameter of the kernel, $u$ and $v$ are two independent variables.

## 3. Results and discussion

For the selection of the most important descriptors, genetic algorithm variable subset selection method was used. According to rule of thumb [40,41] at least five data points (compounds) should be included in the equation for every parameter (descriptor). On the other hand, the ratio of 5 training molecules for each descriptor must be included in the equation. For this reason to select the optimum number of descriptors the influences of the number of descriptors was investigated from one to ten descriptors.

The $R^2$ value can be generally increased by adding the additional predictor variables to the model, even if the added variable does not contribute to the reduction of the unexplained variance of the dependent variable. Therefore, the $R^2$ usage requires special

**Table 1**
The data set and the corresponding observed and predicted $pIC_{50}$ values by GA-MLR and GA-SVM methods.

| Number | Name[a] | Exp ($pIC_{50}$) | GA-MLR | GA-SVM |
|---|---|---|---|---|
| Training set | | | | |
| 1 | II-2c | 3.81 | 4.23 | 4.15 |
| 2 | II-2a | 3.95 | 3.59 | 3.96 |
| 3 | II-2e | 4.03 | 4.03 | 4.06 |
| 4 | II-6b | 4.03 | 3.93 | 4.06 |
| 5 | III-7 | 4.10 | 4.61 | 4.13 |
| 6 | III-4b | 4.26 | 4.5 | 4.53 |
| 7 | II-2f | 4.43 | 4.32 | 4.46 |
| 8 | IX-6d | 4.51 | 4.96 | 4.54 |
| 9 | IX-7a | 4.51 | 4.88 | 4.54 |
| 10 | X-7l | 4.52 | 4.95 | 4.55 |
| 11 | I-2c | 4.55 | 4.56 | 4.58 |
| 12 | IX-3f | 4.57 | 4.35 | 4.60 |
| 13 | I-2d | 4.61 | 4.37 | 4.64 |
| 14 | IV-7 | 4.71 | 5.94 | 5.35 |
| 15 | X-7e | 4.73 | 4.86 | 4.76 |
| 16 | III-10 | 4.75 | 4.53 | 4.73 |
| 17 | X-11a | 4.76 | 5.18 | 4.79 |
| 18 | IX-6a | 4.77 | 4.66 | 4.80 |
| 19 | IX-7d | 4.80 | 4.89 | 4.82 |
| 20 | IX-3a | 4.82 | 5.39 | 4.85 |
| 21 | IX-7b | 4.82 | 4.77 | 4.79 |
| 22 | II-2b | 4.86 | 4.61 | 4.82 |
| 23 | IX-3c | 4.87 | 4.84 | 4.84 |
| 24 | X-7f | 4.90 | 5.25 | 4.93 |
| 25 | IX-6c | 4.91 | 5.47 | 4.94 |
| 26 | IX-3b | 4.93 | 4.92 | 4.96 |
| 27 | IX-7c | 4.93 | 5.08 | 4.90 |
| 28 | X-9a | 4.93 | 6.06 | 5.43 |
| 29 | X-11b | 4.95 | 5.06 | 4.98 |
| 30 | X-7h | 4.95 | 4.92 | 4.98 |
| 31 | IV-1d | 4.96 | 5.07 | 4.95 |
| 32 | IX-3e | 4.99 | 5.16 | 4.98 |
| 33 | X-7b | 5.03 | 4.87 | 5.00 |
| 34 | V-12f | 5.09 | 4.73 | 5.06 |
| 35 | VIII-14b | 5.11 | 5.67 | 5.14 |
| 36 | IV-1c | 5.16 | 4.81 | 5.13 |
| 37 | IV-1f | 5.23 | 5.55 | 5.26 |
| 38 | NS1619 | 5.23 | 4.9 | 5.20 |
| 39 | X-4b | 5.28 | 5.11 | 5.25 |
| 40 | IV-1e | 5.46 | 5.31 | 5.43 |
| 41 | IV-3f | 5.48 | 5.59 | 5.47 |
| 42 | IV-1a | 5.51 | 5.17 | 5.38 |
| 43 | X-7g | 5.51 | 5.53 | 5.48 |
| 44 | IV-3d | 5.52 | 5.13 | 5.49 |
| 45 | IV-1b | 5.6 | 5.94 | 5.63 |
| 46 | IV-3c | 5.61 | 5.47 | 5.58 |
| 47 | IV-3e | 5.67 | 5.77 | 5.64 |
| 48 | V-12d | 5.84 | 5.38 | 5.68 |
| 49 | X-4c | 6.16 | 6.38 | 6.13 |
| 50 | V-12e | 6.35 | 6.58 | 6.38 |
| 51 | V-17 | 6.43 | 5.84 | 6.40 |
| 52 | V-12a | 6.6 | 5.26 | 5.61 |
| 53 | V-12c | 6.72 | 6.71 | 6.64 |
| 54 | VIII-5b | 6.75 | 6.17 | 6.23 |
| 55 | V-12g | 6.87 | 6.25 | 5.96 |
| 56 | VIII-14a | 7.56 | 6.53 | 6.50 |
| 57 | V-16b | 8.49 | 8.16 | 7.59 |
| Test set | | | | |
| 58 | I-2a | 3.98 | 3.55 | 4.17 |
| 59 | II-2d | 4.26 | 4.42 | 4.28 |
| 60 | X-7c | 4.52 | 5.53 | 5.23 |
| 61 | I-2b | 4.73 | 4.13 | 4.56 |
| 62 | I-2f | 4.79 | 4.67 | 4.67 |
| 63 | X-7a | 4.87 | 5.51 | 5.10 |
| 64 | III-5b | 5.06 | 4.84 | 5.16 |
| 65 | X-4a | 5.39 | 3.93 | 4.70 |
| 66 | VI-16 | 5.58 | 5.4 | 5.35 |
| 67 | IV-3a | 5.81 | 5.14 | 5.30 |
| 68 | VIII-6b | 6.19 | 5.47 | 5.39 |
| 69 | VI-3b | 6.55 | 5.77 | 5.94 |
| 70 | IV-3b | 6.84 | 5.81 | 5.70 |
| 71 | V-16a | 7.56 | 5.63 | 5.49 |

[a] The name of compounds are according to the reference [24].

attention. For this reason, it is better to use another statistical parameter, named as adjusted $R^2(R^2_{adj})$, were $R^2_{adj}$ is defined by equation (3).

$$R^2_{adj} = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p-1}\right) \quad (3)$$

$R^2_{adj}$ is interpreted similarly to the $R^2$ value, considering the number of degrees of freedom, as well. It is adjusted by dividing the residual sum of squares and total sum of squares by their respective degrees of freedom. The $R^2_{adj}$ value diminishes if an added variable to the equation does not reduce the unexplained variance [41]. Subsequently, $R^2_{adj}$ is utilized to compare models with different numbers of predictor variables.

Another statistical parameter is the standard error of the estimate ($s$) that measures the dispersion of the observed values about the regression line. When the $s$ value is low, the reliability of the prediction is higher. Fig. 1 shows the plots of $R^2$, $R^2_{adj}$, and $s$ for the training set as a function of the number of descriptors for the 1–10 parameter models. $R^2$ and $R^2_{adj}$ are increased with the increasing number of descriptors. However, the values of $s$ decreased with the increasing number of descriptors. As can be seen, the models with 9 and 10 descriptors did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved with maximum 8 descriptors.

The selected variables and the correlation matrix of these descriptors are visualized shown in Table 2. From Table 2, it could be seen that the correlation coefficient value of each pair descriptors was less than 0.62, which meant that the selected descriptors were independent.

To examine the relative importance as well as the contribution of each descriptor in the model, for each descriptor the value of the mean effect ($MF$) was calculated. This calculation was performed with the equation (4).

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_{j}^{m} \beta_j \sum_{i}^{n} d_{ij}} \quad (4)$$

$MF_j$ represents the mean effect for the considered descriptor $j$, $\beta_j$ is the coefficient of the descriptor $j$, $d_{ij}$ stands for the value of the target descriptors for each molecule and, eventually, $m$ is

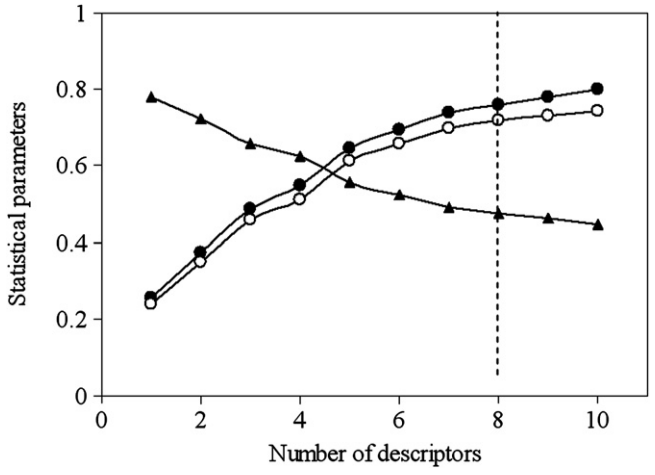


**Fig. 1.** Influences of the number of descriptors on the $R^2$ (●), $R^2_{adj}$ (○) and $s$ (▲) of the regression model.

**Table 2**
Correlation coefficient matrix of the selected descriptors.

| | G(O···Cl) | RDF105m | RDF040e | Mor05v | Mor19v | Mor08p | HATS8m | R7u+ |
|---|---|---|---|---|---|---|---|---|
| G(O···Cl) | 1 | | | | | | | |
| RDF105m | 0.294 | 1 | | | | | | |
| RDF040e | 0.237 | 0.037 | 1 | | | | | |
| Mor05v | −0.091 | −0.271 | −0.615 | 1 | | | | |
| Mor19v | −0.164 | 0.071 | 0.079 | −0.209 | 1 | | | |
| Mor08p | −0.074 | −0.210 | 0.053 | 0.299 | −0.176 | 1 | | |
| HATS8m | 0.588 | −0.036 | 0.327 | −0.185 | −0.329 | −0.124 | 1 | |
| R7u+ | 0.060 | −0.344 | 0.348 | −0.186 | −0.281 | 0.099 | 0.485 | 1 |

the descriptors number in the model. The *MF* value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign exhibits the variation direction in the toxicity values as a result of the increase (or reduction) of this descriptor values. The mean effect values are 0.058, −0.049, −0.649, 1.879, 0.183, −0.353, 0.304 and −0.373 for G(O···Cl), RDF105m, RDF040e, Mor05v, Mor19v, Mor08p, HATS8m and R7u+, respectively. By interpreting the descriptors contained in the model, it is possible to gain useful chemical insights about the activity of compounds. For this reason, an acceptable interpretation of the QSAR results is provided below.

G(O···Cl) (sum of geometrical distances between O···Cl) is one of the geometrical descriptors which has been appeared in the model. The value of this descriptor related to the distances between Oxygen and Chlorine atoms in compounds. As can be seen the G(O···Cl) mean effect has a positive sign which indicates that $pIC_{50}$ is directly related to this descriptor; therefore, increasing the G(O···Cl) of molecules leads to increase in its $pIC_{50}$ values.

The second and third descriptors are RDF105m and RDF040e, which belongs to the radial distribution function (RDF) descriptors. RDF in these forms meets all the requirements for the 3D structure descriptors. It is independent of the atom number (i.e. the size of a molecule), it is unique regarding the three-dimensional arrangement of the atoms and it is invariant against the translation and rotation of the entire molecule. Additionally, the RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space (e.g. to describe the steric hindrance or the structure/activity properties of a molecule). The radial distribution function (RDF) descriptors are based on the distance distribution in the molecule. The radial distribution function of an ensemble of *n* atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius *R*. In these descriptors (RDF105m and RDF040e) weighting schemes are the atomic masses and the atomic Sanderson electronegativities respectively, that shows the mass and the electronegativity of the molecules play main role in this descriptor. RDF105m and RDF040e display a negative sign, which indicates that the $pIC_{50}$ is inversely related to these descriptors. From above discussing it was concluded that by decreasing the molecular mass and molecular electronegativity the value of these descriptors decreased, causing a increasing in their $pIC_{50}$ value.

Mor05v, Mor19v and Mor08p are the other descriptors, appearing in the model which are belong to the 3D-MoRSE descriptors. 3D-MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction) are derived from Infrared spectra simulation using a generalized scattering function [42]. Mor05v and Mor19v were proposed as signal 05/weighted by atomic van der Waals volumes and signal 19/weighted by atomic van der Waals volumes respectively which relate to the van der Waals volume of the molecules. These descriptors display a positive sign, which indicate that the $pIC_{50}$ is directly related to this descriptor. Therefore, increasing the van der Waals volume of the molecules leads to increase in its $pIC_{50}$ values. Mor08p was proposed as signal 08/weighted by atomic polarizabilities which relates to polarizabilitie of the molecules. Mor08p displays a negative sign, which indicates that the $pIC_{50}$ is inversely related to this descriptor. Decreasing the polarizability of the molecules leads to increase in its $pIC_{50}$ values.

HATS8m and R7u+ belong to the GETAWAY descriptors. The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors have been recently proposed as chemical structure descriptors derived from a new representation of molecular structure, the Molecular Influence Matrix (MIM) [43]. HATS8m is the leverage-weighted total index weighted by atomic masse and R7u+ is the *R* maximal autocorrelation of lag 7/unweighted.

In summary, it is concluded that the distances between Oxygen and Chlorine atoms, the molecular mass, the van der Waals volume of the molecules, the molecular electronegativity, and the molecular polarizability play main role in BK-channel activators features which are going to design.

### 3.1. Genetic algorithm-multi-parameter linear regression

Multi-parameter linear correlation of $pIC_{50}$ values for 57 BK-channel activators in training set was obtained using eight selected descriptors by GA, and the following equation was obtained:

$$\begin{aligned} pIC_{50} = {} & 2.83(\pm0.43) + 0.02(\pm0.01)G(O\cdots Cl) - 0.32(\pm0.10) \\ & \times RDF105m - 0.12(\pm0.03)RDF040e - 1.56(\pm0.19) \\ & \times Mor05v + 2.47(\pm0.43)Mor19v + 1.34(\pm0.31) \\ & \times Mor08p + 4.38(\pm0.90)HATS8m - 7.05 \\ & (\pm2.08)R7u+ \end{aligned} \quad (5)$$

Then the built model was used to predict the test set (14 compounds). The prediction results are given in Table 1. Also the

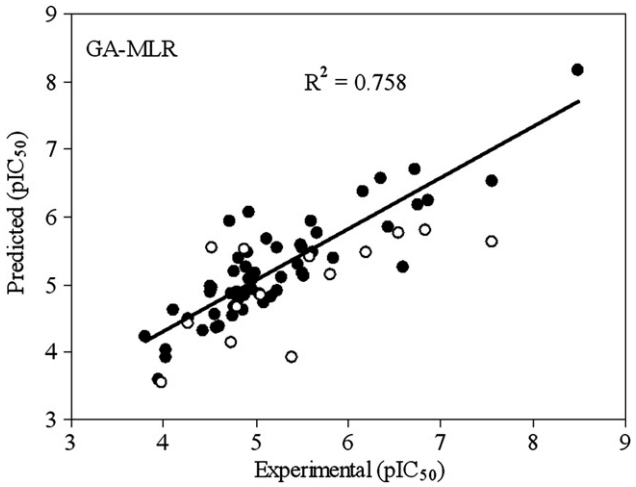


**Fig. 2.** Plot of the calculated values of $pIC_{50}$ from the GA-MLR model versus the experimental values for training (●) and test (○) sets.

**Table 3**
Statistical results of different QSAR models.

| | Training set | | | | Test set | | Whole data set | | |
|---|---|---|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | $Q^2_{LOO}$ | $F$ | RMSE | $R^2$ | RMSE | $R^2$ | $R^2_{CR}$ |
| MLR | 0.436 | 0.758 | 0.675 | 18.824 | 0.874 | 0.456 | 0.550 | 0.650 | 0.173 |
| SVM | 0.295 | 0.924 | 0.712 | 664.433 | 0.795 | 0.659 | 0.428 | 0.832 | 0.201 |

calculated values of $pIC_{50}$ for the compounds in training and test sets using the GA-MLR model have been plotted versus the experimental values of it (Fig. 2). The correlation coefficient $R^2$ was obtained to be 0.758 for the training set and 0.456 for the test set. Table 3 shows the statistical results for whole, training, and test sets.

The model obtained was validated using leave-one-out (LOO) cross-validation process. For LOO cross-validation, a data point is removed from the set, and the model is recalculated. The predicted activity for that point is then compared to its actual value. This is repeated until each data point is omitted once. The cross-validated correlation coefficient ($Q^2_{LOO}$) is 0.675.

The robustness of the resultant models was also validated with the chance correlation procedure. For a set of the various BK-channel activators, the $pIC_{50}$ values were randomly attributed to the molecules. Then, the MLR modeling with the selected descriptors was repeated with the randomized data [44]. The randomization was repeated 20 times. The squared correlation coefficient of the model with the randomly selected data ($R^2_{CR}$) was used as statistical criterion. If the $R^2_{CR}$ values of these models were much lower than those of the original model, it could be considered that the model was reasonable, and had not been obtained by the chance. The low chance correlation value ($R^2_{CR} = 0.173$) confirm this result.

### 3.2. Genetic algorithm-support vector machine

In order to develop a more accurate model, SVM was used to develop a model by the training set compounds based on the same subset of selected descriptors. LOO cross-validation method implied in SVM was used to build the model. Performance of SVM for regression depends on the combination of several factors. They are kernel function type, capacity parameter $C$, $\varepsilon$ of $\varepsilon$-insensitive loss function, and its corresponding parameters.

Firstly, the kernel function should be decided, which determines the sample distribution in the mapping space. The radial basis function (RBF) is commonly used in many studies because of its good general performance and few parameters to be adjusted [45].

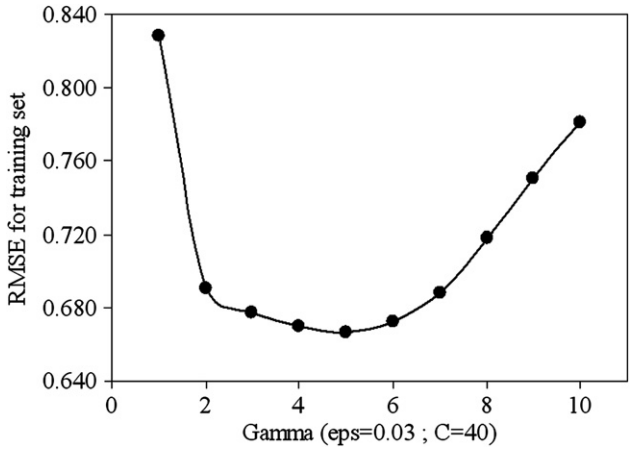


**Fig. 3.** The gamma versus RMSE for the training set ($\varepsilon = 0.03$; $C = 40$).

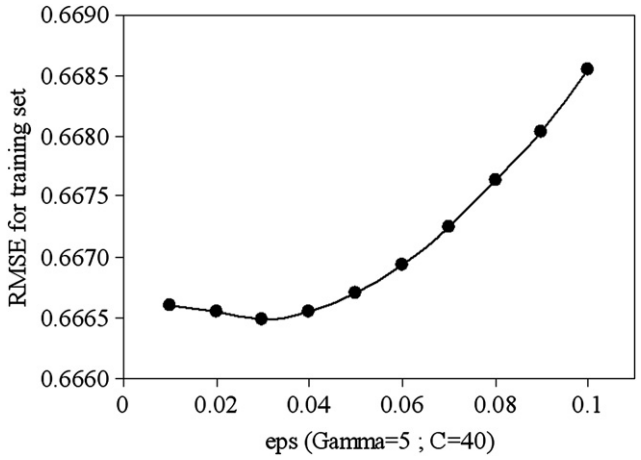


**Fig. 4.** The epsilon versus RMSE for the training set (Gamma = 5; $C = 40$).

In this work, the radial basis function was used, the form of which in $R$ is as follows:

$$\exp\left(-\gamma^{*}|u-v|^2\right)$$

where $\gamma$ is a parameter of the kernel; $u$ and $v$ are two independent variables.

Secondly, corresponding parameters, i.e. $\gamma$ of the kernel function greatly affect the number of support vectors, which has a close relation with the performance of the SVM and training time. Too many support vectors could produce overfitting and increase the training time. In addition, $\gamma$ controls the amplitude of the RBF function and, therefore, controls the generalization ability of SVM. The plot of $\gamma$ versus RMSE on the LOO cross-validation is shown in Fig. 3. As can be seen from the figure, the optimal $\gamma$ was 5.

Parameter $\varepsilon$-insensitive prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. The optimal value for $\varepsilon$ depends on the type of noise present in the data, which is usually unknown. The RMS error of LOO cross-validation on different epsilon is recorded in Fig. 4 and the optimal value was found to be 0.03.

Lastly, the effect of capacity parameter $C$ was tested. It controls the trade-off between maximizing the margin and minimizing the training error. If $C$ is too small then insufficient stress will be placed on fitting the training data. If $C$ is too large then the algorithm will

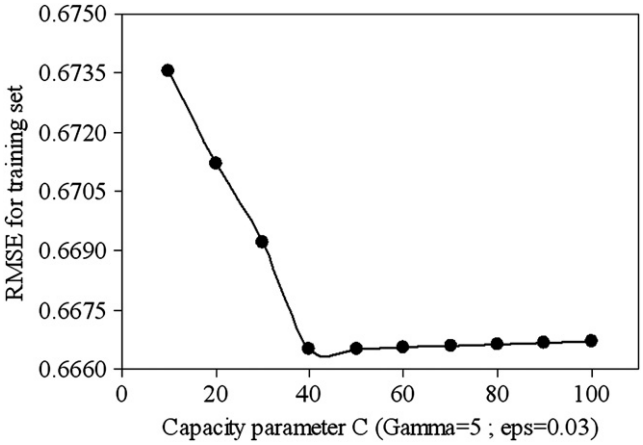


**Fig. 5.** The capacity parameter $C$ versus RMSE for the training set (Gamma = 5; $\varepsilon = 0.03$).

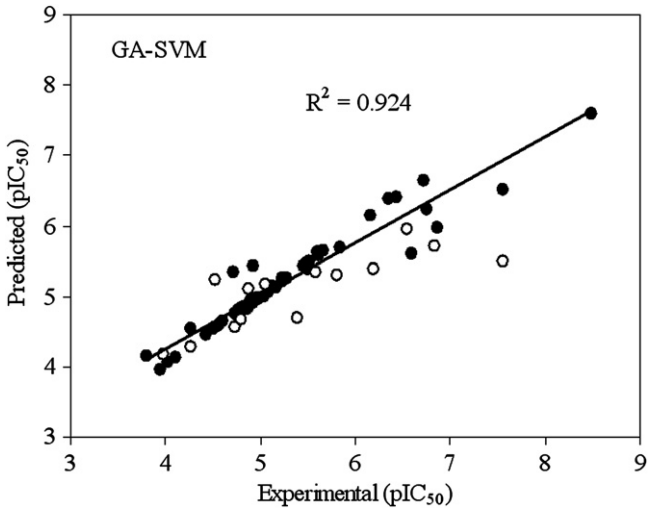


**Fig. 6.** Plot of the calculated values of $pIC_{50}$ from the GA-SVM model versus the experimental values of it for training (●) and test (○) sets.

overfit the training data. However, Ref. [37] indicated that prediction error was scarcely influenced by *C*. To make the learning process stable, a large value should be set up for *C* initially. The plot of RMSE versus *C* value is shown in Fig. 5 with values $\gamma = 5$, $\varepsilon = 0.03$. The optimal value of *C* was 40. Therefore, the best choices for $\gamma$, $\varepsilon$ and *C* were 5, 0.03 and 40. For the optimal model, the cross-validated coefficient $Q^2$ was 0.712. It gave RMSE of 0.295 for the training set, 0.795 for the test set, and the corresponding correlation coefficients ($R^2$) were 0.924 and 0.659, respectively. Then the optimized support vector regression could simulate the complicated nonlinear relationship between $pIC_{50}$ value and the descriptors.

The calculated $pIC_{50}$ values obtained from SVM predictive model are listed in Table 1. Fig. 6 shows the predicted versus experimental values of $pIC_{50}$ for the training and test sets with SVM method. Table 3 presents the statistical parameters of the results obtained from the two studied models for the same set of compounds. The RMSE of SVM model for the training, test and whole data sets were lower than those of models proposed in MLR method. The correlation coefficient ($R^2$) given by SVM was higher than of MLR. And the results of *F*-test were obtained and also are shown in Table 3. From the table, it can be seen that SVM model gives higher *F* values, so this model gives the most satisfactory results, compared with the results obtained from MLR method. It can be seen that although parameters appearing in the GA-MLR model are used as inputs for the generated GA-SVM model, the statistics has shown a large improvement. These improvements are due to the fact that BK-channels activity of compounds shows non-linear correlations with the selected descriptors.

## 4. Conclusion

In the present study, a linear method (GA-MLR) and a non-linear method (GA-SVM) were used to construct a quantitative relation between the $pIC_{50}$ values of BK-channel activators and their calculated descriptors. The results obtained by GA-SVM were compared with those obtained by GA-MLR which confirmed the superiority of the GA-SVM model as a more powerful method to predict the $pIC_{50}$. Since the improvement of the results obtained using non-linear model (GA-SVM) is considerable, it can be concluded that the non-linear characteristics of the descriptors on the $pIC_{50}$ values of the BK-channel activators. Distances between of the Oxygen and Chlorine atoms, the mass, the van der Waals volume, the electronegativity, and the polarizability of the molecules found to be important factors controlling the BK-channels activity. The proposed models can identify and provide an insight into some instructions for further designing of new BK-channel activators.

## References

[1] V. Calderone, Curr. Med. Chem. 9 (2002) 1385–1395.
[2] S.-N. Wu, Curr. Med. Chem. 10 (2003) 649–661.
[3] N.S. Cook, Potassium Channels: Structure, Classification, Function and Therapeutic Potential. Ellis Horwood Limited, Southampton, UK, 1990.
[4] D.L. Hill, The Biochemistry and Physiology of Tetrahymena, first ed. Academic Press, New York, 1972, p. 230.
[5] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Chem. Biol. Drug Des. 72 (2008) 575–584.
[6] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Chem. Biol. Drug Des. 74 (2009) 165–172.
[7] S. Riahi, E. Pourbasheer, M.R. Ganjali, P. Norouzi, Chem. Biol. Drug Des. 73 (2009) 558–571.
[8] S. Riahi, M.R. Ganjali, E. Pourbasheer, P. Norouzi, Chromatographia 67 (2008) 917–922.
[9] S. Riahi, E. Pourbasheer, M.R. Ganjali, P. Norouzi, A. Zeraatkar Moghaddam, J. Chil. Chem. Soc. 55 (2008) 1086–1093.
[10] A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, Monatsh. Chem. 139 (2008) 1423–1431.
[11] P.V. Khadikar, A. Phadnis, A. Shrivastava, Bioorg. Med. Chem. 10 (2002) 1181–1188.
[12] V.K. Agrawal, P.V. Khadikar, Bioorg. Med. Chem. 9 (2001) 3035–3040.
[13] S. Riahi, E. Pourbasheer, M.R. Ganjali, P. Norouzi, J. Hazard. Maternite 166 (2009) 853–859.
[14] A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, Monatsh. Chem. 140 (2009) 15–27.
[15] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, J. Chem. Inf. Comput. Sci. 43 (2003) 1882–1889.
[16] C. Niani, L. Wencong, Y. Jie, L. Gozheng, Support Vector Machine in Chemistry. World Scientific Publishing Co. Pet. Ltd, 2004.
[17] H.X. Liu, R.S. Zhang, F. Luan, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Chem. Inf. Comput. Sci. 43 (2003) 900–907.
[18] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Comput. Chem. 26 (2001) 5–14.
[19] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem, Inf. Comput. Sci. 43 (2003) 1288–1296.
[20] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Chem. Inf. Comput. Sci. 44 (2004) 161–167.
[21] S. Riahi, M.R. Ganjali, E. Pourbasheer, F. Divsar, P. Norouzi, M. Chaloosi, Curr. Pharmaceut. Anal. 4 (2008) 231–237.
[22] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Chem. Biol. Drug Des. 72 (2008) 205–216.
[23] S. Riahi, M.R. Ganjali, A.B. Moghaddam, E. Pourbasheer, P. Norouzi, Curr. Anal. Chem. 5 (2009) 42–47.
[24] A. Coi, F.L. Fiamingo, O. Livi, V. Calderone, A. Martelli, I. Massarelli, A.M. Bianucci, Bioorg. Med. Chem. 17 (2009) 319–325.
[25] V. Calderone, A. Coi, F.L. Fiamingo, I. Giorgi, M. Leonardi, O. Livi, A. Martelli, E. Martinotti, Eur. J. Med. Chem. 41 (2006) 1421–1429.
[26] HyperChem Release 7, HyperCube, Inc., [Online] available. http://www.hyper.com.
[27] R. Todeschini, Milano Chemometrics and QSPR Group [Online] available. http://www.disat.unimib.it/chm.
[28] R. Leardi, R. Boggia, M. Terrile, J. Chemometr. 6 (1992) 267–281.
[29] S.J. Cho, M.A. Hermsmeier, J. Chem. Inf. Comput. Sci. 42 (2002) 927–936.
[30] K. Baumann, H. Albert, M.V. Korff, J. Chemometr. 16 (2002) 339–350.
[31] Q. Lu, G. Shen, R. Yu, J. Comput. Chembiochem. 23 (2002) 1357–1365.
[32] S. Ahmad, M.M. Gromiha, J. Comput. Chem. 24 (2003) 1313–1320.
[33] U. Depczynski, V.J. Frost, K. Molt, Anal. Chim. Acta 420 (2000) 217–227.
[34] The Mathworks Inc, Genetic Algorithm and Direct Search Toolbox Users Guide Massachusetts (2002).
[35] C. Cortes, V. Vapnik, J. Mach. Learn Res. 20 (1995) 273–297.
[36] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000.
[37] T. Joachims, Learning to Classify Text using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer, 2002.
[38] B. Scholkopf, A. Smola, Learning with Kernels. MIT Press, Cambridge, MA, 2002.
[39] R. Herbrich, Learning Kernel Classifiers. MIT Press, Cambridge, MA, 2002.
[40] M. Tute, History and objectives of quantitative drug design in advances in drug research. in: P. Sammes, J. Taylor (Eds.), Comprehensive Medicinal Chemistry, vol. 4. Pergamon, Oxford, 1990, pp. 1–32.
[41] C. Hansch, J. Taylor, P. Sammes, Comprehensive Medicinal Chemistry: the Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds, vol. 6, Pergamon, New York, 1990, pp. 1–19.
[42] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, 2000.
[43] V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. 42 (2002) 682–692.
[44] K. Baumann, QSAR Comb. Scientometrics 24 (2005) 1033–1046.
[45] W.J. Wang, Z.B. Xu, W.Z. Lu, X.Y. Zhang, Neurocomputing 55 (2003) 643–663.