Analytica Chimica Acta 690 (2011) 53–63

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Prediction of $P2Y_{12}$ antagonists using a novel genetic algorithm-support vector machine coupled approach

Ming Hao[a], Yan Li[a,*], Yonghua Wang[b], Shuwei Zhang[a,*]

[a] School of Chemical Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China
[b] Center of Bioinformatics, Northwest A&F University, Yangling, Shaanxi 712100, China

ARTICLE INFO

Article history:
Received 11 November 2010
Received in revised form 26 January 2011
Accepted 1 February 2011
Available online 28 February 2011

Keywords:
$P2Y_{12}$
Genetic algorithm
Support vector machine
Descriptor selection

ABSTRACT

Presently, a genetic algorithm (GA)-support vector machine (SVM) coupled approach is proposed for optimizing the 2D molecular descriptor subset generated for series of $P2Y_{12}$ (members of the G-protein-coupled receptor family) antagonists, with the statistical performance and efficiency of the model being simultaneously enhanced by SVM kernel-based nonlinear projection. As we know, this is the first QSAR study for prediction of $P2Y_{12}$ inhibition activity based on an unusually large dataset of 364 $P2Y_{12}$ antagonists with diversity of structures. In addition, three other widely used approaches, i.e., partial least squares (PLS), random forest (RF), and Gaussian process (GP) routines combined with GA (namely, GA–PLS, GA–RF, GA–GP, respectively) are also employed and compared with the GA–SVM method in terms of several rigorous evaluation criteria. The obtained results indicate that the GA–SVM model is a powerful tool for prediction of $P2Y_{12}$ antagonists, producing a conventional correlation coefficient R^2 of 0.976 and R^2_{cv} (cross-validation) of 0.829 for the training set as well as R^2_{pred} of 0.811 for the test set, which significantly outperforms the other three methods with the average $R^2 = 0.894$, $R^2_{cv} = 0.741$, $R^2_{pred} = 0.693$. The proposed model with excellent prediction capacity from both the internal to external quality should be helpful for screening and optimization of potential $P2Y_{12}$ antagonists prior to chemical synthesis in drug development.

1. Introduction

Nowadays, the most common causes of mortality from cardiovascular and cerebrovascular diseases are endangering human health all over the world. Researchers are trying their best to develop new drugs against these stubborn diseases, the current therapy for which includes antiplatelet agents such as aspirin, dipyridamole, glycoprotein IIb/IIIa antagonists, and thienopyridines [1]. Clopidogrel, a thienopyridine, is an oral prescription antiplatelet drug approved for the reduction of atherosclerotic events (stroke, myocardial infarction, and death) in patients with acute coronary syndrome that acts by blocking the adenosine diphosphate (ADP)-stimulated platelet aggregation. As an important platelet agonist, ADP, induces a primary aggregation response and contributes to the secondary aggregation following release from platelet dense granules upon the activation by other agonists [2]. ADP induces platelet aggregation via the activation of two major ADP receptors, $P2Y_1$ and $P2Y_{12}$ [3], both members of the G-protein-coupled receptor family [4,5]. Experimental studies have demonstrated that a selective blockade of either receptor is sufficient to inhibit the platelet activation [6]. However, $P2Y_{12}$ represents a more attractive therapeutic target for the selective modulation of ADP-induced platelet activation compared to $P2Y_1$, because $P2Y_1$ gene is ubiquitously expressed, whereas $P2Y_{12}$ is primarily a platelet specific receptor.

Current drugs against $P2Y_{12}$ receptor [7], however, suffer some limitations. Take Clopidogrel as an example, the active metabolite of the prodrug can irreversibly and selectively inhibit $P2Y_{12}$ receptor, leading to a delay for the antiplatelet efficacy for several days [8]. This makes it less effective in acute settings and difficult to manage if a patient bleeds, experiences a trauma, or requires emergency surgery. In addition, there are also some other unavoidable shortcomings which can affect the efficiency of this drug, for example, some individuals do not metabolize the prodrug adequately and some might be resistant to clopidogrel [9]. It is anticipated that a direct acting, reversible $P2Y_{12}$ antagonist will achieve an improvement in efficacy and also exhibit an improved safety profile. Several groups have put their efforts aimed at discovering ADP receptor antagonists, including cangrelor (AR-C69931MX) and ticagrelor (AZD-6140) [10], both of which are currently in late stage clinical trials. Recently, John et al. have also reported an unusually large dataset of structural diversity of $P2Y_{12}$ antagonists to address the unmet medical need for safe and effective oral antiplatelet agents

[*] Corresponding authors. Tel.: +86 411 84986062.
E-mail addresses: yanli@dlut.edu.cn (Y. Li), zswei@chem.dlut.edu.cn (S. Zhang).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.02.004

with good human platelet rich plasma potency, selectivity, in vivo efficacy and oral bioavailability.

As we know, novel medicines are typically developed using a trial and error approach which is normally costly and time-consuming. The application of in silico methods such as quantitative structure-activity relationship (QSAR) to this issue has the potential to decrease substantially the time and effort required to discover new drugs or improve current ones in terms of the efficacy [11,12]. Therefore, in silico approaches have been successfully applied to various fields of biochemistry such as the prediction of chromosome aberrations [13], blood–brain barrier-penetrating agents [14], P-glycoprotein substrates and inhibitors [15], etc. However, there is still, to our best knowledge, no report of any computational models for prediction of the $P2Y_{12}$ inhibition activity up to the present. Therefore, it is necessary to build a predictive QSAR model to fill this gap.

Among the QSAR investigations, one of the important factors affecting the quality of the model is the molecular descriptors used to extract the structural information that is suitable for the model development. The software $Mold^2$ [16] enables a rapid calculation of a large and diverse set of descriptors encoding two-dimensional (2D) chemical structure information. By comparative analysis of $Mold^2$ descriptors with those calculated by $Cerius^2$, Dragon or MolconnZ on several data sets it has been demonstrated that $Mold^2$ descriptors convey a similar amount of information as these widely used software packages [16]. Although serving as free available software, $Mold^2$ has been proved suitable not only for the QSAR research [17], but also for virtual screening of large databases in drug development [16].

The selection of appropriate approaches to building the models is another key factor to produce an accuracy prediction. Often used statistical methods include the simple but interpretable multiple linear regression (MLR) [18], PLS [19] and nonlinear, relatively not prone to interpretable but often highly predictive methods such as artificial neural networks (ANN) [20] and recently popular SVM, RF, GP and so forth [21–24], which is just the case in this work. All of these methods have a proven record of successful applications in QSAR modeling. However, several of them also often suffer several limitations. For example, traditional statistical method like MLR can only handle data sets where the number of descriptors (p) is smaller than that of the molecules (n), unless a pre-selection of the descriptors is executed (e.g. by using successive projections or genetic algorithms [25,26], etc.). Also they are not flexible enough and do not account for nonlinear behavior [21]. SVM, a relatively new nonlinear technique employed in classification and regression problems [27], is not robust to the presence of a large number of irrelevant descriptors [21]. PLS is a popular computational method that expresses a dependent variable in terms of linear combinations of the independent variables commonly known as principal components. However, PLS may not be suitable for handling multiple mechanisms of action [21], such as the nonlinear biological behaviors.

Thus, in the present work for purpose of obtaining mathematic models with highly statistical performance and efficiency, GA is selected as the feature selection method to obtain the optimal descriptor subset when dealing with a number of descriptors. As a state-of-art algorithm that has found success in a variety of areas, SVM, has been used as a classifier in the wrapper feature selection method [28]. Among the many wrapper algorithms used, the GA, which solves optimization problems using the evolution with appropriate crossover and mutation operators, has been proven as promising one due to its prominent capability in solving global optimization problems. However, existing GA-based wrapper was primarily developed for optimizing either the feature subset selection [29,30] or optimizing the parameters of various algorithms [31]. However, in SVM regression analysis, a key problem is that the model efficiency is not only largely dependent on the feature subset, but also on the kernel parameters needed to be optimized simultaneously in the model generation process [32]. In light of this, for the first time in this work, we have adopted GA to optimize the SVM parameters (including not only the C and the kernel function γ, but also the variable ε for the radial basis function kernel) and descriptor subset simultaneously to build a reliable model. In addition, for comparison with the GA–SVM, three other popular methods, i.e., GA–PLS, GA–RF and GA–GP, are also applied using the same dataset.

2. Material and experimental methods

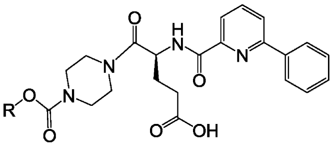
2.1. Data sets

A large, diverse dataset of 364 antagonists of $P2Y_{12}$ with definitive biological values, were collected from literatures [33–36] published by the same research group. The experimental IC_{50} values of all the molecules were from human platelet rich plasma incubated with 20 μM ADP. Here, the converted molar pIC_{50} ($-\log IC_{50}$) values, ranging from 4.013 to 6.678 M, were used as the dependent variables in the QSAR regression analysis to improve the normal distribution of the experimental data points. As to the selection of training/test sets which plays a crucial role in the QSAR modeling, two split formats of approaches including (1) the single pair of training/test sets [37–40] which is usually based on the descriptors space such as Kennard–Stone (KS) algorithm and Kohonen self-organizing mapping, and (2) the repeated splitting of data to training/test sets [21,41]) both have a good record of successful applications in QSAR modeling. Thus in this present work, the whole data set was divided into training (291 compounds) and test (73 compounds) sets in a ratio of 4:1 based on Kennard–Stone algorithm, which guarantees that the points of the training set are distributed evenly within the whole area occupied by representative points, and the closeness condition of the test set points to the training ones is satisfied [42]. Since KS has been reported [43] to be superior to both the random sampling (RS) and Kohonen self-organizing mapping [44,45], it has also been successfully applied to many QSAR researches [39,43,46]. Table 1 shows several representative compounds together with their activity. All information about the 364 compounds with their diverse scaffolds of structures is provided in Table S1 (Supporting information).

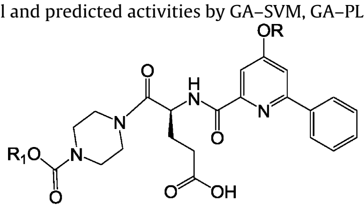
2.2. Descriptors calculation and pre-processing

Construction of the 2D prediction models firstly depends on the generation of molecular descriptors. By simply using various molecular modeling tools, it is possible to calculate thousands of these descriptors directly from the structure of any particular molecule. In the present work, all two-dimensional structures of the dataset were built with the ISIS/Draw 2.3 program [47], and converted SDF format by Open Babel software package (http://openbabel.sourceforge.net/). The final structures were transferred into $Mold^2$ [16], a free program available to public, to calculate molecular descriptors. The $Mold^2$ software package can calculate 777 molecular descriptors solely from 2D chemical structures. Hong et al. have reported that the models generated using $Mold^2$ descriptors were comparable to those generated using descriptors from the compared commercial software packages. In our work, all original 777 molecular descriptors were calculated, which were then preprocessed (also called unsupervised selected) as follows: (1) descriptors containing greater than 85% zero values were removed; (2) zero- and near zero- variance predictors were removed because such descriptors may cause the model to crash or the fit to be unstable; and (3) one of the two descriptors

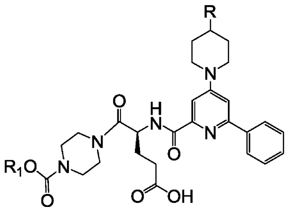
Table 1
Representative chemical structures, actual and predicted activities by GA–SVM, GA–PLS, GA–RF and GA–GP models based on 364 $P2Y_{12}$ antagonists of diverse structures.



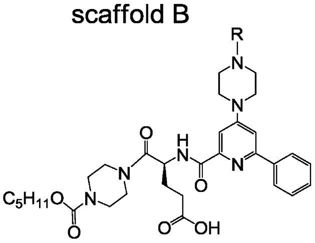
scaffold A



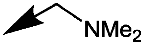
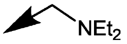
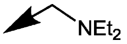
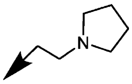
scaffold B



scaffold C



scaffold D

Compd.	Scaffold	Substitute		pIC50 (M)	GA–SVM	GA–PLS	GA–RF	GA–GP	Ref.
		R_1	R						
1	A	–	$(CH_2)_3CF_3$	4.155	4.207	4.206	4.413	4.645	[33]
2	A	–	CH_2cPent	4.086	4.138	3.939	4.208	4.701	[33]
3	A	–	$(CH_2)_2CH(CH_3)_2$	4.056	4.108	4.309	4.232	4.580	[33]
191[a]	B	Bu	$O(CH_2)_2OH$	5.108	5.181	5.040	5.066	5.150	[34]
192	B	Pent	$O(CH_2)_2OH$	4.939	4.991	4.843	4.988	5.020	[34]
193	B	Et	$O(CH_2)_2OMe$	5.237	5.199	4.872	5.170	4.914	[34]
194	B	Allyl	$O(CH_2)_2OMe$	4.876	4.928	4.764	4.929	4.940	[34]
196[a]	B	Bu	$O(CH_2)_2OMe$	5.357	5.487	5.118	5.306	5.162	[34]
297[a]	C	Pent		6.000	5.989	5.806	5.805	6.016	[36]
298	C	Bu		5.620	5.672	5.704	5.694	5.788	[36]
299	C	Pent		5.658	5.751	5.719	5.692	5.758	[36]
305[a]	C	Pent		6.125	6.188	6.107	6.023	5.905	[36]
323[a]	D	–	H	5.721	5.487	5.284	5.642	5.534	[36]
324	D	–	Me	5.638	5.593	5.393	5.656	5.572	[36]
325	D	–	Pr	5.301	5.447	5.366	5.499	5.488	[36]
326[a]	D	–	$(CH_2)_2OH$	5.745	5.657	5.322	5.677	5.680	[36]
329[a]	D	–	$(CH_2)_2NH_2$	5.959	6.210	6.047	5.833	6.061	[36]
330	D	–	$(CH_2)_2NMe_2$	6.292	6.240	5.943	6.128	6.080	[36]

[a] Test set.

that have the absolute correlations above 0.75 was omitted. After these steps, the number of the descriptors was reduced to 106 for further research. In addition, in each case, descriptors were scaled according to the following formula:

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}} \quad (1)$$

where X_{ij} and X_{ij}^{n} are the non-scaled and scaled jth descriptor values for compound i, respectively, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for jth descriptor, respectively. Thus, for all descriptors, $\min(X_{ij}^{n}) = 0$ and $\max(X_{ij}^{n}) = 1$.

2.3. GA–SVM

Genetic algorithm, derived from Darwin's theory of natural selection and evolution, is a highly efficient optimization algorithm which has already been successfully applied in many QSAR analyses [48,49]. GA works with a set of candidate solutions called a population. Based on the Darwinian principle of survival of the fittest, GA obtains the optimal solution after a series of iterative computations (i.e., selection or reproduction, crossover or recombination, and mutation). For variable selection issue, the binary coding form of each chromosome is adopted with 1 and 0 representing selected and non-selected descriptors, respectively. Crossover, the critical genetic operator that allows new solution regions in the search space to be explored, is a random mechanism for exchanging genes between two chromosomes using different crossover strategies [32,50,51]. The simplest one point crossover was employed in our study. In mutation the genes may occasionally be altered, i.e., in binary code genes the code may be changed from 0 to 1 or vice versa. Children replace the old population using the elitism or diversity replacement strategy and form a new population in the next generation. The evolutionary process operates many generations until the termination condition is satisfied [32,52,53]. The detailed methodology about GA has been described everywhere [53,54].

Support vector machines are a relatively new type of learning algorithm originally introduced by Vapnik and co-workers [55]. Because of many attractive features and promising empirical performances, it is gaining increasing popularity in many fields including QSAR analysis [56]. Although SVM is initially developed for binary classification, it has been extended to solve regression problems with the given data set $D = \{(x_i, y_i)\}_{i=1}^{N}$ obtained from a latent function by the introduction of ε-insensitive loss function, where x_i means the sample vector, y_i the corresponding response,

C			γ			ε			f		
C	$\ldots C_i \ldots$	C_{nc}	γ_1	$\ldots \gamma_j \ldots$	$\gamma_{n\gamma}$	ε_1	$\ldots \varepsilon_k \ldots$	$\varepsilon_{n\varepsilon}$	f_1	$\ldots f_m \ldots$	f_{nf}

Fig. 1. The chromosome representation (including four parts of C, γ, ε and the features mask).

and N the number of samples. Since for the present work, the main aim employing SVM is to predict $P2Y_{12}$ inhibition activity which belongs to the regression issue, the support vector regression (SVR) is simply introduced with the standard form of SVR as follows [57]:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$
$$\text{s.t.} \begin{cases} w^T\phi(x_i) + b - z_i \le \varepsilon + \xi_i \\ z_i - w^T\phi(x_i) - b \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \quad i = 1, \ldots, N \end{cases} \tag{2}$$

The dual is:

$$\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha_i - \alpha_i^*)^T Q(\alpha - \alpha_i^*) + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} z_i(\alpha_i - \alpha_i^*)$$
$$\text{s.t.} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0, \quad 0 \le \alpha_i, \quad \alpha_i^* \le C, \quad i = 1, \ldots, N \tag{3}$$

where $Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$.

For SVR, the input x is mapped into a higher dimensional feature space by the use of a kernel function (often used in SVM including linear, polynomial, radial basis function, and sigmoid function), and then a linear model given in Eq. (4) is constructed in this feature space:

$$f(x, \omega) = \sum_{j=1}^{m} \omega_j g_j(x) + b \tag{4}$$

where $g_j(x), j = 1,\ldots, m$ represents a set of nonlinear transformations, ω_j and b are the coefficient and bias terms, respectively.

The generalization performance of SVR depends on a good setting of parameters: C, ε, the kernel type and corresponding kernel parameters. The selection of the kernel function and corresponding parameters is very important because they define the distribution of the training set samples in the high dimensional feature space. There are four possible choices of kernel functions available in normal algorithms [58], i.e., linear, polynomial, radial basis function, and sigmoid function. For regression tasks, the radial basis function kernel is often used because of its effectiveness and speed in training process. It was also used for all SVR models in our study. For the RBF kernel, the most important parameter is the width γ of the radial basis function. The previous references [59] have illustrated that C is a regularization parameter that 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 overfit the training data. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. The value of ε can affect the number of support vectors used to construct the regression function, where the bigger ε is, the fewer support vectors are selected. In the present work, a genetic algorithm approach was employed to simultaneously optimize the parameters (including the C, width γ and ε) and feature subset.

Since RBF has been selected as a kernel function of SVR, its three parameters (C, γ and ε) and the descriptors used as input must be optimized using our proposed GA-based system. The chromosome comprises four parts, C, γ, ε and the features mask as described formation [32] previously. The binary coding system was used to represent the chromosome in this work. Fig. 1 shows the binary chromosome representation in our design process, where the definitions of the parameters are: C_i represents the ith bit's value of bit string that represents parameter C, and n_c is the number of bits representing parameter C; γ_j represents the jth bit's value of bit string that represents parameter γ, and n_γ is the number of bits representing parameter γ; ε_k represents the kth bit's value of bit string that represents parameter ε, and n_ε is the number of bits representing parameter ε; f_m represents the mask value of mth feature, and n_f is the number of bits representing the selected features, respectively. n_c, n_γ and n_f can be modified according to the calculation precision and/or efficiency required ($n_c = 20$, $n_\gamma = 20$, $n_\varepsilon = 20$ and n_f is the total 106 number of descriptors in the present work).

According to Eq. (5), the bit strings representing the genotype of parameters C, γ and ε in Fig. 1 were transformed into their phenotypes. Note that the precision of representing parameter depends on the length of the bit string (n_c, n_γ and n_ε), and the minimum and maximum values of the parameter are determined by the designer. For chromosome representing the feature mask, the bit with value '1' represents that the feature is selected, and '0' indicates that the feature is not selected.

$$p = \min_p + \frac{\max_p - \min_p}{2^l - 1} \times d \tag{5}$$

where p is the phenotype of the bit string; $\min_p$ and $\max_p$ are the minimum and maximum values of the parameter, respectively; d represents the decimal value of bit string and l is the length of the bit string.

To evaluate whether an individual is fit to survive, fitness function is needed in the GA. In the GA–SVM model, we used two criteria, namely mean squared error based on 10-fold cross-validation (MSECV) and the number of selected features, to design the fitness function. The principle is that individuals with low MSECV and small number of features have a high fitness value, and thus a high probability to be passed to the next generation. A single objective fitness function that combines the two goals into one was designed to solve the multiple criteria problem [32], with a formula as below:

$$\text{fitness} = w_\text{a} \times \text{MSECV} + w_\text{b} \times \sum_{i=1}^{n_f} f_i \tag{6}$$

where w_a represents the weight value for MSECV, w_b for the number of features respectively. f_i is the mask value of the ith feature where '1' represents that feature i is selected and '0' represents that feature i is not selected. In this equation, w_a can be adjusted to 100% if MSECV is the most important and generally, w_a can be set from 75% to 100% according to user's requirements. In our study, we set w_a to 100% for the purpose of high predictive ability after we systematically changed the weight values.

Fig. 2 depicts the design of the proposed GA–SVM approach whose detailed explanation is as follows: (1) converting genotype to phenotype. Each parameter and descriptor chromosome was converted from its genotype into a phenotype; (2) feature subset. After the genetic operation and the converting of each feature

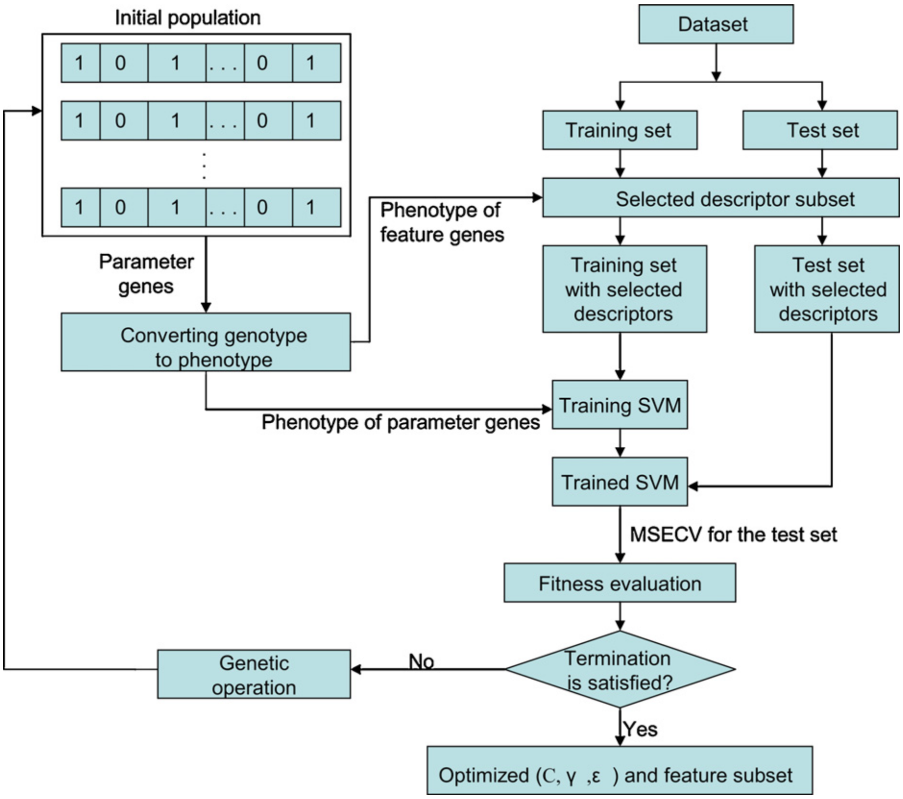

Initial population
1 0 1 . . . 0 1
1 0 1 . . . 0 1
1 0 1 . . . 0 1
Parameter genes
Converting genotype to phenotype
Phenotype of feature genes
Phenotype of parameter genes
Dataset
Training set
Test set
Selected descriptor subset
Training set with selected descriptors
Test set with selected descriptors
Training SVM
Trained SVM
MSECV for the test set
Fitness evaluation
Termination is satisfied?
No
Genetic operation
Yes
Optimized (C, γ ,ε) and feature subset

Fig. 2. System architecture of the proposed GA–SVM.

subset chromosome from the genotype into phenotype, a feature subset can be determined; (3) fitness evaluation. For each chromosome representing C, γ, ε and selected features, training dataset is used to train the SVM, while the testing dataset is used to calculate MSECV. When the MSECV is obtained, each chromosome is evaluated by the fitness function described in Eq. (6); (4) termination criteria. When the termination criteria are satisfied, the process ends; otherwise, we proceed with the next generation; and (5) genetic operation. In this step, the system searches for better solutions by genetic operations, including the selection, crossover, mutation, and replacement [32]. After these steps, the optimal parameters (C, γ, ε) and descriptor subset determined will also be served as input to the other three statistical methods for the further investigation and comparison.

2.4. GA–PLS

PLS is similar to principal components regression but with both the independent and dependent variables involved in the generation of the orthogonal latent variables rather than only independent variables used. PLS is based on the projection of the original multivariate data matrices down onto smaller matrices (T, U) with orthogonal columns, which relates the information in the response matrix Y to the systematic variance in the descriptor matrix X, as shown below:

$$X = \bar{X} + TP' + E \quad (7)$$

$$Y = \bar{Y} + UC' + F \quad (8)$$

$$U = T + H \quad (9)$$

where $\overline{X}$ and $\overline{Y}$ are the corresponding mean value matrices, T and U are the matrices of scores that summarize the x and y variables respectively, P is the matrix of loadings showing the influence of the x variables in each component, C is the matrix of weights expressing the correlation between Y and $T(X)$, E, F, and H are the corresponding residuals matrices, respectively. PLS calculations also give an auxiliary matrix (PLS weights), which expresses the correlation between U and X and is used to calculate the T [60]. Determination of the significant number of model dimensions was made by cross-validation [61].

Up to date, PLS regression algorithms have been extended to various methods such as the kernel algorithm, the wide kernel algorithm, SIMPLS algorithm and the classical orthogonal scores algorithm in the R package pls [62]. In the present study, the kernel algorithm was selected to build the QSAR models, and the optimal principal component with the lowest root-mean-squared-error (RMSE) according to the following equation was selected based on 10-fold cross validation for further analysis.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (10)$$

where y_i is the actual output and $\hat{y}_i$ is the predicted output of the model, and n is the number of compounds in the analyzed set, respectively.

To reduce noise and enhance prediction performance, in the present work, a simple combination use of PLS method with genetic algorithm was also carried out to select the optimal descriptors for building PLS regression model. Here, the minimum MSE based on the 10-fold cross-validation was adopted as the fitness function. And the GA parameters were set as follows: the number of individual (NIND) of 50, the max number of generation to evolve (MAXGEN) of 200, the gap between the two generation (GGAP) of 0.9, and the probability of crossover (P_c) of 0.7 and others were set to default values in the GA toolbox.

2.5. GA–RF

RF models were constructed according to the described original RF algorithm [63]. RF is an ensemble of single decision trees, which ensemble produces a corresponding number of outputs and the outputs of all trees are aggregated to obtain one final prediction. The training algorithm of RF for regression can be briefly summarized as follows: (1) draw N bootstrap samples from the original training set; (2) construct an unpruned tree T_p ($p = 1,\ldots, N$) with each training set B_p. At each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and then choose the best split from among those variables. The tree is grown to the maximum size and not pruned back; and (3) predict the N trees by average for regression. RF algorithm is the same as Bagging when $m_{\text{try}} = p$ and the tree growing algorithm used in RF is CART. RF algorithm is efficient especially when the number of descriptors (p) is very large, with the reason that RF only tests the m_{try} of the descriptors rather than the p, where the default m_{try} is one-third of the number of descriptors (p) for regression. Thus, m_{try} is very small so that the search is very fast. In addition, RF is more efficient than a single tree deriving from that RF does not do any pruning at all, while a single tree needs some pruning using cross validation that can take up a significant portion of the computation time to get the right model complexity.

RF possesses its own reliable statistical characteristics based on OOB set prediction, which could be used for validation and model selection with no cross-validation performed. It was shown that the prediction accuracy of an OOB set and a 5-fold cross validation procedure was near the same [21]. Although RF performs relatively well “off the shelf” without expending much effort on the parameter tuning or variable selection [21], it is also of importance for carrying out some tentative investigations on the changes of m_{try} or descriptor selection to optimize the performance of RF.

Random forest, as a new classification and regression tool, has not been frequently applied in QSAR, QSPR (quantitative structure–property relationship). Thus it is necessary to investigate whether RF can obtain better statistical performance for the current dataset. Here, we just present a brief introduction about RF, for more details please see the corresponding important literatures [21,63]. It has been reported that RF can show excellent performance even when most predictive variables are noise, and be used when the number of variables is much larger than the number of observations, and returns measures of variable importance [21,63]. However, for approaching an ideal regression model (with high prediction accuracy by using less number of descriptors), a variable selection process is still required. To achieve the above object, in this work, the GA variable selection method using MSE based on out-of-bag of RF as the fitness function was carried out to achieve regression task for the current P2Y$_{12}$ antagonists. The GA parameters were set to be the same as those of PLS. And performance measures of the RF model, presently, were employed using the R package randomForest [64].

2.6. GA–GP

Preliminarily used in QSAR field, the Gaussian process (GP) [65], in the present study, was also introduced to predict the P2Y$_{12}$ antagonists activity. Pioneering works were made by Burden [66] who demonstrated that GP can be applied in the QSAR modeling of data sets of compounds active at the benzodiazepine and muscarinic receptors, etc. In addition, researchers [67–70] also reported satisfactory statistical prediction performances of GP on a series of pharmacokinetic properties. Recently, GP was adopted not only for implementing the automatic QSAR modeling of ADME properties [71], but also for executing multivariate spectroscopic calibration [72]. All these works confirmed the feasibility of GP's application on QSAR studies as a promising machine learning tool. In view of this, GP is also introduced in the QSAR modeling of P2Y$_{12}$ inhibition presently.

Being defined simply as a collection of random variables which have a joint Gaussian distribution, a Gaussian process is completely characterized by its mean and covariance function. One usually considers the mean function to be zero everywhere. The covariance function defines nearness or similarity between the values of targets (predictions) at two input points. More details can be found in [65].

In the current work, to determine the best variables for building the GP model, the Netlab toolbox was used to perform the Gaussian process combined with the genetic algorithm (http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/). Here, the squared exponential was selected as covariance function, and other parameters were set by default. During this process, it should be pointed out that for comparison the same GA parameters set as those used in PLS and RF modeling were adopted. Besides, the MSE based on 10-fold cross-validation was employed as the fitness function. After the determination of the optimal descriptors, we carried out further calculation of GP using the R package kernlab [73].

2.7. Evaluation of the statistical performance

The statistical performances of the constructed models are usually evaluated by several critical parameters like the square of correlation coefficient (R^2), the RMSE described above, and the 10-fold cross-validated R^2_{cv} [74,75]. In case of the external validation, the predictive capacity of the model we established was judged by its application for prediction of test set activity values, the predictive R^2 (R^2_{pred}) value of which was calculated as follows:

$$R^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{\text{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{\text{test}}(y_i - \bar{y}_{\text{tr}})^2} \qquad (11)$$

where y_i and $\hat{y}_i$ are the measured and predicted values of the dependent variable (over the test set), respectively, and $\bar{y}_{\text{tr}}$ is the averaged value of the dependent variable for the training set and the summations run over all compounds in the test set.

3. Results and discussion

3.1. Set parameters of GA–SVM method

These parameters include the range of kernel parameter C, γ, ε, NIND, MAXGEN, GGAP, P_c and the probabilities of mutation (P_m), which are set as follows: the minimum $C = 0$, maximum $C = 200$, minimum $\gamma = 0$, maximum $\gamma = 1000$, minimum $\varepsilon = 0.001$, maximum $\varepsilon = 1$, NIND = 50, MAXGEN = 200, GGAP = 0.9, P_c is set to 0.7 and P_m set to default value included in the genetic algorithm tool box.

3.2. Effect of parameter optimization and selected descriptors

The prediction effect of SVM regression systems using the GA method is evaluated by means of 10-fold cross-validation method. The optimal parameters (C, γ, ε) and proper descriptors are determined when the mean-square-root (MSE) reaches the lowest value in terms of 10-fold cross-validation. Thus, for this dataset, the parameter optimization ended in a lowest MSE of 0.063, the C, γ and ε fitted to 7.191, 0.183 and 0.052, respectively. Table S2 shows the final 41 descriptors selected and their corresponding definitions.

3.3. Explanation of the descriptors

Presently, a GA–SVM model was developed successfully, with the final number of descriptors reduced to 41 (Table S2) from the original 106 ones for further study. Here it should be pointed out that since in most QSAR researches a full direct explanation for all the descriptors involved in the related model is difficult, where most similar reported works all gave few detailed analysis of the descriptors involved in their model development, only a few descriptors in this work were explained. According to the suggestion [76] that the number of compounds should be at least 5 times larger than that of the selected independent variables, the models we developed all maintain the recommended ratio. By analyzing these 41 descriptors, some interesting information about $P2Y_{12}$-antagonist interaction is inferred as follows.

First of all, in all selected 41 descriptors, the electronic factors are proved to be an important kind of interactions for the binding of $P2Y_{12}$ with its antagonists. In the present model there are totally 18 descriptors (D464, D467, D472, D475, D478, D499, D500, D504, D508, D509, D525, D529, D551, D552, D559, D585, D590, and D592) that represent certain electronic relevant features possessing about 43% of all selected descriptors, which demonstrates clearly the crucial role of electronic factors playing in describing the inhibition activity of the antagonists. Among these, the descriptors including D464, D467, D472, D475, D478, D499, D500, D504, D508 and D509 all belong to 2D autocorrelation descriptors [77], which represent the topological structure of compounds, but are more complex in nature when compared to the classical topological descriptors. The computation of these descriptors involves the summations of different autocorrelation functions corresponding to different structural lags and leads to different autocorrelation vectors corresponding to the lengths of substructural fragments. Keeping in mind of this aspect, the interpretation of the two-dimensional autocorrelation descriptors is uneasy. Basically, the pool of 2D autocorrelation descriptors defines a wide 2D space. Herein, on behalf of a greater applicability, the physicochemical properties (atomic Sanderson electronegativities and atomic polarizabilities here) are inserted as weighting components. In addition, D525 and D529 refer to the mean molecular charge indices of order-5 and order-9, respectively. The remaining descriptors (i.e., D551, D552, D559, D585, D590, and D592) are BCUT variables, which are the eigenvalues of modified connectivity matrix, the Burden matrix [78,79]. In fact, the BCUT metrics have been successfully applied to several QSAR studies [50,80,81]. In summary, all these descriptors depicted above represent certain electronic relevant information.

Furthermore, several atom-centred fragment and functional group counts descriptors (D715, D719, D732, D744, D745, D746, D651 and D712) may also be correlated with the H-bond formation process. Third, the hydrophobic or hydrophilic effect is also proven to be important for the inhibition activity of the $P2Y_{12}$ antagonists. The developed model selects the descriptor relevant to $\log P$ (D777), which indicates that it is the hydrophobic effect of the ligand that dominates its inhibition activity. Whereas at the same time, the descriptor D775 (hydrophilic factor) which regulates the hydrophilicity of the molecules, is also selected by our model, suggesting that an enhancement of the inhibitory activity may be achieved by substitution with more hydrophilic substituents, while as a complete unit of molecule the antagonist should also hold certain hydrophobic features like the ring-based structures. Finally, it is worthwhile to mention that some other descriptors are also useful in the prediction of $P2Y_{12}$ inhibition potency which can offer us some important information. For example, D142, one of the topological descriptors, appears in the model as a Balaban type of mean square vertex distance index. It relates to the molecular branching in an isomeric series, which index decreases with the increase of the molecular branching [27]. Several other descriptors are found alone may not be very conspicuous, but are valuable in the prediction of $P2Y_{12}$ inhibition activity while coupling with other factors.

In summary, it is concluded that the electronic factors, hydrophobic and H-bond interaction play a central role in the $P2Y_{12}$ inhibition of the studied antagonists. Partially, our results are supported by some references [33–36]. For example, the report [33] indicated the importance of H-bond donors for the inhibition activity and we find that the corresponding descriptor D712 (number of group donor atoms for H-bonds (with N and O)) also plays a part in $P2Y_{12}$ inhibition.

As expectation, an ideal QSAR model would be robust, sparse, predictive, and interpretable. In many cases, however, such ideal is not easy to achieve with current descriptors and response variable mapping methods, though much effort is being expended. Consequently, QSAR modeling tends to be divided into two classes depending on the intended outcome of the study. Predictive QSAR aims to screen large, chemically diverse compound libraries that are often noisy, thus they often present less descriptors explanation, especially, with various descriptors, which is just the case here. In addition, in case of possible multiple mechanisms of action among the molecules, nonlinear machine learning algorithms are sometimes employed (like SVM) with purpose of making the corresponding models they built be as potent predictive as possible so that new candidates can be assessed prior to synthesis or large databases and virtual libraries be screened for hits, which in turn makes the model interpretation much harder. Interpretative modeling often uses linear simulation tools (like MLR), chemically relevant and interpretable descriptors, and smaller, more congeneric data sets that have usually been measured to a higher degree of accuracy. As a result, it is still a difficult task to produce an as well highly predictive as easily interpretable model. Thus, in terms of developing a highly predictive model, the proposed GA–SVM model in this work could implement this task.

3.4. Performance of different statistical methods

After finishing all above work, four different statistical methods (GA–SVM, GA–PLS, GA–RF and GA–GP) are applied on the dataset and their performances are compared with detailed statistics summarized in Table 2. The representative predicted $P2Y_{12}$ inhibitory activities by these models are shown in Table 1, with the full predicted results listed in Table S1.

3.4.1. GA–SVM

Based on the determined optimal parameters by GA, the SVM model presents an RMSE of 0.133 and 0.209 for the training and test sets, respectively. The determined coefficient R^2 reaches as high as 0.976 with $R^2_{cv} = 0.829$ for the training set. The model predictability is evaluated by an external prediction set, which illustrates R^2_{ts} and R^2_{pred} values of 0.806 and 0.811, respectively. The experimental versus predicted $P2Y_{12}$ inhibition activity based on the SVM model is shown in Fig. 3A.

3.4.2. GA–PLS

To investigate whether there is a linear relationship existing between the descriptors and $P2Y_{12}$ inhibition, the widely used PLS approach is also applied in the present work. After GA–PLS, fifty-seven descriptors are determined and used for further calculation. Based on the lowest 10-fold cross-validation RMSE (0.311), a 27-latent variable QSAR model is obtained. The statistical results of the PLS model present a coefficient of determination $R^2 = 0.839$, $R^2_{cv} = 0.737$ and RMSE = 0.243 for the training set, respectively. The model was also evaluated on unseen chemicals, i.e., the test data, resulted in $R^2_{ts} = 0.594$, $R^2{}_{pred} = 0.577$ and RMSE = 0.312 for the test set, respectively. Fig. 3B presents a visual investigation of the PLS

Table 2
Performance comparison between GA–SVM, GA–PLS, GA–RF and GA–GP models.

Model	Calibration set			Prediction set		
	R^2	R^2_{cv}	RMSE	R^2_{ts}	R^2_{pred}	RMSE
GA–SVM	0.976	0.829	0.133	0.806	0.811	0.209
GA–PLS	0.839	0.737	0.243	0.594	0.577	0.312
GA–RF	0.966	0.750	0.128	0.745	0.739	0.245
GA–GP	0.877	0.737	0.234	0.783	0.764	0.234

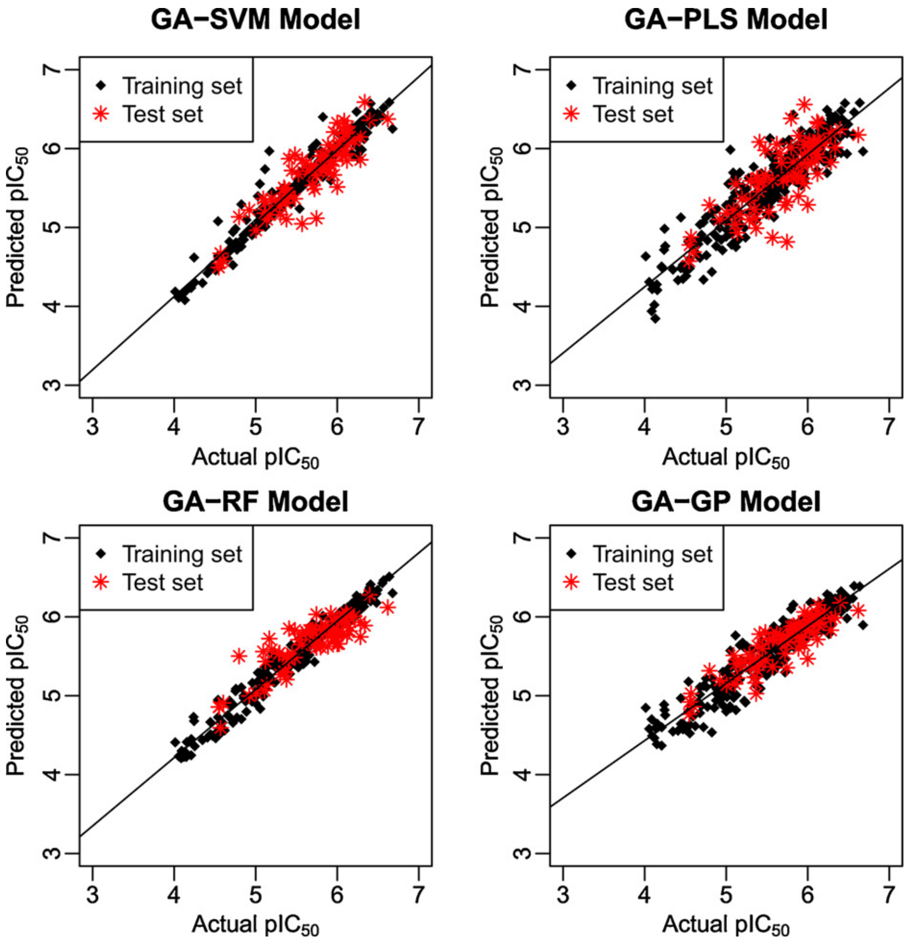


Fig. 3. The predicted versus the actual pIC_{50} values for the P2Y_{12} antagonists. (A) GA–SVM model; (B) GA–PLS model; (C) GA–RF model; and (D) GA–GP model.

scatter plot for predicted versus experimental pIC_{50} values of the training and test sets. In a word, PLS generates a relatively poor QSAR model for these P2Y_{12} antagonists.

3.4.3. GA–RF

Random forest effectively has only one tuning parameter, m_{try}. Since in the present work, there are 38 variables to be selected by GA as the optimal subset of descriptors, the m_{try} value is tried from 1 to 38, the optimal one of which is determined also by the 10-fold cross-validation RMSE (0.309). Thus, RF results are obtained based on the optimal m_{try} (=12) and 500 trees in the forest. For the training and test sets, the RMSE values of 0.128 and 0.245, a coefficient of determination, the R^2 of 0.966 are obtained. In addition, the $R^2{}_{\text{cv}}$ is 0.750 and the R^2_{ts} and R^2_{pred} for the test set are 0.745 and 0.739, respectively. Fig. 3C shows the performance of the RF model for the data sets.

3.4.4. GA–GP

The Gaussian process method, based on clearly defined statistical principles which is easily programmed [66], is also adopted to predict the P2Y_{12} inhibition activity. It can be noted that 44 descriptors are determined finally to build the optimal GP model. The optimal inverse kernel width for the Radial Basis kernel function (sigma) finally fixes to 0.020 based on the sigest function included in the R package kernlab. The resulting GP model gives statistical results of R^2, RMSE, R^2_{cv} values of 0.877, 0.234, 0.737 for the training set, and $R^2_{\text{ts}} = 0.783$, $R^2_{\text{pred}} = 0.764$, RMSE = 0.234 for the test set, respectively. Fig. 3D depicts the scatter plot of the GP model based on the current dataset.

3.5. Further tests on the external predictability

To believe firmly the performance of the prediction, the squared correlation coefficient values between the observed and predicted

Table 3
Comparison of the external predictability of GA–SVM, GA–PLS, GA–RF and GA–GP models for the prediction set.

Model	R^2	R_o^2	$(R^2 - R_o^2)/R^2$	R_m^2	k	k'
GA–SVM	0.806	0.792	0.018	0.710	0.999	0.999
GA–PLS	0.594	0.530	0.108	0.443	0.993	1.004
GA–RF	0.745	0.461	0.381	0.348	0.996	1.002
GA–GP	0.783	0.501	0.361	0.367	0.994	1.004

Table 4
Comparison with and without Y-randomization check of the optimal GA–SVM model.

Model	Calibration set			Prediction set		
	R^2	R_{cv}^2	RMSE	R_{ts}^2	R_{pred}^2	RMSE
GA–SVM[a]	0.976	0.829	0.133	0.806	0.811	0.209
GA–SVM[b]	0.011	−0.131	0.696	0.039	−0.425	0.571

[a] Without Y-randomization check.
[b] With Y-randomization check.

values of the test set compounds with intercept (R_{ts}^2) and without intercept (R_o^2) are also calculated. Table 3 presents the values of the parameters for all models in the present work. According to references [42,82–84], models are considered acceptable if they satisfy all following conditions: (1) $R_{pred}^2 > 0.5$, (2) $R_{ts}^2 > 0.6$, and (3) R_o^2 is close to R_{ts}^2, such that the $[(R^2 - R_o^2)/R^2] < 0.1$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$. When the observed values of the test set compounds (X axis) are plotted against the predicted values of the compounds (Y axis) with the intercept set to zero, slope of the fitted line gives the value of k and interchange of the axes gives the value of k' respectively. All the models (except the SVM one) have not satisfied the requirement of the value of $(R_{ts}^2 - R_o^2)/R_{ts}^2$ being less than 0.1.

Previous report [85] has illustrated that the $R^2{}_{pred}$ may not truly reflect the predictive capability of a model on a new dataset. Also, the squared regression coefficient (R_{ts}^2) between the observed and predicted values of the test set compounds does not necessarily mean that the predicted values are very near to the observed activities (as there may be considerable numerical difference between the values though maintaining an overall good inter-correlation). To better evaluate the external predictive capacity of a model a modified R^2 term (R_m^2) is been defined as follows [86]:

$$R_m^2 = R_{ts}^2 \times (1 - \left|\sqrt{R_{ts}^2 - R_o^2}\right|) \qquad (12)$$

In case of good external prediction capacity, predicted values will be very close to the observed ones and thus the R_{ts}^2 value will be very near to the R_o^2 one. In the best case R_m^2 may be equal to R_{ts}^2, whereas in the worst case R_m^2 value could be zero. Here, the three models (i.e., PLS, RF and GP) are all less than the recommended value (0.5). Only the SVM model achieves a best R_m^2 value of 0.710.

3.6. Comparison of different approaches

After the above discussion, it can be concluded that the developed SVM model based on the GA optimization on the parameters (C, γ, ε) and the descriptor subset outperforms all other three ones in terms of the statistics. Most importantly, this model has passed through every rigorous examinations, especially as it is the only one with an R_m^2 value larger than 0.5. In addition, the PLS model is observed uniformly less accurate both in the training and test sets when compared with other three models, suggesting that the linear relationship of this series of $P2Y_{12}$ data set is not obvious compared to the nonlinear one (Fig. 3 and Table 2). Thirdly, though the performance of RF is better than GP in the training set, the external evaluation of the RF model gives worse results (Table 2), thus proving the better generalization performance of the GP model than RF. In summary, in our modeling process the SVM illustrates the best performances, and is more suitable to achieve further prediction task for unknown $P2Y_{12}$ antagonist data set.

3.7. Outlier test

Outliers from a QSAR are compounds that do not fit the model or are poorly predicted by it [87]. Many reasons may exist for the presence of outliers in the dataset used for in silico modeling. Typically, some outliers are recognized as acting by a different mechanism of action from other molecules, which may be well modeled by QSAR techniques, and thus do not follow the general structure-activity rule established by this modeling. When performed correctly, the removal of outliers will allow for the development of stronger and more significant models, and the outlier test is therefore reasonable and necessary in the derived models. There are a variety of methods to highlight outliers including, at the most basic level, identifying those compounds with significantly high residuals from regression-based techniques. At present, since the proposed SVM model presents wonderful prediction ability over the others, only this one is checked to identify possible outliers. It can be observed that none of residuals in both the training and test sets is more than one log unit, illustrating that the SVM model is both a robust and predictive one. Thus it is reasonable to consider that there are no outliers in the present SVM model.

3.8. Y-randomization check

Presently, the Y-randomization check [84] is implemented for further assurance of the robustness of the optimal GA–SVM model. The dependent variable is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to possess low R_{tr}^2, R_{cv}^2, R_{ts}^2, R_{pred}^2 and high RMSE for the training and test sets, respectively. If the opposite happens, then an acceptable QSAR model cannot be obtained for the specific modeling method and data. In the current work, 500 times of Y-randomization checks are repeated and the ended results are compared with those prediction statistics without such checks, with the average values reported in Table 4. As shown in this table, the correlation coefficients have a significant decline while the RMSE values sharply increase, which indicates that the proposed GA–SVM model is not due to a chance correlation.

4. Conclusion

In the present work, we have developed a GA–SVM method as efficient tool for simultaneous parameters optimization and descriptor subset selection. As far as we know, there is still no similar research reported up to now that carried out such simultaneous feature selection and the parameters optimization (not only C, γ but also ε) for SVM regression analysis. In addition, this is also the first QSAR study for the prediction of an unusually large dataset of 364 P2Y_{12} antagonists with diversity of structures by using the proposed GA–SVM.

In addition, three other widely used approaches including the PLS, RF and GP are also employed combined with GA on the dataset and the models they established are compared with the GA–SVM model in terms of several rigorous evaluation criteria. As a result, the GA–SVM model has gone throughout all rigorous examinations suggested by all relating references [42,83–85], with the best qualities and generalization capabilities than the other approaches, demonstrating its feasibility and reliability to derive highly predictive model for P2Y_{12} antagonists. Results from the GA–SVM model also suggest that the electronic factors, hydrophobic and H-bond interaction play a central role in the P2Y_{12} inhibition. Thus, the proposed models may provide an insight into some instructions for further synthesis of highly potent P2Y_{12} antagonists and should be useful for the predictive tasks to screen for new and potent P2Y_{12} antagonists in early drug development.

Acknowledgements

The authors are very grateful to the University of Sheffield for access to the genetic algorithm tool box. The authors also thank the R Development Core Team for affording the free R2.10 software. We would like to thank Professor Chih-Jen Lin (National Taiwan University) for kindly allowing us to use LibSVM.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.aca.2011.02.004.

References

[1] H. Horiuchi, Ann. Med. 38 (2006) 162–172.
[2] N. Turner, J. Moake, L. McIntire, Blood 98 (2001) 3340–3345.
[3] J. Remijn, Y. Wu, E. Jeninga, M. IJsseldijk, G. van Willigen, P. de Groot, J. Sixma, A. Nurden, P. Nurden, Arterioscler. Thromb. Vasc. Biol. 22 (2002) 686–691.
[4] C. Gachet, Thromb. Haemost. 86 (2001) 222–232.
[5] R. Nicholas, Mol. Pharmacol. 60 (2001) 416–420.
[6] C. Léon, C. Ravanat, M. Freund, J. Cazenave, C. Gachet, Arterioscler. Thromb. Vasc. Biol. 23 (2003) 1941–1947.
[7] G. Hollopeter, H. Jantzen, D. Vincent, G. Li, L. England, V. Ramakrishnan, R. Yang, P. Nurden, A. Nurden, D. Julius, Nature 409 (2001) 202–207.
[8] T. Meadows, D. Bhatt, Circ. Res. 100 (2007) 1261–1275.
[9] P. Gurbel, K. Bliden, B. Hiatt, C. O'Connor, Circulation 107 (2003) 2908–2913.
[10] B. Springthorpe, A. Bailey, P. Barton, T. Birkinshaw, R. Bonnert, R. Brown, D. Chapman, J. Dixon, S. Guile, R. Humphries, Bioorg. Med. Chem. Lett. 17 (2007) 6013–6018.
[11] D. Agrafiotis, D. Bandyopadhyay, J. Wegner, H. van Vlijmen, J. Chem. Inf. Model. 47 (2007) 1279–1293.
[12] X. Sun, Y. Li, X. Liu, J. Ding, Y. Wang, H. Shen, Y. Chang, Mol. Divers. 12 (2008) 157–169.
[13] J. Serra, E. Thompson, P. Jurs, Chem. Res. Toxicol. 16 (2003) 153–163.
[14] S. Doniger, T. Hofmann, J. Yeh, J. Comput. Biol. 9 (2002) 849–864.
[15] Y. Wang, Y. Li, S. Yang, L. Yang, J. Chem. Inf. Model. 45 (2005) 750–757.
[16] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, J. Chem. Inf. Model. 48 (2008) 1337–1344.
[17] M. Hao, Y. Li, Y. Wang, S. Zhang, Int. J. Mol. Sci. 11 (2010) 3413–3433.
[18] Y. Wang, Y. Li, J. Ding, Y. Chang, Mol. Divers. 12 (2008) 93–102.
[19] Y. Wang, Y. Li, B. Wang, Int. J. Mol. Sci. 8 (2007) 166–179.
[20] Z. Wang, Y. Li, C. Ai, Y. Wang, Int. J. Mol. Sci. 11 (2010) 3434–3458.
[21] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.
[22] O. Obrezanova, M. Segall, J. Chem. Inf. Model. 50 (2010) 1053–1061.
[23] P. Zhou, X. Chen, Y. Wu, Z. Shang, Amino Acids 38 (2010) 199–212.
[24] Y. Li, Y. Wang, J. Ding, Y. Wang, Y. Chang, S. Zhang, QSAR Comb. Sci. 28 (2009) 396–405.
[25] M. Pontes, R. Galvãob, M. Araújo, P. Moreira, O. Neto, G. Joséa, T. Saldanha, Chemom. Intell. Lab. Syst. 78 (2005) 11–18.
[26] G. Bakken, P. Jurs, J. Med. Chem. 43 (2000) 4534–4541.
[27] E. Pourbasheer, S. Riahi, M. Ganjali, P. Norouzi, Eur. J. Med. Chem. 45 (2010) 1087–1093.
[28] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Mach. Learn. 46 (2002) 389–422.
[29] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci. 37 (1997) 306–310.
[30] O. Deeb, B. Hemmateenejad, A. Jaber, R. Garduno-Juarez, R. Miri, Chemosphere 67 (2007) 2122–2130.
[31] D.F. Cook, C.T. Ragsdale, R.L. Major, Eng. Appl. Artif. Intell. 13 (2000) 391–396.
[32] C. Huang, C. Wang, Expert. Syst. Appl. 31 (2006) 231–240.
[33] J.J. Parlow, M.W. Burney, B.L. Case, T.J. Girard, K.A. Hall, P.K. Harris, R.R. Hiebsch, R.M. Huff, R.M. Lachance, D.A. Mischke, S.R. Rapp, R.S. Woerndle, M.D. Ennis, J. Med. Chem. 53 (2010) 2010–2037.
[34] J.J. Parlow, M.W. Burney, B.L. Case, T.J. Girard, K.A. Hall, R.R. Hiebsch, R.M. Huff, R.M. Lachance, D.A. Mischke, S.R. Rapp, R.S. Woerndle, M.D. Ennis, Bioorg. Med. Chem. Lett. 19 (2009) 4657–4663.
[35] J.J. Parlow, M.W. Burney, B.L. Case, T.J. Girard, K.A. Hall, R.R. Hiebsch, R.M. Huff, R.M. Lachance, D.A. Mischke, S.R. Rapp, R.S. Woerndle, M.D. Ennis, Bioorg. Med. Chem. Lett. 19 (2009) 6148–6156.
[36] J.J. Parlow, M.W. Burney, B.L. Case, T.J. Girard, K.A. Hall, P.K. Harris, R.R. Hiebsch, R.M. Huff, R.M. Lachance, D.A. Mischke, S.R. Rapp, R.S. Woerndle, M.D. Ennis, Bioorg. Med. Chem. Lett. 20 (2010) 1388–1394.
[37] B. Hemmateenejad, M. Elyasi, Anal. Chim. Acta 646 (2009) 30–38.
[38] B. Hemmateenejad, M. Safarpour, A. Mehranpour, Anal. Chim. Acta 535 (2005) 275–285.
[39] S. Macho, A. Rius, M. Callao, M. Larrechi, Anal. Chim. Acta 445 (2001) 213–220.
[40] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, X. Chen, H.-D. Li, J. Chemom. 24 (2010) 584–595.
[41] B. Hemmateenejad, K. Javadnia, M. Elyasi, Anal. Chim. Acta 592 (2007) 72–81.
[42] A. Golbraikh, A. Tropsha, J. Comput. Aided Mol. Des. 16 (2002) 357–369.
[43] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Chemom. Intell. Lab. Syst. 33 (1996) 35–46.
[44] L.F. Capitán-Vallvey, N. Navas, M. del Olmo, V. Consonni, R. Todeschini, Talanta 52 (2000) 1069–1079.
[45] R.K.H. Galvão, M.C.U. Araujo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Talanta 67 (2005) 736–740.
[46] Y. Zhao, M. Abraham, A. Ibrahim, P. Fish, S. Cole, M. Lewis, M. de Groot, D. Reynolds, J. Chem. Inf. Model. 47 (2007) 170–175.
[47] ISIS Draw 2.3, MDL Information Systems, Inc.
[48] M. Taha, A. Qandil, D. Zaki, M. AlDamen, Eur. J. Med. Chem. 40 (2005) 701–727.
[49] P. Mazzatorta, M.T.D. Cronin, E. Benfenati, QSAR Comb. Sci. 25 (2006) 616–628.
[50] H. Gao, J. Chem. Inf. Comput. Sci. 41 (2001) 402–407.
[51] M.H. Fatemi, M. Jalali-Heravi, E. Konuze, Anal. Chim. Acta 486 (2003) 101–108.
[52] D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[53] L. Davis, M. Mitchell, Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.
[54] J. Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.
[55] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[56] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Chem. Biol. Drug Des. 72 (2008) 205–216.
[57] H. Li, Y. Liang, Q. Xu, Chemom. Intell. Lab. Syst. 95 (2009) 188–198.
[58] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[59] H. Liu, R. Zhang, X. Yao, M. Liu, Z. Hu, B. Fan, Anal. Chim. Acta 525 (2004) 31–41.
[60] J.M. Luco, J. Chem. Inf. Comput. Sci. 39 (1999) 396–404.
[61] S. Wold, Technometrics 20 (1978) 397–405.
[62] R. Wehrens, B.H. Mevik, Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR), 2007, http://cran.r-project.org/web/packages/pls/index.html.
[63] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[64] A. Liaw, M. Wiener, Breiman and Cutler's Random Forests for Classification and Regression, 2010, http://cran.r-project.org/web/packages/randomForest/index.html.
[65] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006.
[66] F. Burden, J. Chem. Inf. Comput. Sci. 41 (2001) 830–835.
[67] D. Enot, R. Gautier, J. Marouille, SAR QSAR Environ. Res. 12 (2001) 461–469.
[68] P. Tiño, I.T. Nabney, B.S. Williams, J. Lösel, Y. Sun, J. Chem. Inf. Comput. Sci. 44 (2004) 1647–1653.
[69] A. Schwaighofer, T. Schroeter, S. Mika, J. Laub, A. Ter Laak, D. Sülzle, U. Ganzer, N. Heinrich, K. Müller, J. Chem. Inf. Model. 47 (2007) 407–424.
[70] T. Schroeter, A. Schwaighofer, S. Mika, A. Ter Laak, D. Suelzle, U. Ganzer, N. Heinrich, K. Müller, ChemMedChem 2 (2007) 1265–1267.
[71] O. Obrezanova, G. Csányi, J.M.R. Gola, M.D. Segall, J. Chem. Inf. Model. 47 (2007) 1847–1857.
[72] T. Chen, J. Morris, E. Martin, Chemom. Intell. Lab. Syst. 87 (2007) 59–71.
[73] A. Karatzoglou, A. Smola, K. Hornik, Kernlab: Kernel-based Machine Learning Lab, 2010, http://cran.r-project.org/web/packages/kernlab/index.html.
[74] V. Svetnik, T. Wang, C. Tong, A. Liaw, R.P. Sheridan, Q.H. Song, J. Chem. Inf. Model. 45 (2005) 786–799.

[75] C.L. Bruce, J.L. Melville, S.D. Pickett, J.D. Hirst, J. Chem. Inf. Model. 47 (2007) 219–227.

[76] L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.

[77] G. Moreau, P. Broto, Nouv. J. Chim. 4 (1980) 359–360.

[78] F. Burden, J. Chem. Inf. Comput. Sci. 29 (1989) 225–227.

[79] F. Burden, Quant. Struct.: Act. Relat. 16 (1997) 309–314.

[80] D.T. Stanton, J. Chem. Inf. Comput. Sci. 39 (1999) 11–20.

[81] B. Pirard, S. Pickett, J. Chem. Inf. Comput. Sci. 40 (2000) 1431–1440.

[82] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.

[83] A. Golbraikh, M. Shen, Z.Y. Xiao, Y.D. Xiao, K.H. Lee, A. Tropsha, J. Comput. Aided Mol. Des. 17 (2003) 241–253.

[84] A. Tropsha, P. Gramatica, V. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.

[85] K. Roy, A. Mandal, J. Enzyme Inhib. Med. Chem. 24 (2009) 205–223.

[86] P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.

[87] W.J. Egan, S.L. Morgan, Anal. Chem. 70 (1998) 2372–2379.