Chem Biol Drug Des 2010; **75:** 506–514

Research Article

doi: 10.1111/j.1747-0285.2010.00953.x

Exploring QSARs for Inhibitory Activity of Non-peptide HIV-1 Protease Inhibitors by GA-PLS and GA-SVM

Omar Deeb[1*] and Mohammad Goodarzi[2,3]

[1]Faculty of Pharmacy, Al-Quds University, P.O. Box 20002, Jerusalem, Palestine
[2]Department of Chemistry, Faculty of Sciences, Islamic Azad University, Arak Branch, P.O. Box 38135–567 Arak, Markazi, Iran
[3]Young Researchers Club, Islamic Azad University, Arak Branch, P. O. Box 38135–567 Arak, Markazi, Iran
**Corresponding author: Omar Deeb, deeb2000il@yahoo.com*

The support vector machine (SVM) and partial least square (PLS) methods were used to develop quantitative structure activity relationship (QSAR) models to predict the inhibitory activity of non-peptide HIV-1 protease inhibitors. Genetic algorithm (GA) was employed to select variables that lead to the best-fitted models. A comparison between the obtained results using SVM with those of PLS revealed that the SVM model is much better than that of PLS. The root mean square errors of the training set and the test set for SVM model were calculated to be 0.2027, 0.2751, and the coefficients of determination (R^2) are 0.9800, 0.9355 respectively. Furthermore, the obtained statistical parameter of leave-one-out cross-validation test (Q^2) on SVM model was 0.9672, which proves the reliability of this model. The results suggest that TE2, Ui, GATS5e, Mor13e, ATS7m, Ss, Mor27e, and RDF035e are the main independent factors contributing to the inhibitory activities of the studied compounds.

Key words: inhibitory activity, HIV-1 protease inhibitors, quantitative structure activity relationship, support vector machine, partial least square, genetic algorithms

Received 12 March 2009, revised 12 January 2010 and accepted for publication 28 January 2010

Human immunodeficiency virus (HIV), the causative agent of AIDS, now infects millions of people worldwide. Although a cure has not been found yet for this fatal disease, rapid advances in molecular biology along with the 3-D elucidation of HIV proteins have led to new drug-targeting approaches for designing antiviral agents that specifically bind to key regulatory proteins that are essential for HIV replication. Thus, by developing new inhibitors of HIV-1 protease activity, the treatment of AIDS can be advanced (1–6).

Peptidic and peptidomimetic non-hydrolyzable transition state mimics were rapidly developed as highly potent HIV protease inhibitors (7,8). These competitive inhibitors possess optimal interactions in substrate binding pockets with low to subnanomolar K_i values. However, their peptidic nature often makes them poor pharmacological agents, with low bioavailability and rapid clearance (8). Several peptidic inhibitors are currently under clinical trials and significant efforts to improve their pharmacology continues. In this study, we picked out small non-peptide HIV protease inhibitors with potentially better pharmacological characteristics based on the structural features of peptidic inhibitors bound to the enzyme (9), and performed quantitative structure activity relationship (QSAR) studies.

QSAR studies resort to several statistical techniques, which can be applied for model construction. For example, multiple linear regressions (MLR) and artificial neural networks (ANN) (10–16) are used for inspection of linear and non-linear correlations between activity and molecular descriptors respectively. Neural networks have some problems inherent to their architecture, such as overtraining, overfitting, network optimization, and reproducibility of results. This is mainly because of random initialization of the networks and variation of stopping criteria (17). Because of these limitations, other more accurate and informative QSAR techniques are used. The support vector machine (SVM) is a new algorithm developed from the machine learning community (18). 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. Owing to its remarkable generalization performance, the SVM has attracted attention and gained extensive application, such as pattern recognition problems (17,19), drug design (20), QSAR (21,22), and quantitative structure–property relationship (QSPR) analysis (23). In most of these cases, results obtained from SVM modeling either matches or is significantly better than that of traditional machine learning approaches. The work presented here employs SVM in predicting inhibitory activity of non-peptidic HIV-1 protease inhibitors.

In this research, we performed QSAR study on the inhibitory activity of non-peptide HIV-1 protease inhibitors using the SVM technique.

Performance of this model was compared with that of the PLS method.

Materials and methods

Data set

The data used in this QSAR study consisted of inhibitory activity data (IC_{50}), which is the half maximal (50%) inhibitory concentration (IC) of a compound, have been reported by Tummino *et al.* (24). The activity data [IC_{50} (μM)] for non-peptide HIV-1 protease inhibitors (Table 1) was converted to the logarithmic scale [−log IC50 (M)] and then used for subsequent QSAR analyses as the response variables. Figure 1 shows the general structure of non-peptide HIV-1 protease inhibitors.

Software

Geometry optimization was performed using HyperChem[a] (Version 7.5 Hypercube, Inc., Gainesville, FL, USA). Dragon[b] 5.0 (Milano Chemometrics and QSAR Research Group, Milano, Italy) software was utilized to calculate the molecular descriptors. The SPSS software (version 16.0, SPSS, Inc., IL, USA) was employed for the simple statistical analysis. The SVM evaluations were carried out using the SVM toolbox for use with Matlab (Version 7.6, Math works, Inc., Natick, MA, USA) that was developed by Gunn (25).

Descriptors calculation and selection

Molecular chemical structure was built using Hyperchem 7.5 software. AM1 method (26) was applied to optimize the molecular structure of the compounds. No molecular symmetry constraint was applied; rather full optimization of all bond lengths and angles was carried out. All calculations were carried out at the restricted Hartree-Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.01 Kcal/mol.

Geometry optimization was run multiple times with different starting points for each molecule, and the lowest energy conformation was considered for the calculation of electronic properties. The resulting geometry was transferred into the Dragon program package developed by Milano Chemometrics and QSPR Group to calculate descriptors in Constitutional, Topological, Geometrical, Charge, GETAWAY (GEometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D Molecular Representation of Structure based on Electron diffraction), Molecular Walk Count, BCUT, 2D-Autocorrelation, Aromaticity Index, Randic molecular profile, Radial Distribution Function, Functional group, and Atom-Centered Fragment classes. Molecular descriptors (1481) belonging to eighteen different theoretical descriptors were calculated for each molecule. The calculated descriptors were first analyzed for the existence of constant or near constant variables. The detected ones were then removed. Correlation among descriptors with the activity of the molecules was examined and collinear descriptors (i.e. $r > 0.9$) were detected. Descriptors that contain a high percentage (> 90%) of identical values were discarded to decrease the redundancy existing in the descriptor data matrix. Among the collinear descriptors, the one presenting the highest correlation with the activity was retained and others were removed from the data matrix. The dataset was splitted into two sets based on activity range; training set (85%) with activity ranges from 3.921 to 8.444 and test set (15%) with activity ranges from 4.538 to 8.208 (27). In this work, genetic algorithm (GA) variable subset selection method (28) was used for the selection of the most relevant descriptors from the pool of remaining descriptors. These descriptors would be

Table 1: Activity data [pIC50 (M)] for non-peptide HIV-1 protease inhibitors[a]

No.	R_1	R_2	R_3	Chirality	n	pIC_{50}
1	–	–	–		–	5.161
2	–	–	–		–	5.091
3	–	–	–		–	4.721
4	–	–	–		–	4.538
5[b]	–	–	–		–	4.538
6	H	H	$(CH_2)_3OPh$			5.638
7	H	H	CH_2Ph			3.921
8	H	H	$CH(Ph)CH_2COCH_3$			4.745
9	CH_3	CH_3	$CH(Ph)CH_2COCH_3$			5.721
10	H	H	$(CH_2)_3SPh$			4.585
11	H	H	$(CH_2)_4Ph$			3.959
12	H	H	(CH2)4-Ph(2-OCH_3)			5.769
13	OH	H	(CH2)4-Ph(2-OCH_3)			6.284
14	H	Ph				5.523
15	H	CH_2Ph				5.769
16	H	CH_2CH_2Ph				5.886
17	4-OH	CH_2CH_2Ph				6.284
18[b]	4-OCH_2CO_2H	CH_2CH_2Ph				6.796
19	H	Ph(2-Me)				6.377
20	H	Ph(2-iPr)				7.432
21	H	Ph(2-tBu)				7.769
22	(3-Me)	Ph(2-iPr)				8.155
23	Ph	Ph				6.108
24[b]	Isobutyl	Ph				6.387
25	Ph	CH_2Ph				6.319
26	Isobutyl	CH_2Ph				6.585
27	CH_2-cyclopropyl	CH_2Ph				7.076
28	CH_2-cyclopropyl	Cyclohexyl				6.824
29	CH_2-cyclopropyl	Cyclopentyl				7.161
30[b]	Isobutyl	Cyclopentyl				7.236
31	Cyclopentyl	Cyclopentyl				6.638
32	H				2	5.677
33	$CH_3(CH_2)_2$				2	5.292
34	$CH_3(CH_2)_4$				2	7.076
35[b]	$(CH_3)_2CH(CH_2)_2$				2	7.018
36	$HO_2C(CH_2)_4$				2	8.301
37	Ph				1	6.585
38	$Ph(CH_2)_2$				1	7.222
39	H	H		6-R,S		6.886
40[b]	Me	H		6-R,S		7.137
41	Isopropyl	H		6-R,S		7.854
42	Isopropyl	Me		6-R,S		8.137
43	Isopropyl	Me		6-R		6.886
44[b]	Isopropyl	Me		6-S		8.208
45	t-Butyl	H		6-R,S		8.444
46	t-Butyl	Me		6-R,S		8.009

[a]See Figure 1 for the general structure of the compounds under study.
[b]Compounds in the test set.

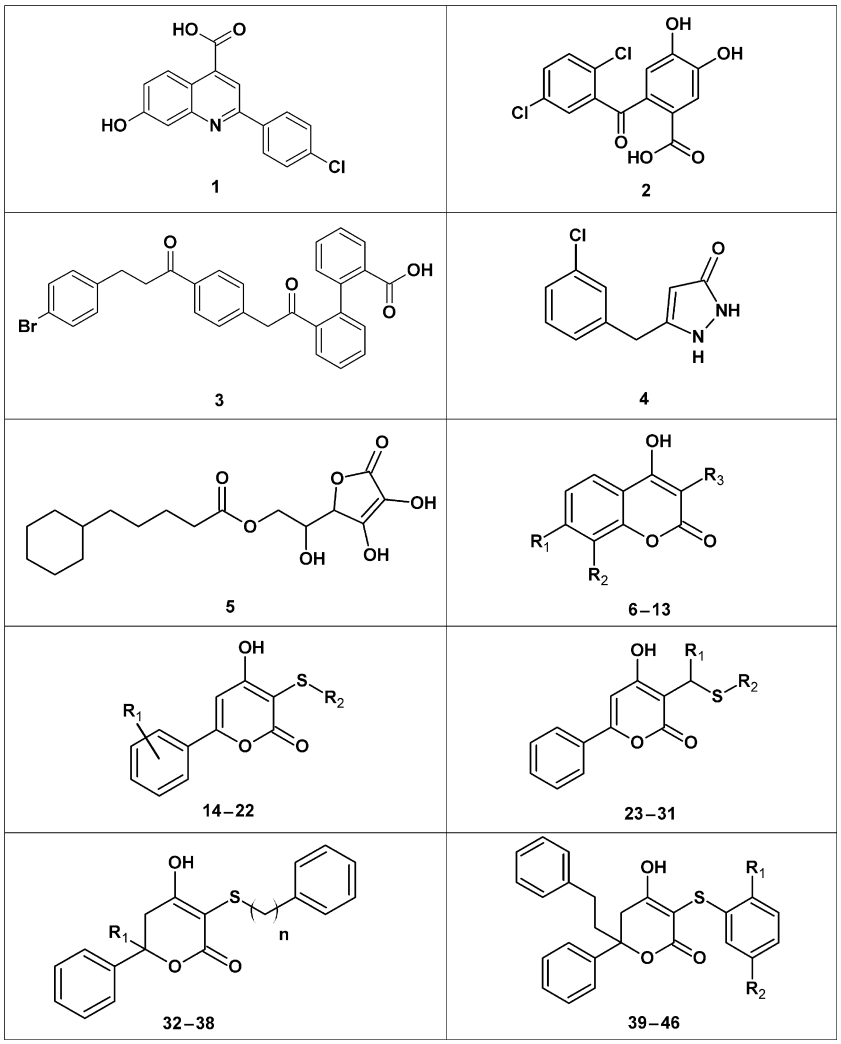

Figure 1: General structure of non-peptide HIV-1 protease inhibtor.

used as inputs of PLS and SVM for the construction of QSAR models.

Genetic algorithm

Genetic algorithms are an interesting and widely used variable selection method. It applies Darwin evolution hypothesis and different genetic functions, i.e. crossover and mutation to optimize problems defined by fitness criteria.

To select the most relevant descriptors, population evolution was simulated (29–31) using the first generation population selected randomly. Each individual member in the population, defined by a chromosome of binary values, represents a subset of descriptors. The number of the genes at each chromosome was equal to the number of the descriptors. A gene was given the value of 1, if its corresponding descriptor was included in the subset; otherwise, it was given zero value. The number of genes with the value of 1 was kept relatively low to have a small subset of descriptors (32). As a result, the probability of generating 0 for a gene was greater (at least 60%) than the probability of generating 1. The operators used here were crossover and mutation. The application probability of these operators varied linearly with a generation renewal (0–0.1% for mutation and 60–90% for crossover). The population size varied between 50 and 250 for different GA runs. For a typical run, the evolution of the population stopped when 90% of the members displayed the same fitness.

Partial least square (PLS)

PLS is the method used for building regression models on the latent variable decomposition relating two blocks, matrices **X** and **Y**, which contain the independent x and dependent y variables respectively. These matrices can be simultaneously decomposed into a sum of *f* latent variables, as follows:

$$X = TP^T + E = \sum t_f p'_{rf} + E \quad (1)$$

$$Y = UQ^T + F = \sum u_f q'_f + F \quad (2)$$

in which *T* and *U* are the score matrices for *X* and *Y* respectively. *P*and *Q* are the loadings matrices for *X* and *Y*respectively. *E* and *F*

are the residual matrices. The two matrices are correlated by the scores T and U, for each latent variable, as follows:

$$u_f = b_f t_f \tag{3}$$

where b_f is the regression coefficient for the f latent variable. The matrix Y can be calculated from u_f, as shown in eqn 4, and the acidity constant of the new samples can be estimated from the new scores T^*, which are substituted in eqn 4, leading to eqn 5

$$Y = TBQ^T + F \tag{4}$$

$$Y_{new} = T^* BQT \tag{5}$$

In this procedure, it is necessary to find the best number of latent variables, which is normally performed using cross-validation that is based on the determination of minimum prediction error. Applications of PLS have been discussed by several researchers (10,33).

Support vector machine

SVM can be applied to regression by the introduction of an alternative loss function and the results appear to be very encouraging. In support vector regression (SVR), the basic idea is to map the data X into a higher-dimensional feature space F via a non-linear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i; y_i)\}_{i=1}^{N}$ (x_i contains independent variables, y_i contains dependent variables, and N is the total number of data patterns). SVM approximates the function in the following form:

$$f(x) = \sum_{i=1}^{N} w_i \Phi(x) + b \tag{6}$$

where w is weight vector, $\{(\Phi_i(x)\}_{i=1}^{N}$ is the set of mappings of input features, $\{w_i\}_{i=1}^{N}$ and b are the slope and offset of the regression function respectively. They are estimated by minimizing the following cost function;

$$R(C) = \frac{1}{2}\|w\|^2 + \frac{C}{N}\sum_{i=1}^{N} L_\varepsilon(y_i - f(x_i), \varepsilon) \tag{7}$$

Where

$$L_\varepsilon(y_i - f(x_i), \varepsilon) = \left\{ \begin{array}{ll} 0 & if |y_i - f(x_i)| \le \varepsilon \\ |y_i - f(x_i)| - \varepsilon & otherwise \end{array} \right\} \tag{8}$$

ε is a precision parameter representing the radius of the tube located around the regression function f(x).

In eqn 7, $[\|\boldsymbol{w}\|^2]$ is the regularization term that controls the trade-off between the complexity and the approximation accuracy of the regression model to ensure that the model possesses an improved generalized performance. The second term $[C/N\sum_i^N L_\varepsilon(y_i - f(x_i), \varepsilon)]$ is the so-called empirical error (risk) measured by ε-insensitive loss function L_ε (y_i - $f(x_i)$), which indicates that it does not penalize errors below $\varepsilon \ge 0$. C is the regularized constant determining the trade-off between the empirical risk and regularization term. Introduction of slack variables 'ξ' leads eqn 7 to the following constrained function: Min.

$$R(w, \xi, \xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{9}$$

Subjected to

$$\begin{aligned} w\Phi(x_i) + b - y_i \le \varepsilon + \xi_i, \\ y_i - w\Phi(x_i) - b_i \le \varepsilon + \xi_i^* \qquad \varepsilon, \xi_i, \xi_i^* \ge 0 \end{aligned} \tag{10}$$

By using Lagrangian multipliers and Karush–Kuhn–Tucker conditions to the eqn 9, it yields the following dual Lagrangian form (34), Maximize:

$$\begin{aligned} PSI(\alpha_i, \alpha_i^*) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \\ - \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)y_i \end{aligned} \tag{11}$$

with the following constrains

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0, \quad 0 \le \alpha_i \le C \quad and \quad 0 \le \alpha_i^* \le C \tag{12}$$

The Lagrangian multipliers in eqn 11 satisfy the equality $\alpha_i \times \alpha_i^* = 0$ (35,36). The Lagrangian multipliers α_i and α_i^* are calculated, and an optimal desired weight vector of the regression hyperplan is $\boldsymbol{w}^* = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j})$. Hence, the general form of the SVR-based regression function can be written as (34,37):

$$f(x, w) = f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b \tag{13}$$

Based on the Karush–Kuhn–Tucker (KKT) conditions (38,39) of quadratic programming, only a number of coefficients α_i ***and*** α_i^* will be non-zero, and the data points associated with these parameters refer to the support vectors of the model.

In eqn 13, $K(x_i,x_j)$ is the kernel function. The value is equal to the inner product of two vectors x_i and x_j in the feature space $\Phi(x)$. That is $K(x_i,x_j) = \Phi(x_i)\Phi(x_j)$. The elegance of using kernel function stems from the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function.

In this work, the Gaussian radial basis function (RBF) kernel was used as kernel function, $k(x_i,x_j) = \exp(-(\|x^i - x_j\|^2)/2\sigma^2)$, where σ^2 is the width of the Gaussian function, so the C and σ that are the relative weights of the regression error and the kernel parameter of the RBF kernel should be optimized by the user, to obtain the

support vector. The parameters of SVMR were optimized by systemically changing their values in the training step and calculating the RMSE and accuracy of the model using 5-fold cross-validation. The optimized values of C, σ^2, and ε were 8, 0.5, and 0.06 respectively obtained based on minimum RMSE and maximum accuracy of model.

Results and discussion

The prediction ability of QSAR/QSPR models is affected by two parameters: the first one is whether the descriptors carry enough structural information to enable the interpretation of the activity/property being investigated. The second is the accuracy of the modeling methods employed. For the selection of the most important descriptors, genetic algorithm variable subset selection method was used. As a rule of thumb, at least five compounds should be included in the equation for every descriptor (40,41). The entire GA process can be summarized as follows: (i) generate random variables subsets; (ii) evaluate each individual subset of selected descriptors for fitness to predict pIC_{50}; (iii) discard worse half of individuals; (iv) breed remaining individuals (or chromosome); (v) allow for mutation; (vi) repeat steps 2–5 until ending criteria are met. The GA finishes when one of two conditions is found: (i) after a finite number of iterations or (ii) after some percentage of the individuals in the population are using identical variable subsets. It is worth mentioning that individuals with noisy variables tend to be discarded and, thus, the variables used by those individuals become less represented in the overall gene population. On the other hand, less noisy variables become more and more represented. Depending on the number of variables and the rate of mutation, many of the individuals eventually contain the same genes. The GA applied to the variable selection in this article uses a binary representation as the coding technique for the given problem; the presence or absence of a descriptor in a chromosome is coded by 1 or 0. The GA performs its optimization by variation and selection via the evaluation of the fitness function (RMSECV). The GA parameters that were used in this study are: population size 64, maximum generations 100, mutation rate 0.005, iteration 100, and cross-over 0.6. Eight descriptors were selected by GA, which are: TE2, Ui, GATS5e, Mor13e, ATS7m, Ss, Mor27e, and RDF035e. Table 2 shows the short description of these descriptors. The correlation matrix of these descriptors is shown in Table S1 in the supplementary material, which shows that the selected descriptors are independent. To demonstrate the absence of chance correlations on the eight descriptors obtained with the previous procedure, we performed a Y-scrambling test on the training set, where the output values of the compounds were shuffled randomly, and the scrambled data set was re-examined by the PLS method against real (unscrambled) input descriptors to determine the correlation and predictivity of the resulting model. The R^2 values obtained from the Y-randomization test are in the range between 0.12 and 0.25. Such low values provide evidence for the absence of any chance of correlation in observed model.

As it was mentioned previously, we employed GA method to find the important descriptors and then we performed PLS and SVM for a linear and non-linear modeling of the dataset, respectively. Table S2 in the supplementary material shows that both methods were achieved for modeling Non-peptide HIV-1 protease inhibitors. We also obtained percent recovery of compounds to show which model was better in predicting the pIC_{50}.

Results of GA-PLS

Model validation was achieved through leave-one-out cross-validation (LOO CV) and external validation (for a test set), and the predictive ability was statistically evaluated through the root mean square errors of calibration (RMSEC) and validation (RMSECV). A leave-one-out cross-validation was carried out using the NIPALS algorithm to find the best number of latent variables (Lv) to be used in calibration and prediction. The calibration and prediction qualities were quantified with R^2 (training set) and Q^2 (leave-one-out cross-validation on training set). The Lv's were chosen according to their Q^2 or the prediction error sum of squares (PRESS) values for cross-validated models. Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the PRESS and SSY (sum of squares of deviations of the experimental values from their mean) statistics, the Q^2 can be easily calculated by eqn 14:

Table 2: Description of the selected descriptors in this study

No	Symbol	Class	Meaning
1	TE2	Charge descriptors	Topographic electronic descriptor(bond resctricted)
2	Ui	Empirical descriptors	Unsaturation index
3	GATS5e	2D autocorrelations	Geary autocorrelation-lag5/weighted by atomic Sanderson electronegativities
4	Mor13m	3D-MoRSE descriptors	3D MoRSE-signal13/weighted by atomic masses
5	ATS7m	2D autocorrelations	Broto-Moreau autocorrelation of a topological structure- lag7/ weighted by atomic masses
6	Ss	Constitutional descriptors	Sum of Kier-Hall electrotopological States
7	Mor27e	3D-MoRSE descriptors	3D MoRSE-signal27/ weighted by atomic Sanderson electronegativities
8	RDF035e	RDF descriptors	Radial distribution function-3.5/ weighted by atomic Sanderson electronegativities

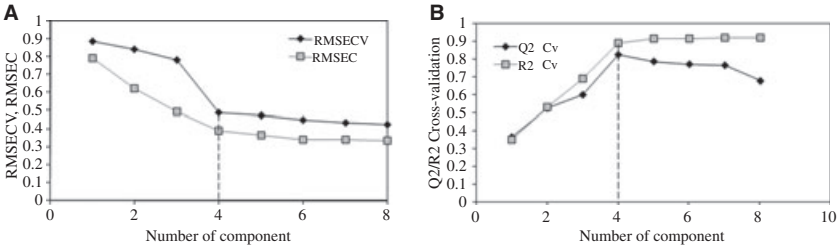

Figure 2: A) Root mean square errors of calibration (RMSEC) and validation (RMSECV) as a function of number of components. B) Cross-s-validation (Q^2/R^2 as a function of the number of components.

$$Q^2_{LOO} = \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n} (y_{exp} - y_{pred})^2}{\sum_{i=1}^{n} (y_{exp} - \bar{y})^2} \qquad (14)$$

The cross-validation method employed was to eliminate only one sample at a time and then PLS calibrates the remaining standard descriptor. By using this calibration, the pIC_{50} of the sample left out was predicted. This process was repeated until each standard had been left out once. One other reasonable choice for the optimum number of factors would be that number that yields the minimum RMSEC. As there are a finite number of samples in the training set, in many cases, the minimum RMSEC value causes underfitting or overfitting for unknown samples that were not included in the model. A solution to this problem has been suggested by Haaland *et al.* and Goodarzi *et al.* (42,43) in which the RMSEC values for all previous factors are compared to the RMSEC value at the minimum. The F-Statistical test can be used to determine the significance of RMSEC values greater than the minimum. In all instances, the number of factors for the first RMSEC values whose F-ratio probability drops below 0.75 was selected as the optimum. We tested two ways for finding the best Lv to make a model where both of them present that four is the best number of component. Figure 2A shows the root mean square errors of calibration (RMSEC) and validation (RMSECV) as a function of number of components, while Figure 2B shows the Q^2/R^2 cross-validation as a function of the number of components. It is clear from Figure 2A that the minimum number of components is four which is the same maximum number of components in Figure 2B. The following equation represents the best model achieved by GA-PLS:

$$\begin{aligned} pIC_{50} = {} & -3.405737(\pm 1.447) + 0.525607(\pm 0.052)TE2 \\ & + 0.911090(\pm 0.236)Ui + 2.586873(\pm 0.369)GATS5e \\ & - 47.069316(\pm 8.558)Mor13e - 0.207581(\pm 0.027)ATS7m \\ & + 13.338116(\pm 3.599)Ss - 0.001142(\pm 0.000)Mor27e \\ & + 49.494231(\pm 7.841)RDF035e \end{aligned} \qquad (15)$$

The best model shown above reveals that the most significant contribution comes from the RDF035e.

Table 3: Results and statistical parameters of GA-PLS and GA-SVM

Parameters		GA-SVM	GA-PLS
NOC[a]			4
Q^2 LOO[b]		0.9672	0.8259
σ		0.5	
ε		0.06	
C		8	
RMSEP	Training set	0.2027	0.3934
	Test set	0.2751	0.3962
RSEP(%)	Training set	3.1520	6.1156
	Test set	4.0216	5.7928
MAE(%)	Training set	6.5080	8.9351
	Test set	18.093	21.745
R^2	Training set	0.9800	0.8935
	Test set	0.9355	0.8603
F statistical	Training set	1815.2	310.26
	Test set	72.481	30.792
t test	Training set	42.606	17.614
	Test set	8.5136	5.5491

RMSEP, root mean square error of prediction, RSEP, relative standard error of prediction, MAE, mean absolute error.
[a]Number of components.
[b]Q^2 Leave-one-out cross-validation.

Results of GA-SVM

The quality of SVM for regression depends on several parameters namely, kernel type k, which determines the sample distribution in the mapping space, and its corresponding parameter σ, capacity parameter C, and ε-insensitive loss function. The three parameters were optimized in a systematic grid search-way and the final optimal model was determined (see Table 3). Optimization of SVM parameters was performed by systematically changing their values in the training step and calculating the RMSE of the model using 5-fold cross-validation. 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

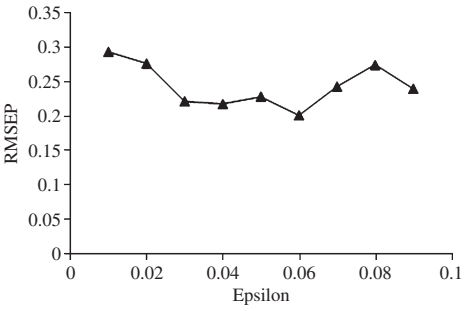

Figure 3: Variation of RMSE versus epsilon values.

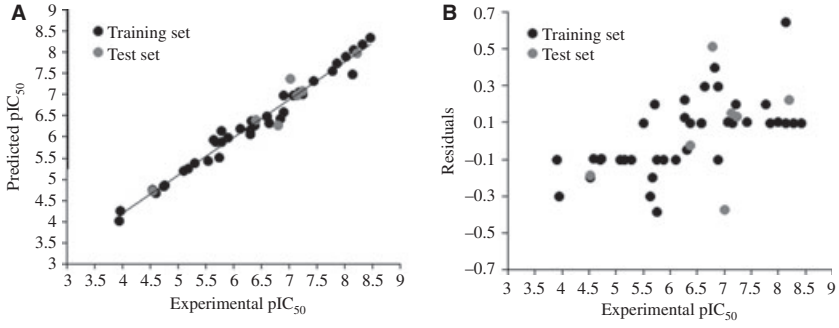

Figure 4: A) Calculated pIC50 against the experimental values using GA-SVM. B) Residual values against experimental pIC50 using GA-SVM.

to select an optimal value for ε, there will be some practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set from meeting boundary conditions and allows the possibility of sparsity in the dual formulations solution. Therefore, choosing the appropriate value of ε is a critical step. To find an optimal value for ε, the RMSE of SVM models with different ε values was calculated. The variation of RMSE versus epsilon values is plotted in Figure 3. As shown in this figure, the optimal value of ε was 0.06. The other parameter is the regularization parameter C that controls the trade-off between maximizing the margin and minimizing the training error. If C value is too small, then insufficient stress will be placed on fitting the training data. On the other hand, if C value is too large, then the SVM model will overfit the training data. To find an optimal value of C, the RMSE of SVM models with different C values was calculated. Moreover, to inspect any interactions between C and epsilon, after optimization of the C value, the epsilon value varied. The results indicate that the value of optimized epsilon did not vary at this stage which illustrates their independency.

Six general statistical parameters were selected to evaluate the prediction ability of the constructed model. These parameters are: root mean square error of prediction (RMSEP), relative standard error of prediction (RSEP), mean absolute error (MAE), square of correlation coefficient (R^2), F-statistical and t test. Table 3 shows the results of GA-PLS and GA-SVM and the calculated statistical parameters. This table shows that the results of the GA-SVM are better than GA-PLS. Figure 4A shows calculated pIC50 against experimental values, while Figure 4B shows their residual values against the experimental pIC50 using GA-SVM. Figure 5A shows the calculated pIC50 against the experimental values, while Figure 5B shows their residual values against the experimental pIC50 using GA-PLS. Figure 1 and Table 1 shows that the inhibitors consist of some different classes with very diverse substituents which could help explain the models and the selected descriptors obtained to predict the inhibitory activity of non-peptide HIV-1 protease inhibitors. For example, selection of topographic electronic descriptor (TE2) reflects to some extent, differences in size, shape, and constitution. Such quantities affect the electronic charge distribution and interatomic distance of the molecules. The empirical descriptor (U_i) represents limited subsets of compounds and cannot be extended to classes of compounds different from those for which they were defined (compounds 1–5 in Table 1). These empirical descriptors are related to specific or local structural factors present in the molecules. The other selected descriptors are related to the geometry of molecules such as 3D atomic coordinates, atomic property, and distance distribution in the geometrical representation of a molecule. In addition, these non-peptide HIV-1 protease inhibitors have shown their ability to participate in complex biological phenomena. It is worth mentioning to the best of our knowledge that no QSAR model was found for this class up to now. Therefore, the development of a robust and interpretable QSAR model that is able to accurately predict the pIC50 is necessary and this article is the first to report such a model.

Conclusion

PLS and SVM were used to develop QSAR models for the prediction of the inhibitory activity of non-peptide HIV-1 protease inhibitors. The built models clearly demonstrate good correlations between the structure and inhibitory activity of the studied compounds. Eight descriptors were selected with genetic algorithm. The

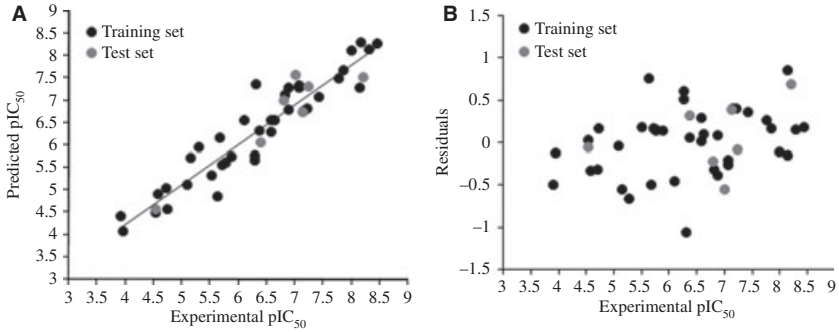

Figure 5: A) Calculated pIC50 against the experimental values using GA-PLS. B) Residual values against experimental pIC50 using GA-PLS.

selected descriptors, which are TE2, Ui, GATS5e, Mor13e, ATS7m, Ss, Mor27e, and RDF035e were found to be important factors controlling the inhibitory activity. Comparison between PLS and SVM methods demonstrates that the performance of SVM model is better than that of PLS, which indicates that the non-linear model is able to describe the relationship between the structural descriptors and the inhibitory activity more accurately. The proposed models will help identifying new HIV-1 protease inhibitors and provide insight to guide their development.

Acknowledgments

The authors would like to acknowledge the computational chemistry laboratory at Al-Quds University for providing Matlab software and for the time dedicated for performing the calculations of the study.

References

1. Kohl N.E., Emini E.A., Schleif W.A., Davis L.J., Heimbach J.C., Dixon R.A., Scolnick E.M., Sigal I.S. (1988) Active human immunodeficiency virus protease is required for viral infectivity. Proc Natl Acad Sci USA;85:4686–4690.
2. McQuade T.J., Tomasselli A.G., Liu L., Karacostas V., Moss B., Sawyer T.K., Heinrikson R.L., Tarpley W.G. (1990) A synthetic HIV-1 protease inhibitor with antiviral activity arrests HIV-like particle maturation. Science;247:454–456.
3. Davies D.R. (1990) The structure and function of the aspartic proteinases. Annu Rev Biophys Biophys Chem;19:189–215.
4. Greenlee W.J. (1990) Renin inhibitors. Med Res Rev;10:173–236.
5. Wlodawer A., Erickson J.W. (1993) Structure-based inhibitors of HIV-1 protease. Annu Rev Biochem;62:543–585.
6. Appelt K.. (1993) Therapeutic approaches to HIV. In: Anderson P.S., Kenyon G.L., Marshall G.R., Editors. Perspectives in Drug Discovery and Destgn. Leiden: ESCOM Science, 1, pp. 23–48.
7. Huff J.R. (1991) HIV protease: a novel chemotherapeutic target for AIDS. J Med Chem;34:2305–2664.
8. West M.L., Fairlie D.P. (1995) Targeting HIV-1 protease: a test of drug-design methodologies. Trends Pharmacol Sci;1:67–75.
9. Lain P.Y., Jadhav P.K., Eyermann C.J., Hodge C.N., Ru Y., Bacheler L.T., Meek J.L., Otto M.J., Rayner M.M., Wong Y.N., Chang C.H., Weber P.C., Jackson D.A., Sharpe T., Ericksonviitanen S. (1994) Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science;263:380–384.
10. Deeb O., Hemmateenejad B., Jaber A., Garduno-Juarez R., Miri R. (2007) Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic-PLS. Chemosphere;67:2122–2130.
11. Deeb O., Hemmateenejad B. (2007) ANN-QSAR Model of Drug-binding to Human Serum Albumin. J Chemical Biology & Drug Design;70:19–29.
12. Ramírez-Galicia G., Garduño-Juarez R., Hemmateenejad B., Deeb O., Estrada-Soto S. (2007) QSAR Study on the Relaxant Agents from Some Mexican Medicinal Plants and Synthetic Related Organic Compounds. J Chemical Biology & Drug Design;70:143–153.
13. Ramírez-Galicia G., Garduño-Juárez R., Hemmateenejad B., Deeb O., Deciga-Campos M., Moctezuma-Eugenio J.C. (2007) QSAR Study on the Antinociceptive Activity of Some Morphinans. J Chemical Biology & Drug Design;70:53–64.
14. Deeb O., Youssef K.M., Hemmateenejad B. (2007) QSAR of Novel Hydroxyphenylureas as Antioxidant Agents. QSAR Comb Sci;27:417–424.
15. Ramírez-Galicia G., Garduño-Juárez R., Deeb O., Hemmatenenejad B. (2008) MLR–ANN and RTO Approach to μ-opioid Receptor-binding Affinity. Pooling Data from Different Sources. J Chemical Biology & Drug Design;71:260–270.
16. Goodarzi M., Freitas M.P. (2008) Augmented three-mode MIA-QSAR modelling for aseries of anti-HIV-1 compounds. QSAR Comb Sci;27:1092–1098.
17. Byvatov E., Fechner U., Sadowski J., Schneider G. (2003) Comparison of support vector machine and artificial neural network Systems for drug/nondrug classification. J Chem Inf Comput Sci;43:1882–1889.
18. Niani C., Wencong L., Jie Y., Gozheng L. (2004) Support Vector Machine in Chemistry. Shanghai: World Scientific Publishing Co. Pet. Ltd.
19. Liu H.X., Zhang R.S., Luan F., Yao X.J., Liu M.C., Hu Z.D., Fan B.T. (2003) Diagnosing Breast Cancer Based on Support Vector Machines. J Chem Inf Comput Sci;43:900–907.
20. Burbidge R., Trotter M., Buxton B., Holden S. (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem;26:5–14.
21. Liu H.X., Zhang R.S., Yao X.J., Liu M.C., Hu Z.D., Fan B.T. (2003) QSAR study of ethyl 2-[(3-Methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-1 and NF-κB mediated gene expression based on support vector machines. J Chem Inf Comput Sci;43:1288–1296.
22. Fatemi M.H., Gharaghani S. (2007) A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine. Bioorg Med Chem;15:7746–7754.
23. Liu H.X., Zhang R.S., Yao X.J., Liu M.C., Hu Z.D., Fan B.T. (2004) Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J Chem Inf Comput Sci;44:161–167.
24. Tummino P.J., Prasad J.V.N.V., Ferguson D., Nouhan C., Graham N., Domagala J.M., Ellsworth E. *et al.* (1996) Discovery and optimization of nonpeptide HIV-1 protease inhibitors. Bioorg Med Chem;4:1401–1410.
25. Gunn S.R. (1997) Support Vector Machines for Classification and Regression. UK: University of Southampton.
26. Dewar M.J.S., Zeoblish E.G., Healy E.F., Stewart J.J. (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc;107:3902–3909.
27. Goodarzi M., Freitas M.P., Jensen R. (2009) Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3- (3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regression. Chemom Intell Lab Sys;98:123–129.
28. Leardi R., Boggia R., Terrile M. (1992) Genetic algorithms as a strategy for feature selection. J Chemom;6:267–281.

29. Hunger J., Huttner G. (1999) Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. J Comput Chem;20:455–471.
30. Ahmad S., Gromiha M.M. (2003) Design and training of a neural network for predicting the solvent accessibility of proteins. J Comput Chem;24:1313–1320.
31. Waller C.L., Bradley M.P. (1999) Development and validation of a novel variable selection technique with application to multidimensional quantitative structure–activity relationship studies. J.Chem.Inf.Comput.Sci;39:345–355.
32. Aires-de-Sousa J., Hemmer M.C., Gasteiger J. (2002) Prediction of ^{1}H NMR chemical shifts using neural networks. Anal Chem;74:80–90.
33. Goodarzi M., Freitas M.P. (2008) Predicting boiling points of aliphatic alcohols through multivariate image analysis applied to quantitative structure–property relationships. J Phys Chem A;112:11263–11265.
34. Vapnik V.N. (2000) The Nature of Statistical Learning Theory. New York: Springer.
35. Feng Pai P., Lin K.P., Lin C.S., Chang P.T. (2010) Time series forecasting by a seasonal support vector regression model. Expert Syst Appl;in press. doi:10.1016/j.eswa.2009.11.076.
36. Xinjun P. (2010) TSVR: An efficient twin support vector machine for regression. Neural Networks;in press. doi:10.1016/j.neunet.2009.07.002.
37. Lu C.J., Lee T.S., Chiu C.C. (2009) Financial time series forecasting using independent component analysis and support vector regression. Decision Support Syst;47:115–125.
38. Kuhn H.W., Tucker A.W. (1951) Nonlinear programming. Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press, 481–492.
39. Cristianini N., Shawe-Taylor J. (2000) An Introduction to Support Vector Machines and other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press.
40. Tute M..(1990) History and objectives of quantitative drug design in advances in drug research In: Sammes P., Taylor J., editors. Comprehensive Medicinal Chemistry. Oxford: Pergamon, 4, pp. 1–32.
41. Hansch C., Taylor J., Sammes P., (1990) Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds. New York: Pergamon, 6, pp. 1–19.
42. Haaland D.M., Thomas E.V. (1988) Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal Chem;60:1193–1202.
43. Goodarzi M., Goodarzi T., Ghasemi N. (2007) Spectrophotometric simultaneous determination of manganese(II) and Iron(II) in pharmaceutical by orthogonal signal correction-partial least squares. Ann Chim;97:303–312.

Notes

[a]HyperChem Release 7.5, HyperCube, Inc, <http://www.hyper.com>.
[b]DRAGON 5.0 evaluation version, <http://www.disat.unimib.it/vhml>.

Supporting Information

Additonal Supporting Information may be found in the online version of this article:

Table S1. Correlation matrix between selected descriptor.

Table S2. Observed pIC_{50} versus predicted using GA-PLS & GA-SVM models as well as recovery (%).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.