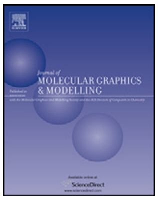
Journal of Molecular Graphics and Modelling 28 (2010) 852–862

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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



Hybrid-genetic algorithm based descriptor optimization and QSAR models for predicting the biological activity of Tipranavir analogs for HIV protease inhibition

A. Srinivas Reddy[a,*], Sunil Kumar[b], Rajni Garg[c]

[a] *Department of Biomedical Engineering, University of California-Davis, Davis, USA*
[b] *Electrical and Computer Engineering Department, San Diego State University, San Diego, CA 92182, USA*
[c] *Computational Science Research Center, San Diego State University, San Diego, CA 92182, USA*

ARTICLE INFO

Article history:
Received 2 December 2009
Received in revised form 4 March 2010
Accepted 14 March 2010
Available online 24 March 2010

Keywords:
Artificial neural networks
Genetic algorithm
Regression analysis
Decision trees
HIV-1 protease inhibitor

ABSTRACT

The prediction of biological activity of a chemical compound from its structural features plays an important role in drug design. In this paper, we discuss the quantitative structure activity relationship (QSAR) prediction models developed on a dataset of 170 HIV protease enzyme inhibitors. Various chemical descriptors that encode hydrophobic, topological, geometrical and electronic properties are calculated to represent the structures of the molecules in the dataset. We use the hybrid-GA (genetic algorithm) optimization technique for descriptor space reduction. The linear multiple regression analysis (MLR), correlation-based feature selection (CFS), non-linear decision tree (DT), and artificial neural network (ANN) approaches are used as fitness functions. The selected descriptors represent the overall descriptor space and account well for the binding nature of the considered dataset. These selected features are also human interpretable and can be used to explain the interactions between a drug molecule and its receptor protein (HIV protease). The selected descriptors are then used for developing the QSAR prediction models by using the MLR, DT and ANN approaches. These models are discussed, analyzed and compared to validate and test their performance for this dataset. All three approaches yield the QSAR models with good prediction performance. The models developed by DT and ANN are comparable and have better prediction than the MLR model. For ANN model, weight analysis is carried out to analyze the role of various descriptors in activity prediction. All the prediction models point towards the involvement of hydrophobic interactions. These models can be useful for predicting the biological activity of new untested HIV protease inhibitors and virtual screening for identifying new lead compounds.

1. Introduction

HIV protease (HIV-PR) is one of the major viral targets for the development of new chemotherapeutics. Currently, many HIV-PR inhibitor drugs are used in combination with other drugs [1–4]. However, the use of current drug regimens is compounded by several issues such as – adherence, tolerability, long-term toxicity, and drug- and cross-resistance. Besides, the mutations also enable HIV to resist currently available treatments [5]. Therefore, there is a continuing need for the development of new chemotherapeutics with improved antiviral potency and favorable pharmacokinetic profile. One way of designing the effective inhibitors is modeling the biological activity to propose new candidate molecules. This can be partially fulfilled by understanding the existing structure–activity relationship (SAR) data to develop QSAR prediction models.

The process of relating the molecular structure of a chemical compound to its biological activity, ADME (absorption, distribution, metabolism, and excretion) properties, or to its chemical reactivity is important in drug design [6–9]. With increase in the amount and complexity of available chemical and biological data, the development of new computational models is becoming increasingly important, for understanding and predicting the interactions between drug molecules and their receptor proteins. The quantitative structure activity relationship (QSAR) models have shown great promise for handling this massive amount of structural and biological data. These models relate the descriptors (also known as parameters or features) of a small molecule or compound, computed based on its physicochemical properties, to its biological activity [10–12]. QSAR is a well known approach for identifying the lead compounds from an existing database of compounds with the known biological activity [12–14]. The QSAR models are developed using linear regression techniques (e.g., multiple linear regression (MLR) [15], partial least squares (PLS)) [16] or non-linear machine

[*] Corresponding author at: Department of Biomedical Engineering, University of California-Davis, 2316 Genome and Biomedical Sciences Building, 451 Health Sciences Drive, Davis, CA 95616-5294, USA. Tel.: +1 5303833763; fax: +1 5307545739.
E-mail addresses: sralla@ucdavis.edu, asvreddy@gmail.com (A.S. Reddy).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.03.005

learning techniques (e.g., artificial neural network (ANN) [17–19], support vector machine (SVM) [20] or decision trees (DT)) [21].

A large number of descriptors are usually computed for a small set of molecules. However, a good descriptor set should contain the descriptors that are highly correlated with the target, yet uncorrelated with each other. Optimization of descriptor set and selecting an appropriate statistical or machine learning technique plays a major role in developing the robust QSAR prediction models. The feature optimization techniques are used to remove the irrelevant and correlated descriptors. The genetic algorithm (GA), which belongs to the class of evolutionary algorithms, has been widely used for feature optimization in QSAR models [22,23]. Solutions generated by GA have less probability of being affected by local minima due to the use of inheritance, mutation, selection, and crossover [24]. Since GA does not carry out the fitness evaluation of the population, different types of fitness functions are used for this purpose, including the MLR [15], partial least square (PLS) [25], correlation-based feature selection (CFS) [26–28], DT [21] and ANN [17–19]. The selected descriptors are then used as input variables for developing QSAR model(s). Similarly, the clustering techniques (e.g., K-means clustering, K-nearest neighborhood) are used for removing the outlier compounds or descriptors [29–31]. The compounds which posses unexpected biological activities and unable to fit in a QSAR model can be treated as outliers. The presence of outliers is not only due to the possibility that the molecules may act by different mechanisms or interact with the receptor in different binding modes but also due to the intrinsic noise associated with both the original data and methodological aspects involved in the construction of a QSAR model [32]. The outliers are removed using various statistical techniques and clustering methods to improve the predictive ability of the developed models by several known researchers [33]. This reduces the dimensionality of data and allows learning algorithms to operate faster and more effectively. In some cases, the classification accuracy can be improved, while in others, the result is a more compact prediction model with easy interpretation.

QSAR models for HIV studies: Many MLR based QSAR models have been used to model the activity of inhibitors of HIV proteins, including models developed by our coauthor Garg et al. [34,35]. Boiani et al. used a DT method to develop QSAR models for N-Oxide containing heterocycles compounds for anti-Trypanosoma cruzi activity [36]. Daszykowski et al. demonstrated the application of CART (Classification and Regression Trees) for the analysis of biological activity of non-nucleoside reverse transcriptase inhibitors (NNRTIs) for HIV reverse transcriptase [37].

Similarly, the ANN QSAR models have also been widely used to predict HIV drug resistance [17,18], to elicit structural information about viral enzymes [38], and to predict the activity of potential drugs [39,40]. Wang and Larder [17] used a three-layer ANN to predict Lopinavir resistance. Draghici et al. [18] studied the HIV protease resistance to drugs (e.g., Indinavir and Saquinavir) by considering the structural features of the HIV protease–drug inhibitor complex as descriptors. Yang and Thomson [38] developed bio-basis function ANNs to predict the protease cleavage sites in proteins. Douali et al. [39] developed the QSAR models for HEPT derivatives by both the linear regression and ANN techniques. From the models obtained using linear regression techniques, they estimated the contribution of each descriptor to the model, and confirmed the hydrophobic requirements for HIV inhibition. Hecht et al. [41] applied the pre-clustering and evolved neural networks to develop prediction models for high-throughput screening of anti-HIV compounds.

The QSAR models, developed using MLR [39,42] or partial least squares (PLS) [25] techniques, are easy to interpret and can provide useful insight into drug-receptor interactions. On the other hand, the QSAR models developed using non-linear techniques (e.g., ANN) have better predictive power but do not provide straightforward interpretation [17,18,38]. The decision tree based QSAR models fall in between the linear and non-linear models in terms of their predictive and interpretation abilities. They are more transparent, easy to understand and convert to a set of prediction rules [21,36–39]. Therefore, *we have used all three above-mentioned approaches in order to develop robust QSAR prediction models for the HIV protease inhibitor dataset, which have good predictive power as well as interpretability*.

In this paper, we discuss the QSAR prediction models developed on a cycloalkylpyranone dataset of HIV protease enzyme inhibitors from which Tipranavir, a U.S. FDA approved HIV protease inhibitor drug was developed [5]. This dataset was developed in-house and curated for quality assurance. For descriptor optimization, we use four variations of hybrid-GA techniques, in which GA is used for searching the descriptor subspace whereas the MLR, CFS, DT and ANN are used for fitness evaluation. The QSAR prediction models are developed using three approaches – MLR, ANN and DT. We use the weight analysis to interpret the importance of descriptors in ANN models. The *major objectives of this study* are: (i) to study and analyze the reduced descriptor sets obtained by the hybrid-GA feature optimization techniques, (ii) *to* develop *robust* QSAR models for biological activity prediction of HIV protease inhibitors, and (iii) to compare the performance of QSAR models developed by using three different classes of techniques, in terms of their predictive power as well as interpretability.

Remainder of the paper is organized as follows: The dataset construction, descriptor computation and selection, QSAR prediction model development and validation are discussed in Section 2. Results of descriptor selection and QSAR prediction model development, including the interpretation of selected descriptors and analysis of developed models, are discussed in Section 3. Conclusion is given in Section 4.

2. Methods

The methodology adopted for this research is illustrated in Fig. 1 and discussed below.

2.1. Dataset construction

The dataset used for this study consists of SAR biological activity (K_i) data of 170 cycloalkylpyranone analogs. Tipranavir, one of the cycloalkylpyranone artifacts is a U.S. FDA approved HIV protease inhibitor drug to treat HIV infection and AIDS [5]. Cycloalkylpyranones are the non-peptidic lead structures obtained by the chemical modification of coumarin. The benzene ring of coumarin was replaced by conformationally-flexible cycloalkyl rings of various sizes. The detailed description of the structural modifications of the data set is illustrated in a roadmap of structural modification (Fig. 2). Here the biological activity (K_i) is the inhibition constant for a drug, which represents the concentration of competing ligand in a competition assay that would occupy 50% of the receptors if no radio-ligand were present.

The data were taken from the published literature and the dataset is made available in supplementary information. The similar compounds published in separate tables or in series of articles were combined, and the repeated data were omitted. In our case, enantiomers and racemic mixtures were treated as outlier compounds, and we decided not to include these compounds in the model development. These compounds differ only optical properties and possess equivalent physicochemical descriptor values; hence the developed statistical models by including them would be biased towards predicting those compounds. The compounds in the dataset have structural variations in their side chains, aromatic groups and heterocyclic rings.

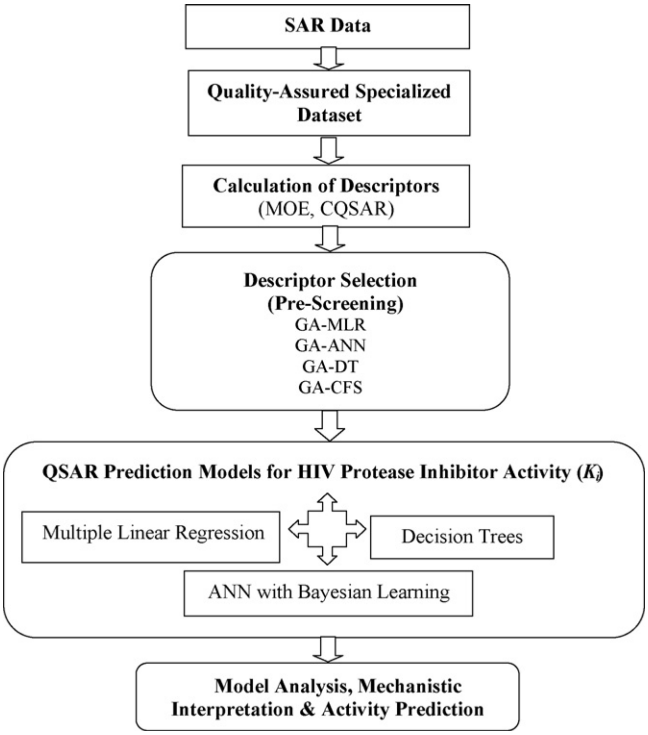

SAR Data
Quality-Assured Specialized Dataset
Calculation of Descriptors
(MOE, CQSAR)
Descriptor Selection
(Pre-Screening)
GA-MLR
GA-ANN
GA-DT
GA-CFS
QSAR Prediction Models for HIV Protease Inhibitor Activity (K_i)
Multiple Linear Regression
Decision Trees
ANN with Bayesian Learning
Model Analysis, Mechanistic Interpretation & Activity Prediction

Fig. 1. Flow chart depicting the proposed methodology.

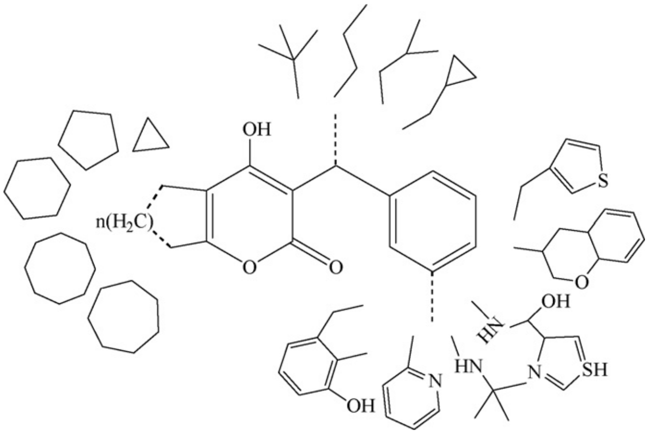


Fig. 2. Structural modifications of cycloalkylpyranone scaffold.

Descriptor calculation: All the compounds were optimized and minimum energy conformations were generated for each compound using MOE (Molecular Operating Environment) software [43]. The 233 descriptors were calculated for each compound using the descriptor generation module of MOE software [43]. These descriptors include both 2D and 3D descriptors and represent various topological (such as topological indices, structural keys, E-state indices, connectivity and shape indices), physicochemical (such as partition coefficient, molecular weight and molar refractivity), and electrostatic (such as partial charges, topological polar surface area and van der Wall surface area) properties as depicted in Fig. 3. Three additional descriptors, namely ClogP (calculated octanol/water partition coefficient), CMR (calculated molar refractivity) and MgVol (McGowan volume), were also calculated using the CQSAR software [44]. These CQSAR descriptors have been found important in many HIV protease QSAR prediction models [25,26].

Pre-screening: In the first step, the compounds and descriptors with missing (or null) values were removed from the dataset. Next, we employed the identity test using Matlab [45] and removed 15 descriptors, which had 90% or more zero values. Finally, the 93 descriptors were removed by the pair-wise correlation analysis using the Matlab, in which the descriptors with more than 90% correlation in their values were identified and one of the correlated descriptors was removed. Since a range of descriptors are being calculated and added for a given dataset by the commercial programs used for calculating these descriptors, among which several of them are redundant or had insignificant difference in their physical reasonableness, it becomes important to decrease the descriptor space by removing the redundancy as suggested by researchers [46,47]. This pre-screening gave us a *quality-assured dataset* of 155 compounds, each with 128 descriptor values which are used for further analysis as discussed below.

2.2. Descriptor set optimization

As discussed in Section 1, identifying a small subset of descriptors, which represent the total set of descriptors, is important for developing a good QSAR prediction model. We have used the hybrid-GA optimization technique, where GA is used for searching the descriptor subspace, whereas the MLR, CFS, DT or ANN is used for fitness evaluation. GA is governed by biological evolution rules and can investigate several possible solutions simultaneously, each of which explores different regions in the descriptor space. Fitness of each solution is evaluated by one of the above-mentioned techniques. The MLR and DT are the linear fitness functions whereas ANN is a non-linear function. The use of ANN can handle the optimization of non-linear descriptor space more efficiently [19]. The major steps employed in the hybrid-GA scheme are illustrated in Fig. 4.

The *first* step in the hybrid GA is to create a population of N individuals (feature subsets). Each individual encodes the same number of randomly chosen descriptors, and the fitness of each individual in this generation is determined. The compounds in the dataset were divided into a training (66%) and test set (34%). The test set is not used during training but serves to test the predictive ability of final models. Here, the root mean squared error (RMSE) is taken as the

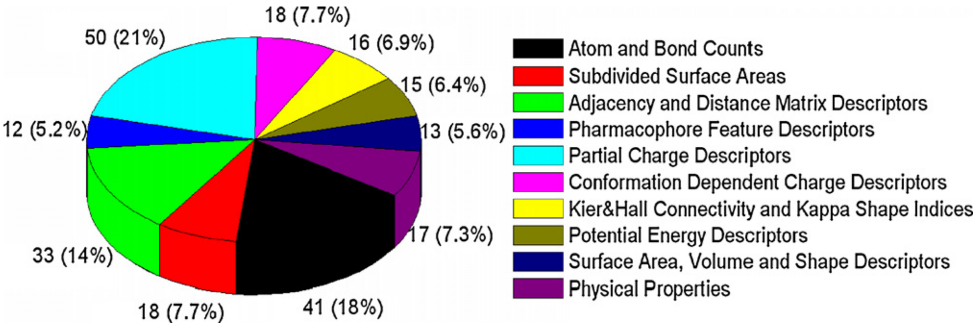


Fig. 3. Pie chart depicting distribution of various categories of 233 MOE [39] descriptors.

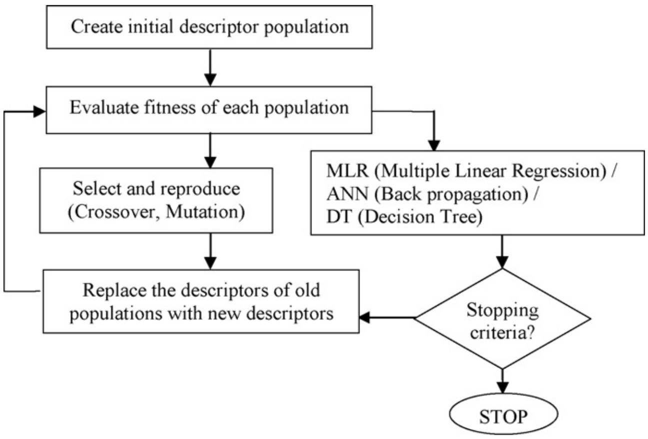

Create initial descriptor population
Evaluate fitness of each population
Select and reproduce (Crossover, Mutation)
MLR (Multiple Linear Regression) / ANN (Back propagation) / DT (Decision Tree)
Replace the descriptors of old populations with new descriptors
Stopping criteria?
STOP

Fig. 4. Schematic representation of hybrid GA algorithm.

fitness measure. *Next*, a fraction of children of the next generation is produced by crossover and the rest by mutation from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from one or both parents, and is evaluated for fitness. The cycle continues for a predetermined number of generations, or until the results do not change continuously for a specified number of generations [24].

The GA-MLR, GA-DT and GA-CFS techniques are implemented using WEKA program [48], and the GA-ANN is implemented in MATLAB. The values of various hybrid-GA parameters and measures are discussed in Section 3.1. Since the MLR, DT and ANN are also used for developing the QSAR models, they are discussed in the next section, whereas the GA-CFS is discussed below.

GA-CFS: As shown in Fig. 5, the correlation-based feature selection (CFS) algorithm evaluates each feature subset by considering the individual predictive ability of each feature along with the degree of redundancy between them, and returns a numeric measure that guides the search [26,27]. The CFS fitness function takes into account the usefulness of individual features for predicting the activity along with the level of inter-correlation to give the goodness of feature subsets. It has wide range of applications for feature selection including QSAR [28,29].

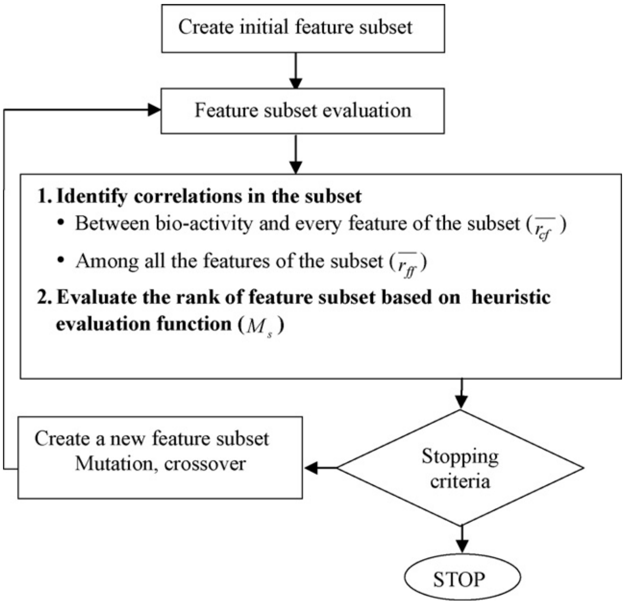

Create initial feature subset
Feature subset evaluation
1. Identify correlations in the subset
• Between bio-activity and every feature of the subset ($\overline{r_{cf}}$)
• Among all the features of the subset ($\overline{r_{ff}}$)
2. Evaluate the rank of feature subset based on heuristic evaluation function (M_s)
Create a new feature subset Mutation, crossover
Stopping criteria
STOP

Fig. 5. Schematic representation of GA-CFS algorithm.

CFS is a simple filter algorithm that ranks feature subsets according to a correlation-based heuristic evaluation function (see Eq. (1)). The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and yet uncorrelated with each other.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

Here M_s is the heuristic "merit" of a feature subset S (fitness function) containing k features, $\overline{r_{cf}}$ is the average feature-activity correlation ($f \in S$), and $\overline{r_{ff}}$ is the average feature–feature inter-correlation [27].

The numerator of (1) provides an indication of how predictive of the class a set of features are. The denominator provides an indication of the redundancy among the features. Eq. (1) forms the core of CFS and imposes a ranking on feature subsets in the search space of all possible feature subsets. To measure the feature–class correlations ($\overline{r_{cf}}$) and feature–feature inter-correlations ($\overline{r_{ff}}$), we adopt the *symmetrical uncertainty* (*U*), which uses a modified *information gain* (*InfoGain*) measure (see Eqs. (2a), (2b) and (3)) [22]. The *symmetrical uncertainty* is defined for features *X* and *Y* as follows:

$$U = 2.0 \times \left[\frac{InfoGain}{H(Y) + H(X)}\right] \tag{2a}$$

where $H(Y)$ is entropy of Y given as,

$$H(Y) = -\sum_{y \in Y} p(y)\log(p(y)) \tag{2b}$$

and $InfoGain = H(Y) - H(Y/X)$

$$InfoGain = -\sum_{y \in Y} p(y)\log(p(y)) + \sum_{x \in X} p(x) \sum_{y \in Y} p\left(\frac{y}{x}\right) \log p\left(\left(\frac{y}{x}\right)\right) \tag{3}$$

2.3. QSAR prediction models

The descriptors obtained after feature optimization are used to develop prediction models. We have used three different approaches to develop QSAR prediction models – linear MLR, non-linear DT, and ANN as discussed below. These belong to different classes and would help us in developing the robust QSAR prediction models for this dataset.

Multiple linear regression (*MLR*): The MLR models serve as the basis for a number of multivariate methods. They establish a quantitative relationship between a group of predictor variables (*X*) and a response *Y* as shown in Eq. (4). This relationship is useful for understanding which predictors have the greatest effect and the direction of the effect [11].

$$Y = X\beta + \varepsilon \tag{4}$$

Here Y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of regressors, β is a $p \times 1$ vector of parameters, and ε is an $n \times 1$ vector of normally distributed noise. The aim of this regression method is to estimate the $\bar{\beta} = (\beta_l, \ldots, \beta_p)$, by using $Min\sum_1^n (Y_i - \beta X_{i1} - \cdots - \beta X_{ip})^2$.

Although MLR is computationally simple and the prediction models give strong mechanistic interpretation, it is criticized for its lack of robustness in handling the non-linear data. It also has certain other limitations, especially when the number of variables is large, or when the degree of correlation between the variables (or samples) is large. We implemented MLR scheme in WEKA [48]. Various parameters and measures used for evaluating the fitness of the MLR models are discussed at the end of this section (see 'Over fitting and model validation').

Decision trees (*DT*): The DTs are widely used machine learning methods used in pharmaceutical industry for predicting the quantitative structure–activity relationships. They are known for their predictive ability and are easy to interpret. DT approximates the discrete-valued functions, is robust to noisy data and capable of learning disjunctive expressions [21].

DTs approach a classification or regression problem in divide-and-conquer fashion [21]. The decision trees, can classify both categorical and numerical data, but the output attribute must be categorical. There are no *a priori* assumptions about the nature of the data, but the multiple output attributes are not allowed. There are various classes of decision trees based on the process of construction, pruning methodology and their application [49]. The classification and regression trees (CART) [50] and random forests (RF) [51] are more widely used in drug design as well as QSAR. The model tree (i.e., M5 decision tree) is mainly used for developing models to predict the values [52]. Unlike other decision trees (e.g., CART), the M5 decision trees store multivariate linear models at their leaf nodes. So, they are analogous to piecewise linear functions, and hence are considered as non-linear in nature.

M5 decision trees are constructed by first using a decision tree induction algorithm to build the initial tree, and then a multivariate linear regression model is constructed for each node of the tree. Each linear model is simplified by eliminating the parameters (by using a greedy search method) to minimize its estimated error. Then pruning of the tree is performed by examining each non-leaf node of the model tree to assign the linear model with lower estimated error to leaf nodes. Lastly smoothing process is applied to improve the prediction accuracy of the tree.

The *advantage of the M5 tree over CART* is that it is much smaller than CART, the decision is clear and the regression functions do not normally involve too many variables [53]. We have used the model tree program implemented in WEKA [48]. Various parameters and measures used for evaluating the fitness of the DT models are discussed at the end of this section (see 'Over fitting and model validation').

Artificial neural network (*ANN*): The ANN is used to identify correlated patterns between the input and target values and can subsequently predict outcomes from fresh inputs [17–19]. The ANNs generally consist of a number of interconnected processing elements or neurons. The descriptors obtained by feature selection are used to build the inputs of the ANN. Each input is associated with some weights (w) and biases (b) depending on the relative importance of the input.

We used a 3-layer ANN with back propagation for fitness evaluation of the population (in hybrid GA-ANN) as well as for the QSAR prediction model development. We investigated the performance of different learning schemes (e.g., Bayesian [54], scaled conjugate gradient [55], gradient descent [56], etc.), and observed that the Bayesian regularization (BR) learning showed the best performance for feature selection as well as biological activity prediction of this dataset.

We used the Bayesian learning algorithm by MacKay [57], and Foresee and Hagan [54], implemented in Matlab ANN toolbox [45]. The training function in Bayesian learning algorithm updates the weight and bias values according to Levenberg–Marquardt optimization procedure [58]. It minimizes a combination of sum of squared errors generated by the outputs (E_D) and sum of squares of weights that reflect the connections of network (E_w), and then determines the correct combination in order to produce a network that generalizes well. Thus the performance index modification (F) involves taking into account the sum of squares of the network weights (Eq. (5)). Then optimization technique is used to minimize F.

$$F = \beta E_D + \alpha E_w \tag{5}$$

Here α and β are black box parameters and do not represent the momentum factor and learning rate.

The other network parameters, like number of hidden nodes and learning rate, were varied to build the best prediction model as discussed in more detail in Section 3.2. The model validation parameters to evaluate the fitness of the model are discussed below.

Over fitting and model validation: To avoid over fitting, we used the 10-fold cross validation method to estimate the fitness of the model and compare all three types of QSAR models (i.e., MLR, DT and ANN based models). In *10-fold cross validation method*, the data are partitioned into 10 sets of size $n/10$ each. Among them, nine sets are used for training and the remaining one set is used for testing. The procedure was repeated 10-times, and average accuracy was computed. *N-fold cross validation* technique is a proven standard cross validation method for model validation [59,60].

Over fitting in ANN is also avoided by using the BR training algorithm [61]. The oversized networks which can cause over fitting can also be avoided by pruning away unnecessary neurons or starting with a low number of neurons and then gradually increasing as necessary while watching the generalization performance [62]. Reducing the number of inputs is also beneficial, which was done with the feature optimization as discussed earlier in this section. Various parameters used in our simulation are discussed in Section 3.

For model validation, various measures such as correlation coefficient (R) (Eq. (6a)), root mean squared error (RMSE) (Eq. (6b)) and cross-validated R-square (R^2_{cv}) (Eq. (6c)) were used. These measures are computed from the difference between the experimental values (Y_i) and the predicted value ($\hat{Y}_i$) for the ith compound as discussed below.

The linear correlation coefficient (R), measures the strength and the direction of a linear relationship between (Y_i) and ($\hat{Y}_i$). Here n is the number of pairs of data.

$$R = \frac{n\sum_{i=1}^{n} Y_i\hat{Y}_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n}\hat{Y}_i}{\sqrt{\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]\left[n\sum_{i=1}^{n}\hat{Y}_i^2 - \left(\sum_{i=1}^{n}\hat{Y}_i\right)^2\right]}} \tag{6a}$$

The root mean squared error (RMSE) and cross-validated R-square (R^2_{cv}) values give the variance of predicted activity from the experimental activity. They are good measures to compare the fitness of the models and are defined below.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{6b}$$

$$R^2_{\text{cv}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{6c}$$

where $\bar{Y}$ is the mean value of the experimental dataset.

3. Results and discussion

In this section, we discuss the performance of the feature optimization techniques and QSAR prediction models on HIV-PR dataset. We also provide interpretation of the descriptors obtained from the feature optimization techniques and a comparative analysis of QSAR prediction models obtained from all three approaches.

3.1. Feature optimization

We used the hybrid-GA techniques (GA-MLR, GA-CFS, GA-DT and GA-ANN) to optimize the 128 descriptors, obtained after prescreening of 155 Tipranavir analogs as discussed in Section 2.1.

The GA parameters, like number of generations and population size, were varied and their performance was analyzed. Using

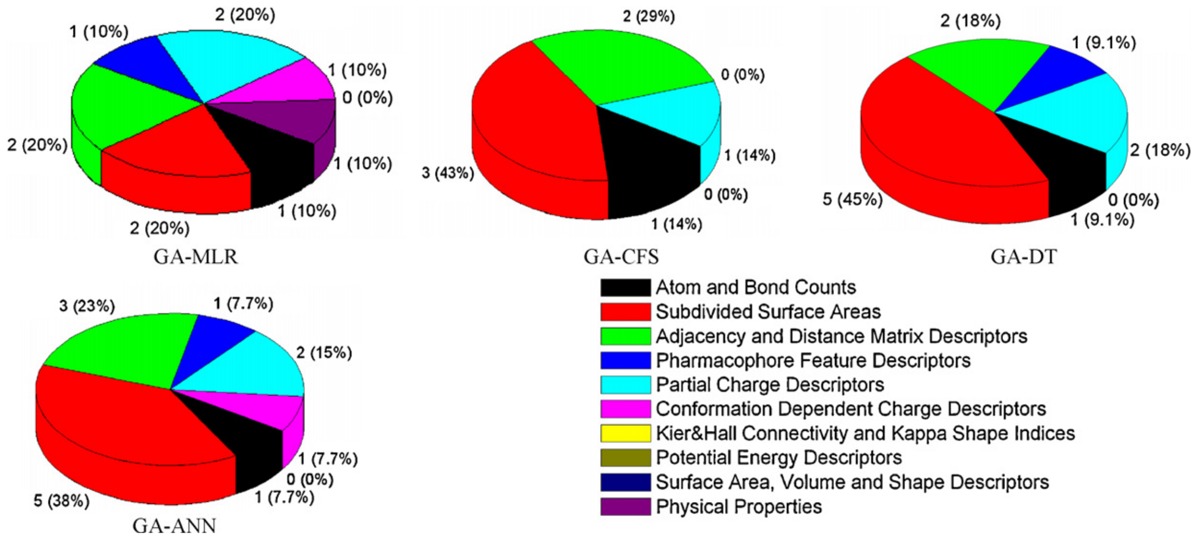


Fig. 6. Pie charts depicting distribution of various categories of descriptors (including CQSAR descriptor ClogP) obtained after feature optimization by GA-MLR, GA-CFS, GA-DT and GA-ANN methods.

the WEKA software, we observed that increasing the number of generations and population size beyond 20 did not improve the results for *GA-MLR*, *GA-CFS* and *GA-DT*. Therefore, we chose only 20 generations and a population size of 20 for these three hybrid-GA schemes. The crossover probability of 0.5 and mutation probability of 0.033 was used. We investigated the mutation and crossover mutation rates for developing GA-MLR, GA-DT, and GA-CFS, and did not report the results as there was not much variation in the performance of the models. Default values used in WEKA were selected for the remaining GA parameters. Please note that MLR, CFS and DT did not require any parameter selection and the default values were chosen from WEKA.

In *GA-ANN*, we observed that increasing the number of generations from 20 to 100 improved the results. Since the improvement was moderate after 100 generations, we used only 100 generations. Similarly, we chose a population size of 40. The elite count was taken as 2 and crossover fraction was selected as 0.8. The rank function was used as the fitness scaling function in GA, which uses rank of each individual, rather than its score. Here, the rank of an individual is its position in the sorted scores and removes the effect of the spread of raw scores. The Gaussian function was used as the mutation function to make small random changes in the population, which provides genetic diversity and enables the GA to search a broader space. The stochastic uniform function was used for parent selection and stall fitness value was fixed as 0.001. When analyzing fitness of the population, we used a 3-layer back propagation ANN (BPNN), trained with 66% of the data while the remaining 34% data were used for testing the model. For the evaluation of fitness function, the BPNN with Bayesian regularization (BR) learning was used. The cross-validated RMSE was used as a fitness value for evaluating the population. Here, the learning rate and momentum factor values were chosen as 0.03 and 0.15, respectively.

Interpretation of the selected descriptors: We obtained nine descriptors using GA-MLR, seven descriptors using GA-CFS, 11 descriptors using GA-DT, and 13 descriptors using GA-ANN. These descriptors were found to represent the overall descriptor space (see Fig. 6), and contribution of each category of descriptors was found to be relatively same before and after the feature optimization (Figs. 3 and 6). The brief description and names of the descriptors obtained after using both methods are given in Table 1. The descriptors related to the adjacency and distance matrix, subdivided surface areas, partial charge, and atom and bond counts were found in all the descriptor sets. Here, the ‘adjacency and distance matrix’ and ‘subdivided surface area’ descriptors illustrate the importance of hydrophobicity, whereas others represent electrostatic/topological interactions for this dataset.

As shown in Fig. 6, the ‘subdivided surface area’ descriptors (especially the van der Wall surface areas of atoms and the atomic contribution to SlogP) are dominating (22% in GA-MLR, 43% in GA-CFS, 45% GA-DT and 38% in GA-ANN). The next prominent category is the ‘adjacency and distance matrix’ descriptors (11% GA-MLR, 29% in GA-CFS, 18% in GA-DT and 23% in GA-ANN), which represent atomic contributions to the hydrophobicity and molar refractivity. In addition, ClogP (a CQSAR descriptor), which models the hydrophobic interaction, is also present as an important descriptor in the GA-ANN model. These classes of descriptors together represent contribution to hydrophobic interactions. The contribution of ‘partial charge’ descriptors contributing to electronic effects is also significant (22% GA-MLR, 14% in GA-CFS, 27% in GA-DT and 15% in GA-ANN). Except for GA-CFS, all other methods have identified surface area dependent descriptors (FASA_H, DCASA and vsa_acid) as vital descriptors. The ‘atom & bond count’ descriptors especially number of nitrogen atoms is chosen by most of the methods. The GA-MLR has predicted more variety of descriptors when compared to other feature optimization techniques.

It is well known that the HIV protease receptor site is hydrophobic in nature [38]. As a result, the inhibitors with hydrophobic groups and side chains can better bind to them. The same is reflected in the obtained descriptors as discussed above. Overall, 33% of GA-MLR, 72% of GA-CFS, 64% of the GA-DT and 61% of GA-ANN selected descriptors model hydrophobic nature of the inhibitors whereas rest of them model electrostatic/topological interactions.

It is noteworthy that these descriptors are human interpretable and are able to explain the interactions between ligands (Tipranavir analogs) and their receptor protein (HIV protease). Such features can thus be used to design and synthesize a new compound with potency and specificity.

3.2. Prediction model development

The descriptors obtained using various feature optimization techniques (i.e., GA-MLR, GA-CFS, GA-DT and GA-ANN) were used to develop the prediction models. Three different approaches – linear MLR, non-linear DT and ANN – were used to develop these

Table 1
The descriptors obtained after feature optimization by GA-MLR, GA-CFS, GA-DT and GA-ANN methods.

	Descriptor type	GA-MLR	GA-CFS	GA-DT	GA-ANN
Hydrophobic	Adjacency and distance matrix descriptors	Opr_violation	BCUT_SLOGP_3 GCUT_SMR_0	PetitjeanSC Opr_violation	BCUT_PEOE_1 GCUT_PEOE_1 balabanJ
	Subdivided surface areas	SlogP_VSA8 SMR_VSA0	SlogP_VSA0 SlogP_VSA1 SMR_VSA3	SlogP_VSA5 SlogP_VSA8 SlogP_VSA9 SMR_VSA0 SMR_VSA3	ClogP[a] SlogP_VSA1 SlogP_VSA7 SlogP_VSA8 SMR_VSA3
Electrostatic	Partial charge descriptors	PEOE_VSA+2 Q_VSA_FHYD	PEOE_VSA−6	PEOE_VSA−4 Q_VSA_PNEG	PEOE_VSA+2 Q_VSA_PPOS
	Conformation dependent charge	FASA_H			DCASA
	Pharmacophoric descriptors	a_don		vsa_acid	a_don
Others	Atom counts and bond counts	a_nN	a_nN	a_nN	b_rotN
	Physical properties	PM3_dipole			

Note: Opr_violation = The oprea rules violation count and lead-like assessment; BCUT and GCUT descriptors are evaluated from the atomic contribution to the given properties like partial charges (BCUT_PEOE_1, GCUT_PEOE_1), partition coefficient (BCUT_SLOGP_3) and molar refractivity (GCUT_SMR_0); PetitjeanSC = Petitjean graph Shape coefficient; balabanJ = Connectivity topological index; SlogP_VSA = Sum of the proximate accessible van der Waals surface area, v_i, calculated for each atom over all the atoms, such that partition coefficient for atom i is in a specified range (a, b]; SMR_VSA = Sum of the proximate accessible van der Waals surface area v_i, calculation for each atom over all the atoms i, such that molar refractivity for atom i is in a specified range (a, b]; ClogP = Partition coefficient (log P(o/w)); Partial charge descriptors (PEOE_VSA+2, PEOE_VSA−6, PEOE_VSA−4, Q_VSA_FHYD, Q_VSA_PNEG, Q_VSA_PPOS) are sum of the proximate accessible van der Waals surface area v_i calculation for each atom over all the atoms i, such that partial charge of atom i is in a specified range; vsa_acid is the approximation to the sum of VDW surface areas of acidic atoms; FASA_H, DCASA are calculated from the water accessible surface areas of all the atoms; a_don is number of hydrogen bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen bond donors and acceptors such as –OH); b_rotN is number of rotatable bonds; and a_nN is number of nitrogen atoms; PM3_diple is the dipole moment calculated using the PM3 Hamiltonian.

[a] Obtained using CQSAR [40] software, others from MOE [39] software.

Table 2
Correlation coefficient (R) and RMSE values obtained using various QSAR prediction models for full training set and 10-fold cross validation (CV) set.

Method		Training set		10-fold CV	
		R	RMSE	R	RMSE
GA-MLR	MLR	0.880	0.516	0.857	0.566
	DT	0.901	0.475	0.844	0.585
	ANN	0.955	0.321	0.812	0.635
GA-CFS	MLR	0.845	0.582	0.823	0.618
	DT	0.910	0.453	0.846	0.582
	ANN	0.892	0.492	0.866	0.560
GA-DT	MLR	0.836	0.597	0.790	0.670
	DT	0.927	0.414	0.896	0.484
	ANN	0.960	0.310	0.826	0.608
GA-ANN	MLR	0.858	0.558	0.808	0.644
	DT	0.915	0.441	0.857	0.561
	ANN	0.974	0.250	0.886	0.527

models. These models are discussed, analyzed and compared below in this section, to test their performance for this dataset. Table 2 shows the correlation coefficient (R) and root mean squared error (RMSE) values of the MLR, DT and ANN prediction models on the descriptor sets obtained using all four feature optimization techniques. The values are reported for the training set as well as the 10-fold cross validation.

For the full training set, the ANN prediction models achieved higher correlation coefficient (R) values than the MLR and DT prediction models for various feature optimization techniques, except the GA-CFS. For the 10-fold cross validation, superior prediction was achieved when the same fitness function of the hybrid-GA technique was used to develop the prediction model. In the case of GA-CFS descriptors, the ANN technique outperformed MLR and DT for 10-fold cross validation.

3.2.1. MLR prediction models

Eqs. (7) and (8) below show the MLR models developed using the descriptors obtained by the GA-MLR and GA-CFS feature optimization methods, respectively. The plots of the experimental vs. predicted biological activity ($\log(1/K_i)$) for the developed models are shown in Fig. 7. These plots also show the values of correlation coefficient (R) by using the 10-fold cross validation, along with the standard deviation (SD), and cross-validated R-square (R^2_{cv}) values. It is observed from the plots that the MLR models developed using the GA-MLR descriptor set (Eq. (7)) have slightly better prediction than the GA-CFS descriptor set (Eq. (8)) as evidenced by their higher R = 0.857 vs. 0.823, lower SD = 0.500 vs. 0.525 and higher $R^2_{\text{cv}} = 0.735$ vs. 0.677 values. Similarly, the calculated RMSE value of the model obtained using GA-MLR descriptor set is lower than the model obtained using GA-CFS descriptor set (0.566 vs. 0.618) as shown in Table 2.

$$\begin{aligned}\mathrm{Log}(1/K_i)_{\text{GA-MLR}} = \;& 0.7205 * a\_nN + 0.0157 * PEOE\_VSA + 2 + 9.2487 * Q\_VSA\_FHYD - 0.1838 * opr\_violation \\ & -0.6868 * a\_don - 0.0427 * PM3\_dipole + 0.0049 * S\log P\_VSA8 + 0.0258 * SMR\_VSA0 \\ & -7.4414 * FASA\_H + 4.7546\end{aligned} \tag{7}$$

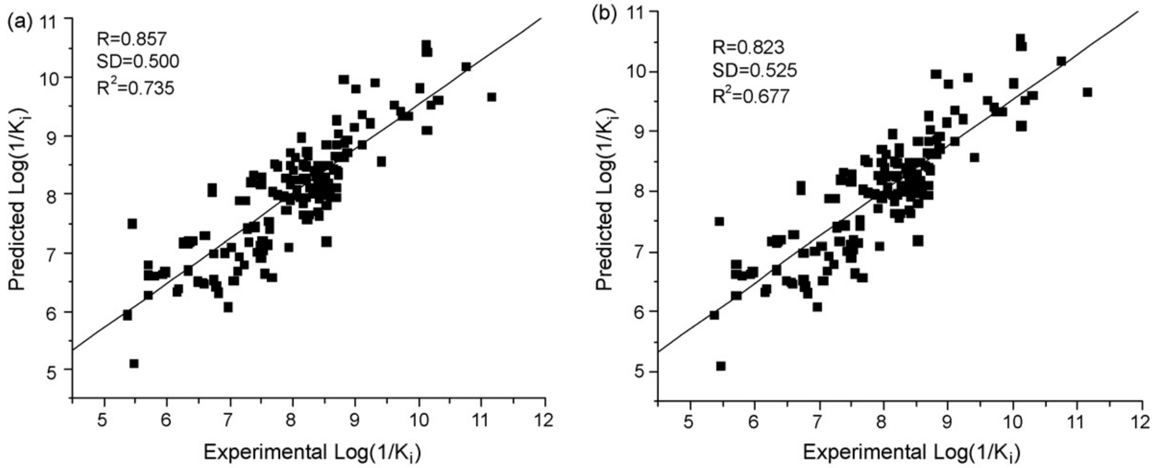


Fig. 7. Plot of experimental vs. predicted activity (log(1/K_i)) of 10-fold cross-validated MLR prediction models based on the descriptors obtained using (a) GA-MLR, and (b) GA-CFS feature optimization methods.

$$\mathrm{Log}(1/K_i)_{\text{GA-CFS}} = \begin{array}{l} 1.6454 * \mathit{BCUT\_SLOGP\_}3 - 66.9872 * \mathit{GCUT\_SMR\_}0 + 0.3333 * \mathit{a\_nN} + 0.0147* \\ \mathit{PEOE\_VSA\_}6 - 0.0171 * S\log P\_\mathit{VSA}0 + 0.0133 * S\log P\_\mathit{VSA}1 + 0.0297 * \mathit{SMR\_VSA}3 - 34.6025 \end{array} \quad (8)$$

In the developed linear QSAR models, higher positive coefficients of a given descriptor suggests that the compounds with the higher value of that descriptor exhibits higher activity, whereas the negative coefficient of a given descriptor suggests that the compounds with lower value of that descriptors would posses higher activity [32]. In both the linear QSAR model developed, the positive coefficient of the all the subdivided surface area descriptors (SlogP_VSA8, SMR_VSA0, SlogP_VSA1, SMR_VSA3) except SlogP_VSA0, partial charge descriptors (PEOE_VSA+2, Q_VSA_FHYD, PEOE_VSA_6), negative coefficient of adjacency and distance matrix descriptors (opr_violation, GCUT_SMR_0) except BCUT_SLOGP_3 shows that the compounds with higher hydrophobic nature exhibits higher HIV-1 protease activity. Interestingly GCUT_SMR_0 has a large negative coefficient in Eq. (8). FASA_H (water accessible surface area) descriptors show that these descriptors contribute negatively to the activity. In both the equations a_nN contributes significantly.

As discussed in Section 3.1, these descriptors are able to explain hydrophobic, electronic and topological interactions of the compounds in this dataset with HIV protease receptor. Similarly, Eq. (8) is also able to explain interactions between ligands and receptor.

3.2.2. Decision tree prediction models

As discussed in Section 2.3, the M5 decision trees implemented in WEKA [48] were used to construct the pieces of linear equations (linear models (LM)) at the leaf nodes of decision tree. The six DT models using the GA-DT descriptors are shown in Fig. 8.

Out of the 11 descriptors generated by the GA-DT feature selection technique, only five descriptors are involved in the decision making, which include SMR_VSA0, SlogP_VSA8, SlogP_VSA9 (hydrophobic), Q_VSA_PNEG (electronic) and a_nN (number of nitrogen atoms), at the root node atoms. All the linear models generated at the leaf nodes have a good similarity in their descriptors, but their coefficient values are different. It is noteworthy that the hydrophobic descriptors are involved in the decision making at each root node.

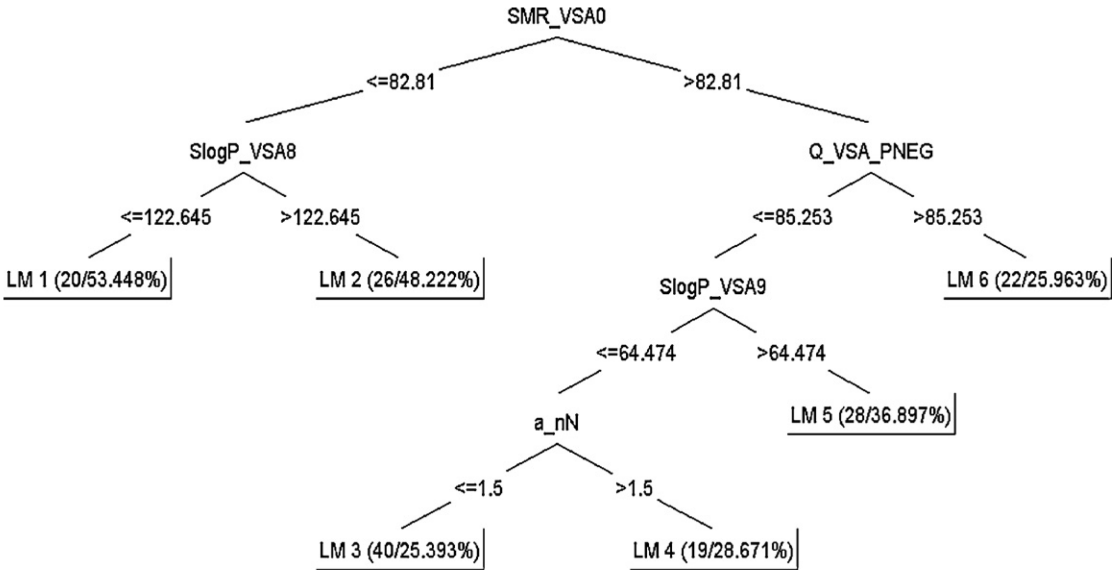


Fig. 8. Decision tree and the regression equations developed using the descriptors obtained using GA-DT feature optimization method.

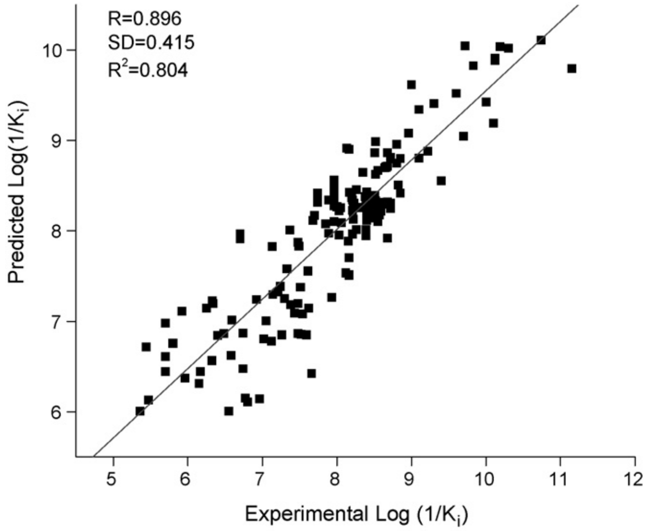


Fig. 9. Plot of experimental vs. predicted activity ($\log(1/K_i)$) values of 10-fold cross-validated DT model, based on the descriptors obtained using GA-DT feature optimization method.

The plot of the experimental vs. predicted biological activity ($\log(1/K_i)$) by the 10-fold CV method for the developed DT model is shown in Fig. 9. As shown in Fig. 10, the R = 0.896 with standard deviation = 0.415 and R^2 = 0.804. The DT model developed using hybrid GA-DT descriptor set has slightly better prediction values than the models developed using other feature selection techniques.

3.2.3. ANN prediction models

Like in the GA-ANN feature optimization technique, a 3-layer back propagation ANN with Bayesian learning (BRNN) was used to develop the prediction model, and was fine-tuned for the parameters like number of hidden nodes, learning rate and momentum factor [63]. Upon varying the learning rate and momentum factors from 0.01 to 0.5, we found that the higher learning rate values decreased the ANN performance in modeling the dataset activity, whereas variation in the momentum did not considerably influence the learning process. The learning rate and momentum factor were fixed at 0.03 and 0.3, respectively. As discussed earlier, the Bayesian learning is less likely to suffer from over-fitting as compared to other learning schemes. We used the 10-fold cross validation to study the accuracy of the predicted ANN model. To understand the model performance, the parameters like RMSE and R^2_{cv}, defined in Section 2.3 were used.

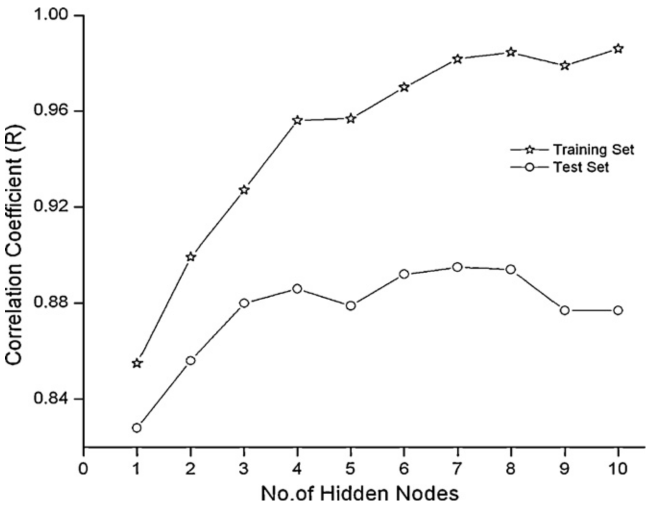


Fig. 10. Variation of correlation coefficient values of ANN prediction models with number of hidden neurons, for descriptors obtained using GA-ANN method.

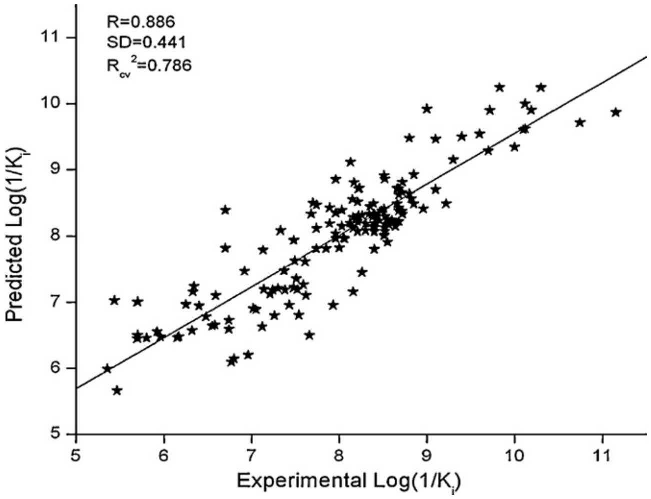


Fig. 11. Plot of experimental vs. predicted activity ($\log(1/K_i)$) values using 10-fold cross-validated ANN prediction model, based on the descriptors obtained using GA-ANN method.

The probability of over-training the ANN is high when the number of layers and the hidden layer neurons is much larger than what is actually necessary to represent the structure of an underlying learning problem. We used the network with one hidden layer and fine-tuned the number of hidden nodes. Fig. 10 depicts the variation in the correlation coefficient (R) values with the number of hidden nodes. The R value increases with the number of hidden nodes for the training set indicating the overtraining. However four hidden nodes seem to be optimal for a good prediction on the test data. This is also in agreement with the number of hidden nodes for 2nd order accuracy computed by a technique recently proposed by Trenn [63].

As shown in Fig. 11 and Table 2, the final ANN model has good prediction accuracy as is evident from the 'R' values of 0.886 and 0.866 for the test set using GA-ANN and GA-CFS descriptors, respectively. The same is reflected in the lower SD value (0.441 and 0.460) and higher R^2_{cv} value (0.786 and 0.750) from the plot and lower RMSE (Table 2). Furthermore, the model developed using the GA-ANN descriptors gives 0.02 extra fit as compared to the GA-CFS descriptor model, in terms of its 'R' value. Overall, the ANN prediction models have better prediction accuracy for our dataset than the other two (MLR and DT) models.

3.2.4. ANN weight analysis

Although ANN provides better prediction models, they suffer from lack of interpretability and can't be used for understanding the structure property (i.e., descriptor vs. activity) trends. We have used a method by Guha et al. [64] to determine the contribution of each descriptor in ANN model, by considering its final weights as briefly discussed below.

This approach considers the overall contribution of a hidden neuron to the output by using all the effective weights associated with it. The hidden neurons are ranked based on the weights, as they do not contribute to the output value equally. In general, the effective weight between the kth input neuron and the output neuron, via the lth hidden layer neuron, will be $w_{kl}w_l^H$. Here, w_{kl} is the weight between the kth input and lth hidden neuron, and w_l^H is the weight between the lth hidden neuron and the output.

Table 3
Sum of weights calculated for each hidden node (input-hidden) of ANN prediction model obtained for GA-ANN optimized descriptors.

Descriptor	Hidden node number			
	1	3	2	4
BCUT_PEOE_1	0.7787	−0.5586	−0.3139	−0.0475
GCUT_PEOE_1	0.0982	−0.5959	−0.1956	0.7651
b_rotN	0.5193	−0.3449	0.4358	−0.3112
balabanJ	0.4643	−0.3881	−1.1189	0.9853
PEOE_VSA+2	−0.4195	−0.5002	0.7517	0.1825
Q_VSA_PPOS	−0.2590	0.1427	−0.0828	0.1835
a_don	1.1124	−0.4039	−0.7429	−0.1935
SlogP_VSA1	0.5280	0.6645	−0.9109	0.2814
SlogP_VSA7	−0.1556	0.8763	0.0510	−0.5787
SlogP_VSA8	−1.2608	0.0349	1.2674	−0.0822
SMR_VSA3	0.3542	−0.0682	0.0179	−0.1583
DCASA	−0.2226	−0.4192	0.7029	0.1690
ClogP	0.9107	0.2462	0.4119	−1.6023
Summed contribution (SC)	0.6414	0.3298	0.0288	0.0001

The contribution of lth hidden neuron is computed as,(9)$C_l = \frac{1}{n_I}\sum_{l=1}^{n_I} w_{kl} w_l^H$ Here, n_I represents the total number of input neurons. The relative contribution of each hidden neuron 'l' is computed in terms of its 'squared contribution' (SC) value as,(10)$\mathrm{SC}_l = \frac{C_l^2}{\sum_{l=1}^{n_H} C_l^2}$Here the SC_l values will sum to 1.0.

The SC value of each hidden neuron and its summed weights (effective weights) for each given descriptor are reported in Table 3. The SC value of 1st hidden Node is 0.6414, which has the highest contribution. From the SC values, we observe that the total contribution of the 1st and 3rd hidden nodes is 98%. At the 1st hidden Node, the descriptors SlogP_VSA8, a_don and ClogP have higher effective weights, whereas at the 3rd hidden node which has 2nd highest SC value, the SlogP_VSA7 and SlogP_VSA1 descriptors have higher contribution. Overall, the hydrophobic descriptors are found to be dominant, which is same as seen from the MLR and DT models. The weight of the SlogP_VSA1 descriptor has positive value at hidden nodes 1, 2 and 4, which is in agreement with the DT and MLR models. Next, the ClogP descriptor also has positive weight at hidden nodes 1, 2, and 3. The same is also observed for the DT models.

In summary, we developed and tested the QSAR prediction models, using (i) the full dataset as training as well as test set, and (ii) 10-fold cross validation test (described in Section 2.3 'Methodology'). The corresponding correlation coefficient (R) and RMSE values were reported in Table 2. We observed that the prediction models developed using GA-ANN descriptors are superior in terms of R and RMSE values when the full set is used for training as well as testing. The same is also true when 10-fold cross validation is used, except for the MLR prediction models which show better performance with the GA-CFS descriptor set.

4. Conclusion

We applied four hybrid-GA descriptor selection techniques (GA-MLR, GA-DT, GA-CFS, GA-ANN) coupled with three statistical and machine-learning QSAR approaches (MLR, DT and ANN) for the prediction of biological activity of compounds on a quality-assured dataset of HIV protease inhibitors (*Tipranavir analogs*). This dataset was compiled in-house and has not been studied earlier using these methods.

All the four feature optimization approaches performed consistently, and the selected descriptors represented the whole descriptor space (Figs. 3 and 6). Although the descriptors obtained using these approaches are different, they could account well for the binding nature of the considered dataset. All the three QSAR techniques yielded the models with good prediction performance (Table 2). However, the ANN prediction models are slightly better than the MLR and DT models, whereas the MLR models have superior mechanistic interpretation. The weight analysis of the GA-ANN descriptors was carried out to interpret these models. Analysis of all three models provided useful insights about the role of various descriptors in predicting biological activity of compounds, including the involvement of hydrophobic interactions.

The three methods described in this paper differ in the type of method (linear, MLR and DT; Non-linear, ANN) used for model development. WE noticed that GA-ANN and GA-DT methods give superior results as compared to GA-MLR. Advantage of using more complex method is that it can address the more complex data/descriptor space. Use of three different methods, validates mutually and supports our results. All models together gave similar set of descriptors, which lead to the development of good prediction models. Also these models were developed after considering a wide selection of descriptors which encompass hydrophobic, electronic, geometrical, topological, and quantum mechanical properties of the molecules they would be able to describe the characteristics of new HIV-active structures better. This study showed that the use of a hybrid GA-based descriptor selection technique in combination with a QSAR technique can provide *robust prediction models with much better predictability as well as mechanistic interpretation*. These models will further enhance our understanding of the hydrophobic and other interactions between HIV protease and its inhibitors.

Acknowledgement

The authors thank Dr. Barun Bhhatarai for help in compiling the dataset.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2010.03.005.

References

[1] J. Coffin, A. Haase, J.A. Levy, L. Montagnier, S. Oroszlan, N. Teich, H. Temin, K. Toyoshima, H. Varmus, P. Vogt, Science 232 (1986) 697.
[2] C. Fortin, V. Joly, P. Yeni, Expert Opin. Emerg. Drugs 11 (2006) 217.
[3] A. Hughes, T. Barber, M. Nelson, J. Infect. 57 (2008) 1.
[4] A.K. Ghosh, G. Schiltz, R.S. Perali, S. Leshchenko, S. Kay, D.E. Walters, Y. Koh, K. Maeda, H. Mitsuya, Bioorg. Med. Chem. Lett. 16 (2006) 1869.
[5] V.A. Johnson, F. Brun-Vézinet, B. Clotet, H.F. Günthard, D.R. Kuritzkes, D. Pillay, J.M. Schapiro, D.D. Richman, Top. HIV Med. 16 (2008) 138.
[6] G.A. Patani, E.J. LaVoie, Chem. Rev. 96 (1996) 3147.
[7] C. Hansch, A. Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, John Wiley & Sons, New York, 1979.
[8] C. Hansh, T. Fujita, J. Am. Chem. Soc. 86 (1964) 1616.
[9] M.T. Khan, I. Sylte, Curr. Drug Discov. Technol. 4 (2007) 141.
[10] E. Estrada, Mini Rev. Med. Chem. 8 (2008) 213.
[11] N.B. Chapman, J. Shorter, Correlation Analysis in Chemistry: Recent Advances, Plenum Press, New York, 1978.
[12] C.W. Yap, H. Li, Z.L. Ji, Y.Z. Chen, Mini Rev. Med. Chem. 7 (2007) 1097.
[13] M. Shen, C. Béguin, A. Golbraikh, J.P. Stables, H. Kohn, A. Tropsha, J. Med. Chem. 47 (2004) 2356.
[14] E. Zvinavashe, A.J. Murk, I.M. Rietjens, Chem. Res. Toxicol. 21 (2008) 2229.
[15] R. Garg, B. Bhhatarai, QSAR and molecular modeling studies of HIV protease inhibitors, in: S.P. Gupta (Ed.), QSAR and Molecular Modeling Studies in Heterocyclic Drugs I, vol. 3, Springer-Verlag, Heidelberg, Germany, 2006, p. 181.
[16] V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, J. Chem. Inf. Comp. Sci. 40 (2000) 659.
[17] D. Wang, B. Larder, J. Infect. Dis. 188 (2003) 653.
[18] S. Drăghici, R.B. Potter, Bioinformatics 19 (2003) 98.
[19] D. Weekes, G.B. Fogel, BioSystems 72 (2003) 149.
[20] X.J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci. 44 (2004) 1257.
[21] P.E. Blower, K.P. Cross, Curr. Top. Med. Chem. 6 (2006) 31.
[22] S.P. Niculescu, J. Mol. Str. THEOCHEM 622 (2003) 71.
[23] R. Leardi, Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier Limited, Genova, Italy, 2003.

[24] M.D. Vose, The Simple Genetic Algorithm: Foundations and Theory, MIT Press, Cambridge, MA, 1999.
[25] J. Ghasemi, S. Ahmadi, Ann. Chim. 97 (2007) 69.
[26] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proc. 17th Int'l Conf. Machine Learning, 2000, p. 359.
[27] M.A. Hall, G. Holmes, IEEE Trans. Knowl. Data Eng. 15 (2003) 1437.
[28] T.-S. Chou, K.K. Yen, J. Luo, N. Pissinou, K. Makki, Proc. IEEE MILCOM 29 (2007) 1.
[29] G. Gini, E. Benfenati, D. Boley, Proc. International Conf. Knowledge-Based Intelligent Engineering Systems and Allied Technologies, vol. 1, 2000, p. 166.
[30] R. Guha, J.R. Serra, P.C. Jurs, J. Mol. Graph. Model. 23 (2004) 1.
[31] M. Vracko, Curr. Comput.-Aided Drug Des. 1 (2005) 73.
[32] (a) R.P. Verma, C. Hansch, Chem. Rev. 109 (2009) 213;
(b) R.P. Verma, C. Hansch, Mol. Pharm. 5 (2008) 745.
[33] M. Casalegno, G. Sello, E. Benfenati, J. Chem. Inf. Model. 48 (2008) 1592.
[34] A. Kurup, S.B. Mekapati, R. Garg, C. Hansch, Curr. Med. Chem. 10 (2003) 1819.
[35] R. Garg, S.P. Gupta, H. Gao, M.S. Babu, A.K. Debnath, C. Hansch, Chem. Rev. 99 (1999) 3525.
[36] M. Boiani, H. Cerecetto, M. Gonzalez, J. Gasteiger, J. Chem. Inf. Model. 48 (2008) 213.
[37] M. Daszykowski, B. Walczak, Q.S. Xu, F. Daeyaert, M.R. de Jonge, J. Heeres, L.M. Koymans, P.J. Lewi, H.M. Vinkers, P.A. Janssen, D.L. Massart, J. Chem. Inf. Model. 44 (2004) 716.
[38] Z.R. Yang, R. Thomson, IEEE Trans. Neural Net. 16 (2005) 263.
[39] L. Douali, D. Villemin, D. Cherqaoui, J. Chem. Inf. Comput. Sci. 43 (2003) 1200.
[40] B. Bhhatarai, R. Garg, Bioorg. Med. Chem. 13 (2005) 4078.
[41] D. Hecht, B. Fogel, IEEE/ACM Trans. Comput. Bio. Bioinfo. 4 (2007) 476.
[42] R. Garg, D. Patel, Bioorg. Med. Chem. Lett. 13 (2005) 3767.
[43] MOE software, Chemical Computing Group, Montreal, Canada.
[44] CQSAR program, Biobyte Corp., Claremont, CA, USA.
[45] MATLAB, The MathWorks, Inc., Natick, MA.
[46] D.A. Konovalov, N. Sim, E. Deconinck, Y.V. Heyden, D. Coomans, J. Chem. Inf. Model. 48 (2008) 370.
[47] M. Shen, Y. Xiao, A. Golbraikh, V.K. Gombar, A. Tropsha, J. Med. Chem. 46 (2003) 3013.
[48] E. Frank, M. Hall, L. Trigg, L. Holmes, I.H. Witten, Bioinformatics 20 (2004) 2479.
[49] J. Minges, Mach. Learn. 4 (1989) 227.
[50] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, The Wadsworth Statistics/Probability Series, Belmont, CA, Wadsworth, 1984.
[51] T.K. Ho, Random decision forest, in: Proc. 3rd Int'l Conf. Document Analysis and Recognition, Montreal, Canada, August 14–18, 1995, pp. 278–282.
[52] Y. Wang, I. Witten, Inducing model trees for continuous classes, in: 9th European Conf. Machine Learning, Prague, 1997, p. 128.
[53] J.R. Quinlan, Learning with continuous classes, in: 5th Australian Joint Conf. Artificial Intelligence, Singapore, 1992, p. 343.
[54] F.D. Foresee, M.T. Hagan, Gauss–Newton approximation to Bayesian regularization, in: Proc. International Joint Conf. on Neural Nets, 1997, p. 1930.
[55] M.F. Moller, Neural Nets 6 (1993) 525.
[56] N. Qian, Neural Nets 12 (1999) 145.
[57] D.J.C. Mackay, Neural Comput. 4 (1992) 415.
[58] K. Levenberg, Quart. Appl. Math. 2 (1944) 164.
[59] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Comput. Sci. 43 (2003) 579.
[60] D.A. Konovalov, L.E. Llewellyn, Y.V. Heyden, D. Coomans, J. Chem. Inf. Model. 48 (2008) 2081.
[61] I.V. Tetko, D.J. Livingstone, A.I. Luik, J. Chem. Inf. Comput. Sci. 35 (1995) 826.
[62] X.-H. Yu, G.-A. Chen, Neural Nets 10 (1997) 517.
[63] S. Trenn, IEEE Trans. Neural Nets 19 (2008) 836.
[64] R. Guha, D.T. Stanton, P.C. Jurs, J. Chem. Inf. Model. 45 (2005) 1109.