European Journal of Medicinal Chemistry 44 (2009) 2913–2922

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques

Kunal Roy*, Partha Pratim Roy

Drug Theoretics and Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

ARTICLE INFO

Article history:
Received 2 June 2008
Received in revised form
26 November 2008
Accepted 8 December 2008
Available online 16 December 2008

Keywords:
QSAR
Cytochrome 3A4
FA-MLR
PLS
GFA
ANN

ABSTRACT

Twenty-eight structurally diverse cytochrome 3A4 (CYP3A4) inhibitors have been subjected to quantitative structure–activity relationship (QSAR) studies. The analyses were performed with electronic, spatial, topological, and thermodynamic descriptors calculated using Cerius 2 version 10 software. The statistical tools used were linear [multiple linear regression with factor analysis as preprocessing step (FA-MLR), stepwise MLR, partial least squares (PLS), genetic function algorithm (GFA), genetic PLS (G/PLS)] and non-linear methods [artificial neural network (ANN)]. All the five linear modeling methods indicate the importance of *n*-octanol/water partition coefficient (log *P*) along with different topological and electronic parameters. The best model obtained from the training set (stepwise regression) based on highest external predictive R^2 value and lowest RMSEP value also showed good internal predictive power. Other models like FA-MLR, PLS, GFA and G/PLS are also of statistically significant internal and external validation characteristics. The best model [according to r^2_m for the test set, as defined by P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313] obtained from ANN showed a good r^2 value (determination coefficient between observed and predicted values) for the test set compounds, which was superior to those of other statistical models except the stepwise regression derived model. However, based upon the r^2_m value (test set), which penalizes a model for large differences between observed and predicted values, the stepwise MLR model was found to be inferior to other methods except PLS. Considering r^2_m value for the whole set, the G/PLS derived model appears to be the best predictive model for this data set. For choosing the best predictive model from among comparable models, r^2_m for the whole set calculated based on leave-one-out predicted values of the training set and model-derived predicted values for the test set compounds is suggested to be a good criterion.

1. Introduction

Nowadays different *in silico* techniques are applied for screening new chemical entities in terms of metabolism and toxicology [1]. For accelerating drug discovery, the pharmaceutical screening process requires the use of these techniques in an earlier stage of drug development to identify the potential interaction of the new chemical entity with particular drug metabolizing enzymes, namely cytochrome P-450s (CYPs) [2]. Cytochrome P-450s (CYPs) comprise of a superfamily of heme-thioloate monooxygenase enzymes that catalyze the oxidation of a wide variety of xenobiotic chemicals, including drugs and carcinogens [3–5]. Cytochrome P450 3A4 (abbreviated as CYP3A4), previously known as nifedipine oxidase, is a member of the cytochrome P450 mixed-function oxidase system [6]. CYP3A4 is one of the most important enzymes involved in the metabolism of xenobiotics in our body (approximately 30% of all known xenobiotic oxidations). CYP3A4 has the highest abundance in the human liver (~40%) and it metabolizes more than 50% of the clinically used drugs [7,8]. CYP3A4 catalyzes the reactions like alkyl carbon hydroxylation, O- and *N*-dealkylation, epoxidation and less frequently, aromatic ring hydroxylation [9]. It is widely recognized that CYP is the major class of oxidative enzymes involved in drug metabolism [10]. CYP3A4 is known to metabolize a large variety of compounds varying in molecular weight from lidocaine (MW = 234) to cyclosporine (MW = 1203) [5,11]. The active pocket in CYP3A4 is approximately 520 Å [12]. As CYP3A4 is considered the major oxidative enzyme

* Corresponding author. Tel.: +91 98315 94140; fax: +91 33 2837 1078.
E-mail address: kunalroy_in@yahoo.com (K. Roy).
URL: http://www.geocities.com/kunalroy_in

0223-5234/$ – see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.12.004

involved in drug metabolism, numerous drug–drug interactions have been reported, where the inhibition of this enzyme by a drug results in the decreased clearance of other drugs [2]. CYP3A4 substrates bind with Asn74 residue of CYP3A4 with hydrogen bonding. Structural requirements of CYP3A4 substrates have been suggested to include a hydrogen bond acceptor atom 5.5–7.8 Å from the site of metabolism and 3 Å from the oxygen molecule [13].

Among different techniques available for screening of new chemical entity, quantitative structure–activity relationship (QSAR) is the one practiced very often. QSARs represent predictive models derived from application of statistical tools correlating biological activity (including therapeutic and toxic) of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or property. The success of any QSAR model depends on the accuracy of input data, selection of appropriate descriptors and statistical tools, and most importantly validation of the developed model [14–18]. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged. For validation of QSAR models usually four strategies are adopted [19]: (a) internal validation or cross-validation; (b) validation by dividing the data set into training and test compounds; (c) true external validation by application of model on external data and (d) data randomization or Y-scrambling.

Optimum physicochemical properties for inhibitors of the CYP3A4 enzyme may be explored through QSAR studies. Good correlation has been reported for the binding affinity of azole compounds towards CYP3A4 enzyme with their hydrophobic parameter (log *P*) [20]. A support vector machine (SVM) approach was applied to classify human cytochrome P450 3A4 inhibitors by Kriegl et al. [21] SVMs were also used to predict the potency of structurally diverse drug-like molecules to inhibit CYP3A4 enzyme [22]. Lewis et al. reported QSAR for free energy changes of CYP3A4 binding of diverse group of compounds [23]. In a previous study, the present group of authors highlighted the importance of molecular properties like electronic, shape and spatial properties of azoles towards the binding affinity with CYP3A and CYP2B enzymes [24]. In the present paper, we have developed QSAR models for structurally diverse chemicals by using the CYP3A4 enzyme inhibition data reported in the literature [23]. We have validated the models by dividing the data set into training and test sets by K-means clustering technique. Different statistical techniques (both linear and non-linear) were used to develop the models to highlight the structural requirements for an ideal CYP3A4 inhibitor. The objectives of the present paper have been twofold: (1) to explore the structure–activity relationships of CYP3A4 binding affinity of diverse compounds and (2) to select the best predictive model from among comparable chemometric models for the CYP3A4 binding affinity.

2. Material and methods

2.1. The data set and descriptors

The CYP3A4 inhibitory activity of twenty-eight diverse compounds (Table 1) reported in the literature [23] has been used as the model data set for the present study. The IC_{50} values are converted to logarithmic scale (M). The analyses were performed

Table 1
Observed and calculated values of CYP3A4 inhibitory activity of structurally diverse compounds.

Sl. No		obs[a]	cal[b]	cal[c]	cal[d]	cal[e]	cal[f]	cal[g]
Training set								
1	Erythromycin	3.879	4.575	4.308	4.657	3.908	4.073	4.251
2	Chloroquine	3.456	4.051	3.939	4.450	3.657	3.240	4.305
3	Diltiazem	3.663	4.799	4.498	4.490	4.287	4.147	4.285
5	Bromocriptine	5.523	5.215	4.952	5.276	5.387	4.967	5.431
6	Dihydroergotamine	5.523	5.366	5.096	5.198	5.626	5.495	5.371
8	Sulfamethiazole	3.078	3.534	3.509	3.543	2.913	3.129	3.643
9	Timoprazole	3.510	4.046	3.923	3.847	3.681	3.823	3.576
11	Tazanolast	3.538	4.209	4.076	3.992	3.821	3.904	3.980
12	Cimetidine	3.000	3.583	3.572	3.368	3.214	3.014	2.742
13	Nifedipine	4.328	4.213	3.919	3.950	4.410	4.261	4.493
14	Omeprazole	4.108	4.645	4.410	4.374	4.002	3.763	4.077
16	2-Methylimidazole	2.902	2.714	2.741	2.770	3.017	2.923	2.870
17	3-Hydroxypyridine	3.115	2.848	2.841	2.887	2.905	3.264	3.109
20	Triadimefon	5.032	4.618	4.886	4.703	5.410	5.569	5.092
21	Propiconazole	5.983	5.252	5.537	5.251	6.360	6.323	5.626
22	Ketoconazole	6.699	5.936	6.264	6.148	6.488	6.650	6.484
23	Metyrapone	5.307	3.832	3.697	3.641	4.607	4.885	4.622
24	Glipizide	5.128	4.841	4.637	4.570	4.323	4.728	4.854
25	Miconazole	6.070	6.463	6.579	6.525	6.338	6.448	6.106
27	Fluconazole	4.602	3.682	4.747	4.151	4.650	4.577	4.523
28	Econazole	6.366	6.386	6.535	6.170	5.807	5.626	6.124
Test set								
4	Verapamil	4.119	5.300	4.879	5.125	5.719	5.556	5.150
7	Troleandomycin	5.097	5.764	5.293	5.396	6.400	4.981	4.552
10	Piroxicam	3.000	4.343	4.142	3.947	3.278	4.098	4.184
15	1-Methylimidazole	2.569	2.584	3.402	3.125	2.883	2.728	3.508
18	2-Anilinopyridine	3.366	4.654	4.362	3.967	4.413	4.841	4.291
19	Thiopermide	5.215	5.063	4.799	4.309	3.953	4.477	4.566
26	Clotrimazole	7.301	6.047	6.274	5.640	6.508	6.650	6.686

[a] Observed (Ref. [23]).
[b] Calculated from model (1).
[c] Calculated from model (2).
[d] Calculated from model (3).
[e] Calculated from model (4).
[f] Calculated from model (5).
[g] Calculated from model (N1).

using electronic (Apol, Dipole, HOMO, LUMO and Sr), spatial (radius of gyration, area, PMI-mag, density, V_m) and thermodynamic ($A \log P$, $A \log P98$, Molref) descriptors as well as different topological parameters (E-state index, kappa shape index, molecular connectivity index, flexibility index, Wiener, Zagreb, subgraph count indices). Definitions of different descriptors have been given in Table 2. All the descriptors were calculated using descriptor + module of the Cerius 2 version 4.10 software [25]. Along with these descriptors, octanol–water partition coefficient ($\log P$) of the compounds reported in Ref. [23] has also been used as an additional descriptor.

2.2. Model development

Most of the QSAR modeling methods implement the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedure. The outcome from the cross-validation procedure is cross-validated R^2 (LOO-Q^2 or LSO-Q^2) which is used as a criterion of both robustness and predictive ability of the model. Cross-validated determination coefficient R^2 (LOO-Q^2) is calculated according to the formula

$$Q^2 = 1 - \frac{\sum\left(Y_{obs} - Y_{pred}\right)^2}{\sum(Y_{obs} - \overline{Y})^2}$$

In above equation, $\overline{Y}$ means average activity value of the entire data set while Y_{obs} and Y_{pred} represent observed and predicted activity values. Often, a high Q^2 value ($Q^2 > 0.5$) is considered as a proof of high predictive ability of the model [26]. But it has been found that if a test set with known values of biological activities is available for prediction, there may not exist any correlation between LOO- (or LSO-) cross-validated R^2 (Q^2) and determination coefficient r^2 between the predicted and observed activities for the test set [27,28]. However, cross-model validation has been suggested to give the better measure of model predictive ability than simple cross-validation [29]. However, models which have been additionally externally validated can be considered predictive for new chemicals [27,28].

In many cases, truly external data points being unavailable for prediction purpose, original data set compounds are divided into training and test sets [30]. Equations are developed based on training set compounds and predictive capacity of the models is judged based on the predictive R^2 (R^2_{pred}) values calculated according to the following equation:

$$R^2_{Pred} = 1 - \frac{\sum\left(Y_{pred(Test)} - Y_{(Test)}\right)^2}{\sum\left(Y_{(Test)} - \overline{Y}_{training}\right)^2}$$

In the above equation, $Y_{pred(Test)}$ and $Y_{(Test)}$ indicate predicted and observed activity values, respectively, of the test set compounds and $\overline{Y}_{training}$ indicates mean activity value of the training set.

2.2.1. Statistical methods

It is our priority to construct QSAR models which is statistically robust both internally as well as externally. We have classified the data set into clusters using K-means clustering technique [31] based on the standardized (values between 0 and 1) predictor variables. This approach (clustering) ensures that the similarity principle can be employed for the activity prediction of the test set. To begin the model development process, the data set ($n = 28$) was divided into training (75% of the total number of compounds) and test (25% of the total number of compounds) sets with

Table 2
Definition of different variables.

Type of descriptors	Descriptor name	Definition	Comment, if any
Electronic	Apol	Sum of atomic polarizabilities	
	Dipole	Dipole moment	
	HOMO	Highest occupied molecular orbital energy	It measures nucleophilicity of a molecule
	LUMO	Lowest unoccupied molecular orbital energy	It measures electrophilicity of a molecule
	Sr	Superdelocalizability	It may be used to predict relative reactivity in a series of molecules
Spatial	RadOfGyration	$\sqrt{\frac{(x_i^2+y_i^2+z_i^2)}{N}}$	N is the number of atoms and x, y, z are the atomic coordinates relative to the center of mass
	Density	The ratio of molecular weight to molecular volume	It reflects the types of atoms and how tightly they are packed in a molecule
	PMI-mag	It calculates the principal moments of inertia about the principal axes of a molecule	
	V_m	Molecular volume inside the contact surface	
	Area	van der Waals area of a molecule	
Thermodynamic	$A \log P$	log of the partition coefficient	A.K. Ghose, G.M. Crippen, J. Comput. Chem. 1986, 7, 565–577
	MolRef	Molar refractivity	A.K. Ghose, G.M. Crippen, J. Comput. Chem. 1986, 7, 565–577
	$A \log P98$	log of partition coefficient	A. Ghose, V.N. Viswanadhan, J.J. Wendoloski, J. Phys. Chem., 1998, 102, 3762–3772
Topological	Balaban's J index	Average distance-based connectivity index	
	Molecular connectivity index	These indices are based on graph-theoretical invariant introduced by Randic	
	Kappa shape index	These indices capture different aspects of molecular shape	
	Zagreb	Sum of the squares of vertex valencies	
	Subgraph count index	This is the number of subgraphs of a given type and order	
	Flexibility index	This is a descriptor based on structural properties that restrict a molecule from being "infinitely flexible"	
	Wiener	Total number of bonds between all pairs of atoms in the hydrogen suppressed graph	
	E-state parameters	Electrotopological state parameters of atoms having different electronic and topological environment	

proportionate representation of compounds from all clusters in the training and test sets. QSAR models were developed using the training set compounds (optimized by Q^2), and then the developed models were validated (externally) using the test set compounds. For the development of equations, five methods were used: (1) stepwise multiple linear regression (stepwise MLR) [32], (2) multiple linear regression with factor analysis as the data-preprocessing step (FA-MLR) [33,34] and (3) partial least squares (PLS) [35], (4) MLR with genetic function approximation (GFA) [36,37], (5) genetic partial least squares (G/PLS) [38,39], (6) artificial neural network (ANN) [40]. The stepwise regression and FA were performed using the statistical software SPSS [41]. K-means clustering, standardization of the variables and PLS were performed using statistical software MINITAB [42]. ANN was performed using the software STATISTICA [43].

In stepwise regression [32], a multiple-term linear equation was built step-by-step. The basic procedures involve (1) identifying an initial model, (2) iteratively "stepping," that is, repeatedly altering the model at the previous step by adding or removing a predictor variable in accordance with the "stepping criteria," ($F = 4$ for inclusion; $F = 3.5$ for exclusion) and (3) terminating the search when stepping is no longer possible given the stepping criteria, or when a specified maximum number of steps have been reached. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the equation. That variable is then included in the model, and the process starts again. The F value used for inclusion or exclusion of a variable in the stepwise regression process is a test for partial regression coefficient and it is obtained by dividing the difference between reductions of sum of squares with and without the variable being included or excluded with error mean square of the equation [44]. The F value for inclusion or exclusion of a variable in an MLR equation during stepwise process is the square of the t value of the regression coefficient of the variable being included or excluded. A limitation of the stepwise regression search approach is that it presumes that there is a single "best" subset of X variables and seeks to identify it. There is often no unique "best" subset, and all possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to study whether some other subsets of X variables might be better.

In case of FA-MLR, classical approach of multiple linear regression (MLR) technique was used as the final statistical tool for developing classical QSAR relations and factor analysis (FA) [33,34] was used as the data-preprocessing step to identify the important descriptors contributing to the response variable and to avoid collinearities among them. FA-MLR is different from principal component regression analysis (PCRA) where factor scores are used as variables instead of the original descriptors. In a typical factor analysis procedure, the data matrix is first standardized, and correlation matrix and subsequently the reduced correlation matrix is constructed. An eigen value problem is then solved and the factor pattern can be obtained from the corresponding eigen vectors. The principal objectives of factor analysis are to display multidimensional data in a space of lower dimensionality with minimum loss of information (explaining >95% of the variance of the data matrix) and to extract the basic features behind the data with ultimate goal of interpretation and/or prediction. Factor analysis was performed on the data set(s) containing biological activity and *all* descriptor variables, which were to be considered. The factors were extracted by principal component method and then rotated by VARIMAX rotation (a kind of rotation which is used in principal component analysis so that the axes are rotated to a position in which the sum of the variances of the loadings is the maximum possible) to obtain Thurston's simple structure. The simple structure is characterized by the property that as many variables as possible fall on the coordinate axes when presented in common factor space, so that the largest possible number of factor loadings becomes zero. This is done to obtain a numerically comprehensive picture of the relatedness of the variables. Only variables with non-zero loadings in such factors where biological activity (or response property) also has non-zero loading were considered important in explaining variance of the activity (or response). Furthermore, variables with non-zero loadings in different factors were combined in a multivariate equation. Clearly, variables are selected based on the importance of different factors (features) to the response variable (biological activity), and thus the method of selection is quite different from stepwise selection of variables as is done in case of stepwise regression.

PLS is a generalization of regression, which can handle data with strongly correlated and/or noisy or numerous X variables [35]. It gives a reduced solution, which is statistically more robust than MLR. The linear PLS model finds "new variables" (latent variables or X scores) which are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross-validation is a reliable and commonly used method for testing this significance [36]. However, recently it has been shown that from the viewpoint of external predictability, choice of variables for PLS based on internal validation may not be optimum [45]. Application of PLS allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables. PLS is normally used in combination with Cross-validation to obtain the optimum number of components. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data [46]. Based on the standardized regression coefficients, the variables with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components.

Genetic function approximation (GFA) technique [36,37] was used to generate a population of equations rather than one single equation for correlation between biological activity and physicochemical properties. GFA involves the combination of multivariate adaptive regression splines (MARS) algorithm with genetic algorithm to evolve population of equations that best fit the training set data. It provides an error measure, called the lack-of-fit (LOF) score that automatically penalizes models with too many features. It also inspires the use of splines as a powerful tool for non-linear modeling. A distinctive feature of GFA is that it produces a population of models (e.g., 100), instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors.

The genetic partial least squares (G/PLS) algorithm [38,39] may be used as an alternative to a GFA calculation. G/PLS is derived from two QSAR calculation methods: GFA and partial least squares (PLS). The G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model and PLS regression as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of G/PLS thus allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables.

GFA can build models not only with linear polynomials but also with higher order polynomials, splines, and Gaussians. By using spline-based terms, GFA can perform a form of automatic outlier removal and classification. The splines used are truncated power splines and are denoted with angular brackets. For example, $\langle f(x) - a \rangle$ is equal to zero if the value of $(f(x) - a)$ is negative, else it is equal to $(f(x) - a)$. The constant 'a' is called the knot of the spline. A spline partitions the data samples into two classes, depending on

the value of some feature. The value of the spline is zero for one of the classes and non-zero for the other classes. Splines are interpreted as performing either range identification or outlier removal. If there are many members in the non-zero partition, then the spline is identifying a range of effect. If there are only a few members of the non-zero set, this indicates that the spline is identifying outliers [37].

Artificial neural network (ANN) [40] is an information-processing pattern that is inspired by the way biological nervous systems, such as the brain, process information. Majority of the networks contain at least three layers – input, hidden and output. The layers of input neurons receive the data either from input files or directly from electronic sensors in real-time applications. The output layer sends information directly to the outside world, to a secondary computer process or to other devices such as a mechanical control system. Between input and output layers there may be many hidden layers. These internal layers contain many of the neurons in various interconnected structures. Based on the function, there are different types of neural networks like feed-forward backpropagation, counter propagation, probabilistic neural network, self-organizing map, etc. But here in the present study, for the development of our non-linear models, feed-forward back-propagation method was used. Multilayer perceptron (MLP) method under "Custom Network Designer" has been selected to design the network. In the first phase backpropagation method was selected for formation of the network using training set. The error term, i.e., difference between output of the network and the desired output is back propagated to the transfer function (sigmoid function) for adjustment of weight. The output [47] can be represented by the following equation.

$$O_j = f(i_j) = \frac{1}{1+\exp(-\beta i_j)}$$

where O_j is the output of node j and β is a gain, being able to adjust the form of the function. Usually β is taken as 1. Using the error signal to adjust the connected weights, the following adjusted weights are obtained for the output layer.

$$W_{i_j}(\text{new}) = W_{i_j}(\text{old}) + \eta\delta_i O_j + \alpha\left[\Delta W_{i_j}(\text{old})\right]$$

In backpropagation method, the learning of the network has followed the Delta Rule, which starts with the calculated difference between the actual outputs and the desired outputs. Using this error, connection weights are increased in proportion to the error times a scaling factor for global accuracy. The complex part of this learning mechanism is to determine which input contributed the most to an incorrect output and how does that element get changed to correct the error. During the learning process, a forward sweep is made through the network, and the output of each element is computed layer by layer. The difference between the output of the final layer and the desired output is back propagated to the previous layer until the input layer is reached. In second phase conjugate gradient descent was used. This method is a good secondary and advanced method of training multilayer perception. It is generally used for the network of large numbers of weights and/or multiple output units. It is a batch update algorithm whereas backpropagation adjusts the weights of the network. Learning rate and momentum of each epoch are adjusted and weight decay is regularized. Cross-validated resampling was used as sampling procedure during formation of network. When a particular number of resampling is selected, the available objects are divided into 3 subsets (training, selection and test sets). Training subset is used to optimize the network. The second subset, i.e., selection set is used to prevent the training from becoming over learned. Finally, a test subset is used to estimate the performance of that network.

Although the use of a test subset set allows us to generate unbiased performance estimates, these estimates may exhibit high variance. Ideally, one would like to repeat the training procedure a number of different times, each time using new training, selection and test cases drawn from the population – then, one could average the performance prediction over the different test subsets, to get a more reliable indicator of generalization performance. In reality, one seldom has enough data to perform a number of training runs with entirely separate training, selection and test subsets.

2.2.2. Statistical qualities

The statistical qualities of the equations were judged by the parameters such as *explained variance* (R_a^2), *determination coefficient* (R^2) and *variance ratio* (F) at specified *degrees of freedom* (*df*) [44]. R_a^2 is defined as

$$R_a^2 = \frac{\left(n-1\right)R^2 - p - 1}{n - p - 1}$$

where n is the number of compounds, p is the number of predictor variables.

F is derived from this equation

$$F = \frac{\frac{\sum(Y_{cal}-\overline{Y})^2}{p}}{\frac{\sum(Y_{obs}-Y_{cal})^2}{n-p-1}}$$

In the above equation, Y_{obs} and Y_{calc} are observed and calculated biological activity values, respectively, while $\overline{Y}$ is the mean of the observed activity values. For G/PLS equations, least-square errors (LSEs) were taken as an objective function to select an equation, while lack-of-fit (LOF) was noted for the GFA derived equations. The generated QSAR equations were validated by leave-one-out *cross-validation* R^2 (Q^2) and *predicted residual sum of squares* (PRESS) [19,48,49] and then were used for the prediction of enzyme inhibition potency values of the test set compounds and the prediction qualities of the models were judged by statistical parameters like predictive R^2 (R_{pred}^2), determination coefficient between observed and predicted values with (r^2) and without (r_0^2) intercept. It was previously shown that use of R_{pred}^2 and r^2 might not be sufficient to indicate the external validation characteristics [45]. Thus, an additional parameter r_m^2 (defined as $r^2*(1-\sqrt{r^2-r_0^2})$, which penalizes a model for large differences between observed and predicted values, was also calculated. Root mean square error of prediction (RMSEP) values of different models was also noted. It may be noted that correlation is dependent on range while RMSEP is not. Two additional parameters $r_{m(LOO)}^2$ and $r_{m(overall)}^2$ were calculated for models internal as well as the overall quality. Finally developed regression based models were subjected to a randomization test for validation purpose.

3. Results and discussion

Membership of compounds in different clusters generated using K-means clustering is shown in Table 3. The PCA score plot (Fig. 1) of first three components of the descriptor matrix shows distribution of training and test set compounds in 3D space and their cluster membership (1, 2 or 3). It may be noted that distribution of the data set into training and test sets has been done by K-means clustering and not using the PCA score plot. However, the plot shows that each test set member is close to at least one training set member in the multidimensional space. The test set size was set to 25% to the total data set size [50] and the test set members are shown in Table 1. The values of important descriptors are shown in Table 4.

Table 3
K-means clustering using topological, structural, spatial, thermodynamic descriptor space.

Cluster No.	No. of compounds in cluster	Compounds (Sl nos.) in each clusters
1	11	2 3 4 10 13 14 21 24 25 26 28
2	12	8 9 11 12 15 16 17 18 19 20 23 27
3	5	1 5 6 7 22

3.1. FA-MLR

The factor analysis of the data matrix shows that 11 factors could explain 95.7% of the variance of the total data matrix (all descriptors along with the response variable). The response variable is highly loaded with factor 2 (which is in turn highly loaded in log *P* and *A* log *P*98) while it is moderately loaded with factor 1 (highly loaded in Balaban J_x index, different kappa shape indices, connectivity indices, area, molar volume, hydrogen bond acceptor, dipole moment and molar refractivity). Based on the results of factor analysis (relative importance of the factors for the response variable), the following best equation was obtained.

$$\begin{aligned}&\mathrm{pIC}_{50} = 5.001\,(\pm 0.776) - 0.829\,(\pm 0.329) J_x + 0.511\,(\pm 0.109)\log P\\ &n_{\mathrm{Training}} = 21, R^2 = 0.711, R^2_a = 0.679, F = 22.19\,(df\,2,18),\\ &Q^2 = 0.644, \mathrm{PRESS} = 10.477,\\ &n_{\mathrm{Test}} = 7, R^2_{\mathrm{pred}} = 0.573, r^2 = 0.667, r^2_0 = 0.637, r^2_m = 0.551\end{aligned} \quad (1)$$

The standard errors of regression coefficients are given within parenthesis. Two variables J_x and log *P* in Eq. (1) could explain 67.9% of the variance (adjusted coefficient of variation) of the activity. The leave-one-out predicted variance was found to be 64.4%. The difference between R^2 and Q^2 is less than 0.3 signifying robustness of the model [51]. While Eq. (1) was applied for prediction of test set compounds, the predictive R^2 value for the test set was found to be 0.573. However, simple determination coefficient between the observed and predicted values of the test set compounds was found to be 0.667. Setting intercept to zero, the determination coefficient was found to be 0.637. As r^2 and r^2_0 values are not much different, an acceptable value of r^2_m (0.551) was obtained. The intercorrelation (r) among predictor variables is shown in Table 5, which suggests the absence of high intercorrelation.

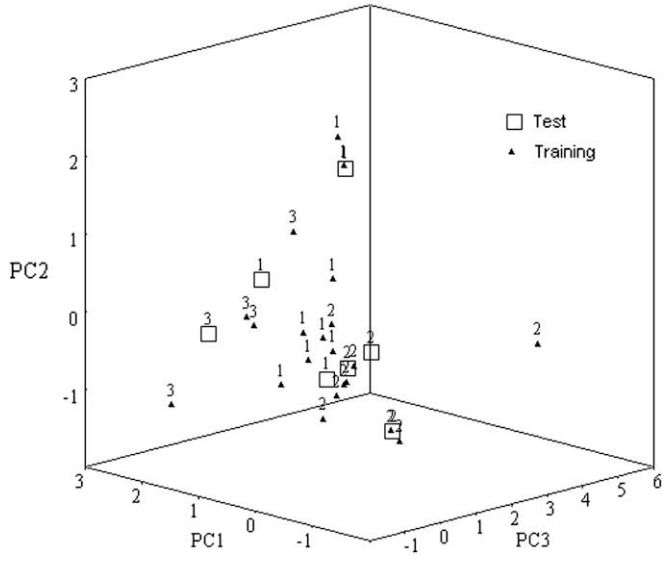

Fig. 1. PCA score plot of first three components of the descriptor matrix shows distribution of training and test set compounds in 3D space and their cluster membership (1, 2 or 3).

In Eq. (1), partition coefficient (log *P*) has positive contribution towards the CYP3A4 inhibition activity. The positive coefficient of this term indicates that increase in hydrophobic property increases the enzyme inhibition activity. For example, miconazole and econazole, having high log *P* values, show higher CYP3A4 inhibition activity. A high value of the parameter J_x, which characterizes the shape of the molecule (covalent radii), is detrimental for the activity. Compounds having high value of this parameter (like compound **17**) have lower inhibitory activity.

3.2. Stepwise MLR

Setting the "stepping criteria" ($F = 4$ for inclusion; $F = 3.5$ for exclusion), the following equation was obtained.

$$\begin{aligned}&\mathrm{pIC}_{50} = 5.002\,(\pm 0.699) + 0.390\,(\pm 0.111)\log P - 0.810\,(\pm 0.296) J_x\\ &-0.400\,(\pm 0.176) S\_aasN\\ &n_{\mathrm{Training}} = 21, R^2 = 0.779, R^2_a = 0.740, F = 19.958\,(df\,3,17),\\ &Q^2 = 0.696, \mathrm{PRESS} = 8.94,\\ &n_{\mathrm{Test}} = 7, R^2_{\mathrm{pred}} = 0.701, r^2 = 0.915, r^2_0 = 0.724, r^2_m = 0.515\end{aligned} \quad (2)$$

The standard errors of regression coefficient are given within parenthesis. Eq. (2) could explain 74.0% of the variance (adjusted coefficient of variation). Eq. (2) differs from Eq. (1) in having an additional term of *S_aasN* (E-state index of the fragment of an aromatic nitrogen with substitution). The presence of an additional term causes a reduction in variance ratio with respect to Eq. (1) while the predicted variance (leave-one-out) increases. The predictive R^2 value for the test set was found to be 0.701. The determination coefficient between the observed and predicted values of the test set compounds was very high (0.915); however, when the intercept was set to zero, this value was substantially reduced to 0.724. The difference between r^2 and r^2_0 is reflected in lower value of r^2_m. However, a high value of R^2_{pred} for this equation reconfirms the inadequacy of this parameter (R^2_{pred}) as indicator of external predictivity [45]. The intercorrelation (r) among predictor variables is shown in Table 5, which suggests the absence of high intercorrelation. The negative coefficient of the term *S_aasN* indicates that compounds having high value of *S_aasN* values have detrimental effect on the inhibitory potency (like compound **27**).

3.3. PLS

Initially starting with all descriptors, the variables with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components and finally the following equations was obtained.

$$\begin{aligned}&\mathrm{pIC}_{50} = 3.681 - 0.361\,J_x + 0.212 S\_aasN + 0.028 S\_sCl + 0.156\log P\\ &+0.128 A\log P98 + 0.001 PMI\_mag + 0.0003 Apol\\ &n_{\mathrm{Training}} = 21, R^2 = 0.731, R^2_a = 0.664, F = 51.67\,(df\,1,19),\\ &Q^2 = 0.660, \mathrm{PRESS} = 9.99,\\ &n_{\mathrm{Test}} = 7, R^2_{\mathrm{pred}} = 0.613, r^2 = 0.690, r^2_0 = 0.611, r^2_m = 0.496\end{aligned} \quad (3)$$

Eq. (3) could explain 66.4% of the variance (adjusted coefficient of variation). The number of latent variables for this PLS equation is 1. The leave-one-out predicted variance of Eq. (3) is lower than the stepwise regression derived model. The predictive R^2 (0.613) and r^2_m

Table 4
Values of important descriptors.

Sl	log *P*	J_X	S_sCH3	S_ssssC	S_aasN	*A* log *P*98	Apol	Sr	PMI-mag	S_sOH	S_sCl
Training set											
1	2.54	2.08	21.51	−4.84	0	1.79	23843.46	1.44	2972.84	57.61	0
2	1.23	1.90	6.68	0	0	4.35	12994.08	1.74	1140.83	0	6.04
3	2.80	1.97	6.88	0	0	3.09	16662.20	1.31	1452.46	0	0
5	2.37	1.20	9.62	−3.82	0	4.36	21156.52	0.75	3592.57	12.15	0
6	2.40	1.04	3.57	−3.84	0	3.20	21939.72	1.57	3467.11	12.06	0
8	0.54	2.10	1.75	0	0	0.77	11241.26	0.19	782.05	0	0
9	1.33	1.97	0	0	0	2.28	10591.44	1.65	474.31	0	0
11	1.39	1.81	1.97	0	0	1.75	10648.56	2.07	1586.11	0	0
12	0.40	1.96	2.05	0	0	−0.50	8603.32	1.56	1058.80	0	0
13	2.86	2.71	6.51	0	0	2.94	11966.50	1.66	523.34	0	0
14	2.23	1.80	7.09	0	0	2.90	14039.98	1.26	1572.99	0	0
16	0.24	2.91	1.92	0	0	−0.02	3321.18	1.11	39.81	0	0
17	0.52	2.92	0	0	0	0.44	3867.96	1.42	58.67	8.57	0
20	2.77	2.17	5.51	−0.55	1.40	3.64	11512.28	1.44	773.10	0	5.83
21	3.50	1.85	2.12	−0.96	1.68	3.80	13363.38	0.64	768.46	0	12.37
22	3.73	1.17	0	−1.08	1.89	3.53	20425.66	0.71	3728.72	0	12.64
23	1.48	2.32	3.80	−0.59	0	1.85	9412.74	0.49	386.44	0	0
24	1.91	1.37	1.81	0	0	1.71	16680.36	3.01	2823.57	0	0
25	5.70	1.75	0	0	1.93	5.65	18135.46	3.81	1655.61	0	24.54
27	0.50	1.90	0	−1.70	2.72	0.75	10610.98	1.15	681.12	10.96	0
28	5.51	1.73	0	0	1.96	4.98	16595.68	2.17	1419.28	0	18.28
Test set											
4	3.79	1.98	12.98	−0.56	0	5.66	16544.98	0.75	2662.03	0	0
7	4.16	1.64	19.79	−1.21	0	2.12	26616.16	1.20	4546.99	11.06	0
10	1.98	2.02	1.22	0	0	1.44	13525.06	0.47	731.34	10.23	0
15	−0.06	2.88	1.94	0	1.89	−0.08	3273.02	1.14	40.61	0	0
18	2.75	2.11	0	0	0	2.77	7598.36	1.86	302.70	0	0
19	2.41	1.41	0	0	0	2.50	11275.86	0.76	794.58	0	0
26	5.48	2.12	0	−0.59	2.12	5.22	15786.72	0.49	671.05	0	6.69

(0.496) values are also inferior to those of Eq. (2). There are four additional terms in this equation compared to Eq. (2). The terms are *S_sCl*, *A* log *P*98, *PMI_mag* and atomic polarizability (*Apol*), and all these terms have positive contributions to the inhibition activity. Compounds with high values of E-state parameter of chloro fragment, partition coefficient (*A* log *P*98), principal moment of inertia and polarizability should have higher CYP3A4 inhibitory activity.

3.4. GFA

Eq. (4) is one of the best obtained from the genetic function approximation (5000 iterations). The terms in the model were linear and linear splines.

$$\begin{aligned} \mathrm{pIC}_{50} &= 4.525\ (\pm 0.247) - 0.038\ (\pm 0.008)S\_sOH + 0.475 \\ &\quad \times (\pm 0.072)\log P + 0.269\ (\pm 0.149)Sr \\ &\quad -1.994\ (\pm 0.291)\ < S\_ssssC + 0.963 > \\ n_{\mathrm{Training}} &= 21, \mathrm{LOF} = 0.561, R^2 = 0.913, R^2_{\mathrm{a}} = 0.891, \\ F &= 41.76\ (df\,4, 16), Q^2 = 0.836, \mathrm{PRESS} = 4.814, n_{\mathrm{Test}} = 7, \\ R^2_{\mathrm{pred}} &= 0.520, r^2 = 0.590, r^2_0 = 0.586, r^2_{\mathrm{m}} = 0.553 \end{aligned} \quad (4)$$

The standard errors of regression coefficients are given within parenthesis. Eq. (4) could explain 89.1% of the variance (adjusted coefficient of variation). Though leave-one-out cross-validation Q^2 is very high (0.836), the external prediction statistics are only moderate. The intercorrelation (r) among predictor variables is shown in Table 5, which suggests the absence of high intercorrelation. The negative coefficient of $\langle S\_ssssC + 0.963 \rangle$ indicates that values of E-state index of ⟩C⟨ fragment (*S_ssssC*) lower than −0.963 are conducive for the inhibition activity. For example, compounds **1, 5, 6, 22** and **27** having E-state values lower than 0.963 have higher CYP3A4 inhibition activity. The positive coefficient of Sr (superdelocalizability) in Eq. (4) indicates that increase in electrophilic property is conducive for the CYP3A4 inhibitory activity (compound **25**). Higher value of E-state index of the hydroxyl group (*S_sOH*) is detrimental for the activity (like compound **1**).

3.5. G/PLS

Eq. (5) is one of the best equations obtained from genetic partial least squares (2000 crossovers, scaled variables, number of components 4, initial equation length 5, no fixed length of the final

Table 5
Intercorrelation (r) matrix.

	J_x	S_sCH3	S_ssssC	S_aasN	log *P*	*A* log *P*98	Sr	*Apol*	PMI_mag	S_sCl	S_sOH
J_x	1.000	0.011	0.409	−0.227	−0.388	−0.475	−0.152	−0.768	−0.832	−0.247	−0.026
S_sCH3	0.011	1.000	−0.634	−0.376	0.023	0.084	−0.150	0.478	0.347	−0.304	0.781
S_ssssC	0.409	−0.634	1.000	0.019	−0.070	−0.109	0.225	−0.655	−0.647	0.176	−0.802
S_aasN	−0.227	−0.376	0.019	1.000	0.518	0.373	0.131	0.147	0.008	0.701	−0.110
log *P*	−0.338	0.023	−0.070	0.518	1.000	0.845	0.429	0.615	0.367	0.824	−0.020
A log *P*98	−0.475	0.084	−0.109	0.373	0.845	1.000	0.319	0.608	0.370	0.698	−0.117
Sr	−0.152	−0.150	0.225	0.131	0.429	0.319	1.000	0.177	0.123	0.432	−0.068
Apol	−0.768	0.478	−0.655	0.174	0.615	0.608	0.177	1.000	0.881	0.287	0.462
PMI_mag	−0.832	0.347	−0.647	0.008	0.367	0.370	0.123	0.881	1.000	0.103	0.384
S_sCl	−0.274	−0.304	0.176	0.701	0.824	0.698	0.432	0.287	0.103	1.000	−0.209
S_sOH	−0.026	0.781	−0.802	−0.110	−0.020	−0.117	−0.068	0.462	0.384	−0.209	1.000

equation and other default settings). The terms in the model were linear and linear splines.

$$\begin{aligned} \mathrm{pIC}_{50} &= 6.650 + 0.502 < Sr - 1.573 > -0.084 S\_sCH3 - 2.400 \\ &\quad < S\_ssssC + 0.553 > -0.642 < 3.73 - \log P > \\ n_{\mathrm{Training}} &= 21, \mathrm{LSE} = 0.170, R^2 = 0.916, R_a^2 = 0.842, \\ &\quad F = 62.48\ (df 3, 17), Q^2 = 0.827, \mathrm{PRESS} = 5.085, n_{\mathrm{Test}} = 7, \\ &\quad R^2_{\mathrm{pred}} = 0.600, r^2 = 0.665, r_0^2 = 0.649, r_m^2 = 0.581 \end{aligned} \quad (5)$$

Eq. (5) could explain 84.2% of the variance (adjusted coefficient of variation). The internal validation statistic Q^2 for this model is very good (0.827). The predictive R^2 value for the test set is found to be 0.600. The positive coefficient of ⟨Sr-1.573⟩ indicates that the value of superdelocalizability should be greater than 1.573 for the better inhibition activity. The negative coefficient of ⟨*S_ssssC* + 0.553⟩ and ⟨3.73−log *P*⟩ indicates that the values of *S_ssssC* and log *P* should be lower than −0.553 and 3.73, respectively, for the compounds to be effective inhibitors.

3.6. ANN

For the development of better predictive models, non-linear modeling with artificial neural network was also tried. We have formed the network with the training set using backpropagation in the first phase and conjugate gradient descent in the second phase. The network so developed was used for prediction of CYP3A4 enzyme inhibitory activity values of the test set compounds. Using different iterations of backpropagation and conjugate gradient descent, varying numbers of hidden layers and its units per layer, a number of models were developed. Neural networks were optimized using a training subset. A separate subset (the selection subset) was used to halt training to mitigate over-learning, or to select from a number of models trained with different parameters. Then, a third subset (the test subset) was used to perform an unbiased estimation of the network's likely performance. Initially neural network was developed with all available descriptors. Here, in Table 6, we have presented 5 best ANN models using different iterations and different number of elements in the hidden layer. Initially neural network models were developed with all available descriptors. Then we have developed models based on descriptors found important in other statistical methods (FA-MLR, stepwise MLR, PLS, GFA, G/PLS). Table 6 shows five network models out of which the first three were developed by taking all descriptors and remaining two were from selected descriptors (thirteen in number). Numbers of iterations selected for backpropagation and conjugate gradient descent for different models are listed in Table 6. Initialization method selected for network was random uniform. Weight decay was regularized in both phases (decay factor = 0.01, scale factor = 1). Learning rate and momentum of each epoch were adjusted to 0.01 and 0.3, respectively. The numbers of cross-validated resampling were set to 10, 15 and 20. During 15 resampling, numbers of cases selected for training, selection and test were 9, 5 and 1, respectively, and for 20 resampling, the values were 12, 7 and 1, respectively. In case of 10 resampling, numbers of cases selected for training, selection and test were 12, 6 and 2, respectively. The best model according to r^2 was model N4 (selected set of descriptors) while the best model according to r_m^2 was model 1 (all descriptors). In the best network (model N4 using selected descriptors, based on the determination coefficient between the observed and predicted values of the test set compounds), 1 hidden layer of 2 elements was used. Numbers of iterations selected for backpropagation and conjugate gradient descent were 450 and 300, respectively. Initialization method selected for network was random uniform. For model N1 based on all descriptors (the best model according to r_m^2 value), 1 hidden layer of 6 elements was used. Numbers of iterations selected for backpropagation and conjugate gradient descent were 450 and 300, respectively. The external validation qualities of the ANN model N1 were compared to those of linear models in Table 7.

3.7. Further test on external validation

The models were also subjected to the test for criteria of external validation as suggested by Golbraikh and Tropsha [18]. These authors [18] have recommended that in addition to a high value of cross-validated R^2 (Q^2), the correlation coefficient r between the predicted and observed activities of compounds from an external test set should be close to 1. At least one (but better both) of the determination coefficient for regressions through the origin (observed versus predicted activities, or, predicted versus observed activities), i.e., r_0^2 or $r_0'^2$ should be close to r^2. Furthermore, at least one slope of regression lines (k or k') through the origin should be close to 1. Models are considered acceptable, if they satisfy all of the following conditions: (i) $Q^2 > 0.5$, (ii) $r^2 > 0.6$, (iii) r_0^2 or $r_0'^2$ is close to r^2, such that $[(r^2 - r_0^2)/r^2]$ or $[(r^2 - r_0'^2)/r^2] < 0.1$ and $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$. A list of values of different parameters for different models as recommended by Golbraikh and Tropsha [18] are given in Table 8. The ANN model N1, stepwise regression and PLS derived models do not pass the test (*vide* values for Sl. No. 6, Table 8).

In case of the genetic models, differences between leave-one-out Q^2 and R^2_{pred} values are considerably high. However, it was previously shown that there may not exist any correlation between internal validation and external validation parameters [18]. The regression derived models were also subjected to randomization test at 90% confidence level and the results are shown in Table 9. For each regression model, the mean value of random models is significantly lower that the corresponding value of the nonrandom model. This suggests that the models are not obtained by chance.

4. Overview

Different statistical methods like stepwise MLR, PLS, FA-MLR, GFA-MLR, G/PLS have been applied for linear modeling of CYP3A4

Table 6
Comparative features of selected ANN models.

Model No.	No. of hidden layers	No. of units in hidden layer	No. of cross-validated resampling	No. of epochs in backpropagation followed by conjugate gradient descent	Absolute error mean	Squared correlation coefficient (r^2) between Obs. and Pred. values of the test set	r_0^2 (test set)	r_m^2 (test set)
N1	1	6	20	450,150	0.841	0.791	0.693	0.543
N2	1	6	15	450,200	0.897	0.675	0.614	0.508
N3	1	8	15	600,200	0.828	0.825	0.672	0.502
N4	1	2	15	450,300	0.769	0.880	0.706	0.513
N5	1	2	10	300,100	0.834	0.803	0.681	0.523

Table 7
Comparison of statistical qualities of different models (the best value of each column is shown in bold face).

Type of statistical method	R^2 (training set)	Q^2 (training set)	R^2_{pred} (test set)	r^2 (test set)	r^2_0 (test set)	$r^2_{m(test)}$ (test set)	$r^2_{m(LOO)}$ (training set)	$r^2_{m(overall)}$ (whole set)	RMSEP (test set)
FA-MLR	0.711	0.644	0.573	0.667	0.637	0.551	0.485	0.525	0.993
Stepwise	0.779	0.696	**0.701**	**0.915**	**0.724**	0.515	0.530	0.597	**0.831**
PLS	0.731	0.660	0.613	0.690	0.611	0.496	0.628	0.625	0.945
GFA	0.913	**0.836**	0.520	0.590	0.586	0.553	0.691	0.633	1.052
G/PLS	**0.916**	0.827	0.600	0.665	0.649	**0.581**	**0.771**	**0.735**	0.960
ANN (N1)			0.672	0.791	0.693	0.543			0.870

enzyme inhibitors using physicochemical, structural, spatial, electronic and topological descriptors (including the E-state indices). The whole data set was divided into training set (21 compounds) and test set (7 compounds) based on K-means clustering of the standardized descriptor matrix and models were developed from the training set. The predictive ability of the models was judged from the prediction of the CYP3A4 inhibition activity test set compounds. A comparison of statistical quality of different models is given in Table 7. In all the five modeling techniques, *n*-octanol/water partition coefficient (log *P*) emerged as an important descriptor which is in agreement with previously published paper modeling free energy changes of binding with CYP3A4 enzyme using ln *P* [23]. The G/PLS model showed an optimum range of log *P* values for an effective inhibitor. The above modeling techniques showed the importance of different topological parameters (like Balaban J_x, E-state indices like *S_aasN*, *S_ssssC*, *S_sOH*, *S_sCl*) and electronic parameters (Sr, *Apol*, *PMI_mag*), etc. The best linear model (based on external validation parameter R^2_{pred}) obtained from the training set (stepwise regression) showed good internal ($Q^2 = 0.696$) and external predictive power ($R^2_{pred} = 0.701$). Other models like FA-MLR ($Q^2 = 0.644$, $R^2_{pred} = 0.573$), PLS ($Q^2 = 0.660$, $R^2_{pred} = 0.613$), GFA ($Q^2 = 0.836$, $R^2_{pred} = 0.520$) and G/PLS ($Q^2 = 0.827$, $R^2_{pred} = 0.600$) are of statistical significance but their performance is inferior to the stepwise regression derived model in terms of external validation statistics (considering only R^2_{pred} values). The GFA derived model shows the best internal validation statistic ($Q^2 = 0.836$). Multicollinearity for MLR models was also checked (Table 10). In all cases, variable inflation factor was found to be less than 10 and tolerance value more than 0.1 which suggests absence of multicollinearity. Again, the best model (according to r^2_m) obtained from ANN (model N1) showed a good r^2 value (squared regression coefficient between observed and predicted values) for the test set compounds (0.791) which were superior to models derived from other statistical methods except the stepwise regression ($r^2 = 0.915$). The best r^2 (0.880) value from ANN models was found in case of model N4 derived from selected descriptors. Again, based upon the r^2_m values, which penalizes a model for large differences between observed and predicted values, the G/PLS derived model was found to be superior ($r^2_m = 0.581$) in comparison to the other models listed in Table 7. The ANN model N1, stepwise regression and PLS derived models do not pass the test of external validation as recommended by Golbraikh and Tropsha [18]. Considering different external validation parameters (R^2_{pred}, r^2_m) and criteria recommended by Golbraikh and Tropsha [18], the G/PLS derived model appears to be the best predictive model for this data set. The calculated values of CYP3A4 inhibition activity of all compounds according to different models are shown in Table 1.

Previously the concept r^2_m was applied only to the test set prediction [45], but it can as well be applied for training set if one considers the correlation between observed and leave-one-out (LOO) predicted values of the training set compounds. More interestingly, this can be used for the whole set considering LOO-predicted values for the training set and predicted values of the test set compounds. The advantages of such consideration are:

1. Unlike external validation parameters (R^2_{pred} etc.), the r^2_m (whole) statistic is not based only on limited number of test set compounds. It includes prediction for both test set and training set (using LOO predictions) compounds. Thus, this statistic is based on prediction of comparably large number of compounds. In many cases, test set size is considerably small and regression based external validation parameter may be less reliable and highly dependent on individual test set observations. In such cases, the r^2_m (whole) statistic may be advantageous.
2. In many cases, comparable models are obtained where some models show comparatively better internal validation parameters and some other models show comparatively superior external validation parameters. This may create problem in selecting the final model. The r^2_m (whole) statistic may be used for selection of the best predictive models from among comparable models.

For the present QSAR study, we have determined r^2_m values for both training (based on LOO-predicted values) and test sets and also for the whole set for the regression based models and the results are shown in Table 7. Based on the r^2_m (whole) statistic, the G/PLS model is found to be the best model.

Table 8
External validation characteristics of different models according to Golbraikh and Tropsha [18].

Statistical parameters	Model number						
Sl. No.	Parameters	1	2	3	4	5	N1
1	r^2	0.667	0.915	0.690	0.590	0.665	0.791
2	r^2_0	0.637	0.724	0.611	0.586	0.649	0.693
3	$r_0'^2$	0.169	0.043	−0.215	0.429	0.297	0.131
4	$(r^2 - r^2_0)/r^2$	0.045	0.209	0.144	0.007	0.024	0.124
5	$(r^2 - r_0'^2)/r^2$	0.747	0.953	1.316	0.273	0.553	0.834
6	Minimum of 4 and 5	0.045	0.209	0.144	0.007	0.024	0.124
7	k	0.920	0.949	0.967	0.919	0.929	0.950
8	k'	1.045	1.023	0.991	1.040	1.036	1.018

Table 9
Results of randomization test of the developed models.

Eq. No.	(1)	(2)	(3)	(4)	(5)
Modeling technique	FA-MLR	Stepwise	PLS	GFA	G/PLS
R from nonrandom model	0.843	0.883	0.855	0.956	0.957
Confidence level	90%	90%	90%	90%	90%
Mean value of R from random trials ± standard deviation	0.227 ± 0.139	0.229 ± 0.120	0.099 ± 0.186	0.444 ± 0.101	0.135 ± 0.254

Table 10
Test for multicollinearity of MLR models.

Eq. number	Variable	VIF	Tolerance
(1)	log P	1.178	0.849
	J_x	1.178	0.849
(2)	log P	1.527	0.655
	J_x	1.179	0.848
	S_aasN	1.368	0.731
(4)	log P	1.521	0.657
	S_sOH	1.440	0.694
	Sr	1.884	0.531
	⟨S_ssssC + 0.963⟩	2.067	0.484

5. Conclusions

Among the developed linear and non-linear models for CYP3A4 inhibitory activity of diverse functional compounds, the best equation based on internal validation was obtained with GFA while the best model based on external validation (R^2_{pred} as objective function) was obtained from stepwise regression. However, the stepwise regression model does not pass the criteria recommended by Golbraikh and Tropsha [18]. Considering different external validation parameters, the G/PLS derived model appears to be the best predictive model. The G/PLS model also shows the highest value of r^2_m statistic for the whole set considering LOO-predicted values of the training set and predicted values of the test set. The parameter $r^2_{m(overall)}$ appears to be advantageous over other internal and external validation parameters in that it is based on prediction of both training and test set compounds and thus involving more compounds in the prediction process, and it is useful for the selection of the best model from among the comparable models with different patterns of internal and external validation parameters. The developed models in this communication show importance of log P along with topological and electronic parameters in describing the CYP3A4 inhibitory activity. Non-linear models were not found to be superior to linear models for this data set. The developed models may be helpful in designing potent CYP3A4 inhibitors and predicting CYP3A4 inhibition potential of novel drug candidates.

Acknowledgements

This research work is supported by a Major Research Grant of University Grant Commission (UGC), New Delhi. One of the authors (P.P. Roy) thanks the UGC, New Delhi for a fellowship.

References

[1] S.A. Wrighton, M.V. Branden, J.C. Stevens, L.A. Shipley, Drug Metab. Rev. 25 (1993) 453–484.
[2] S. Ekins, G. Bravi, S. Binkley, J.S. Gillespie, B.J. Ring, J.H. Wikel, S.A. Wrighton, J. Pharmacol. Exp. Ther. 290 (1999) 429–438.
[3] F.J. Gonzalez, Pharmacol. Ther. 45 (1990) 1–38.
[4] F.P. Guengerich, FASEB J. 6 (1992) 745–748.
[5] S. Rendic, Drug Metab. Rev. 34 (2002) 83–448.
[6] K.J. Silvers, T. Chazinski, M.E. McManus, S.L. Bauer, F.J. Gonzalez, H.V. Gelboin, P. Maurel, P.C. Howard, Cancer Res. 15 (1992) 17–624.
[7] S. Rendic, F.J. Di Carlo, Drug Metab. Rev. 29 (1997) 413–580.
[8] T. Shimada, H. Yamazaki, M. Mimura, Y. Inui, F.P. Guengerich, J. Pharmacol, Exp. Ther. 270 (1994) 414–423.
[9] B.S. Singh, Q.L. Shen, J.M. Walker, P.R. Sheridan, J. Med. Chem. 46 (2003) 1330–1336.
[10] S.A. Wrighton, J.C. Stevens, Crit. Rev. Toxicol. 22 (1992) 1–21.
[11] F.P. Guengerich, Annu. Rev. Pharmacol. Toxicol. 39 (1999) 1–17.
[12] P.A. Williams, J. Cosme, D.M. Vinkovic, A. Ward, H.C. Angove, P.J. Day, C. Vonrhein, I.J. Tickle, H. Jhoti, Science 305 (2004) 683–686.
[13] D.F.V. Lewis, B.G. Lake, Toxicology 125 (1998) 31–44.
[14] W. Tong, H. Hong, Q. Xie, L. Shi, H. Fang, R. Perkins, Curr. Comput. Aided Drug Des. 2 (2005) 195–205.
[15] L. He, P.C. Jurs, J. Mol. Graphics Mod. 23 (2005) 503–523.
[16] T. Ghafourian, M.T.D. Cronin, SAR QSAR Environ. Res. 16 (2005) 171–190.
[17] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[18] A. Golbraikh, A. Tropsha, J. Mol. Graphics Mod. 20 (2002) 269–276.
[19] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 312–317.
[20] D. Itokawa, T. Nishioka, J. Fukushima, T. Yasuda, A. Yamauchi, H. Chuman, QSAR Comb. Sci. 26 (2007) 828–836.
[21] J.M. Kriegl, T. Arnhold, B. Beck, T. Fox, J. Comput. Aided Mol. Des. 19 (2005) 189–201.
[22] J.M. Kriegl, T. Arnhold, B. Beck, T. Fox, QSAR Comb. Sci. 24 (2005) 491–502.
[23] D.F.V. Lewis, B.G. Lake, M. Dickins, J. Enzyme Inhib. Med. Chem. 21 (2006) 127–132.
[24] K. Roy, P.P. Roy, Chem. Biol. Drug Des., 71 (2008) 464–473.
[25] Cerius2 Version 4.10.http://www.accelrys.com/cerius2.
[26] H. Kubinyi, F.A. Hamprecht, T. Mietzner, J. Med. Chem. 41 (1998) 2553.
[27] J.T. Leonard, K. Roy, QSAR Comb. Sci. 25 (2006) 235–251.
[28] A. Golbraikh, A. Tropsha, Mol. Divers. 5 (2000) 231–243.
[29] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Chemom. Intell. Lab. Syst. 84 (2006) 69–74.
[30] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Chemom. Intell. Lab. Syst. 33 (1996) 35.
[31] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Edward Arnold, London, 2001.
[32] R.B. Darlington, Regression and Linear Models, McGraw-Hill, New York, 1990.
[33] R. Franke, Theoretical Drug Design Methods, Elsevier, Amsterdam, 1984, pp. 184–195.
[34] R. Franke, A. Gruska, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 113–163.
[35] S. Wold, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 195–218.
[36] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, J. Med. Chem. 44 (2001) 3254–3263.
[37] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.
[38] W.J. Dunn III, D. Rogers, in: J. Devillers (Ed.), Genetic Algorithms in Molecular Modeling, Academic Press, London, 1996, pp. 109–130.
[39] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci. 37 (1997) 306–310.
[40] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, 1999.
[41] SPSS: Statistical Software, SPSS Inc.; IL, USA.
[42] MINITAB: Statistical Software, MINITAB Inc., PA, USA.
[43] STATISTICA: Statistical Software, STATSOFT Inc., OK, USA.
[44] G.W. Snedecor, W.G. Cochran, Statistical Methods, Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi, 1967, pp. 381–418.
[45] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.
[46] S.S. Kulkarni, V.M. Kulkarni, J. Med. Chem. 42 (1999) 373–380.
[47] Y. Tang, H.L. Jiang, K.X. Chen, R.Y. Ji, Ind. J. Chem. 35B (1996) 325–332.
[48] A.K. Debnath, in: A.K. Ghose, V.N. Viswanadhan (Eds.), Combinatorial Library Design and Evaluation, Marcel Dekker, Inc., New York, 2001, pp. 73–129.
[49] K. Roy, Exp. Opin. Drug Discov. 2 (2007) 1567–1577.
[50] P.P. Roy, J.T. Leonard, K. Roy, Chemom. Intell. Lab. Syst. 90 (2008) 31–42.
[51] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, M. McDowell, P. Robert, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.