J. Chem. Sci. Vol. 125, No. 3, May 2013, pp. 483–493. © Indian Academy of Sciences.

Diverse models for the prediction of CDK4 inhibitory activity of substituted 4-aminomethylene isoquinoline-1, 3-diones

MONIKA GUPTA[a] and A K MADAN[b,*]

[a]Faculty of Pharmaceutical Sciences, M D University, Rohtak 124 001, India
[b]Faculty of Pharmaceutical Sciences, Pandit B D Sharma University of Health Sciences, Rohtak 124 001, India
e-mail: madan_ak@yahoo.com

MS received 15 June 2012; revised 6 November 2012; accepted 17 December 2012

Abstract. In the present study, both *classification* and *correlation* approaches have been successfully employed for development of models for the prediction of CDK4 inhibitory activity using a dataset comprising of 52 analogues of 4-aminomethylene isoquinoline-1,3-(2H,4H)-dione. Decision tree, random forest, moving average analysis (MAA), multiple linear regression (MLR), partial least square regression (PLSR) and principal component regression (PCR) were used to develop models for prediction of CDK4 inhibitory activity. The statistical significance of models was assessed through specificity, sensitivity, overall accuracy, Mathew's correlation coefficient (MCC), cross validated correlation coefficient, F test, r^2 for external test set (pred_r^2), coefficient of correlation of predicted dataset (pred_ r^{2}Se) and intercorrelation analysis. High accuracy of prediction offers proposed models a vast potential for providing lead structures for the development of potent therapeutic agents for CDK4 inhibition.

Keywords. Decision tree; E-state contribution indices; augmented eccentric connectivity topochemical index; molecular connectivity index; connective eccentricity topochemical index.

1. Introduction

Dysregulation of cell-cycle leading to an uncontrolled cellular proliferation is a universal characteristic of cancer.[1] Progression of cells through the cell-cycle is dependent on the formation of specific protein kinase complexes called as cyclin dependent kinases (CDKs) which form in a cyclical fashion.[2] The CDKs are a family of heterodimeric Ser/Thr protein kinases each consisting of a catalytic CDK subunit and an activating cyclin subunit.[3,4] Activities of CDKs are controlled by association with cyclins and reversible phosphorylation reactions.[5] The association of the CDKs with their requisite cyclin partner results in the CDKs adopting a substrate-specific catalytic subunit.[6] The biological activity of CDKs is negatively controlled by direct interactions with proteins referred to as CDK inhibitors. CDK inhibitors are divided into two major families: the Ink4 family, which specifically inhibit cyclin D-associated kinases (CDK 4 and 6) and the Cip/Kip family which inhibit most of the CDKs.[7] The CDK4 initiate the functional change from quiescence (G_0) to proliferation. The major Cdk4/cyclin D substrate is the product of the retinoblastoma gene (pRB).[8] During G1, pRB induces the members of a family of cell cycle regulatory transcription factors, collectively referred to as E2Fs, to activate the transcription of genes whose products are required for S phase.[9,10] Alterations in the cascade involving CDK4, CDK6, cyclin D, Ink4, pRB and E2F have been observed in more than 80% of human cancers. Therefore, the development of selective CDK4 inhibitors is a promising approach for cancer therapy.[11–14] In addition, the identification of specific amino acid residues of the kinase superfamily around the ATP-binding pocket of CDK4 have enabled the researchers to develop potent and selective CDK4 inhibitors.[15] The structure-based design of potent and selective CDK4 inhibitors led to the development of several classes of compounds, including pyrido[2,3-d] pyrimidines,[16] 2-anilinopyrimidines, diaryl ureas, benzoyl-2,4-diaminothiazoles, indolo[6, 7-a] carbazoles,[17] pyrrolo[3,4-c]carbazoles[18,19] and oxindoles.[20] The various CDK inhibitors that are currently in pre-clinical and clinical trials are UCN-01, PD 0183812, Flavopiridol,[21] R547,[22] AT7519,[23] SNS-032, Roscovitine (CYC202) JNJ-7706621, AG-024322, AT7519, AZD5438, P1446A-05, P276-00 and PD-0332991. Among the above mentioned, the compound PD 0332991 and P1446A-05 are selective CDK4 inhibitors.[24,25] Despite more than a decade of investigation, none of the CDK inhibitors resulted in drug approval owing to low activity and toxicity in the

[*] For correspondence

clinical trials.[12,25] Therefore, there is a strong need to develop potent and selective CDK inhibitors.

The search for novel active compounds and the optimization of these compounds to increase their activity or reduce their toxicity requires huge sum of money and time as a large number of compounds are needed to be synthesized and subsequently evaluated for biological activity.[26] The technological contributions like high-throughput synthesis and screening have enhanced the impact of computational chemistry on the drug discovery process by efficiently managing costly resources and dramatically shortening the drug discovery cycle time.[27] A contemporary trend over the years has been the expansion of the QSAR concept to encompass a variety of pharmaceuticals. The major goal of QSAR research is to assign to the structure a number or a set of numbers which must correlate well with the biological activity value measured experimentally. This numerical representation of the structure which describes the structure is called a molecular descriptor. Molecular descriptors can be derived in either empirical or non-empirical ways.[28] The traditional QSAR is normally based on large number of empirical parameters. Non-empirical parameters of chemical structure derived from graph theoretic formalism are being used more frequently by many researchers in QSAR studies pertaining to molecular design and pharmaceutical drug design. When a single number represents a graph invariant, it is known as topological index.[29] In contrasting graph theoretical schemes to traditional quantitative structure-activity relationship (QSAR) methods, one cannot fail to observe the complementarity of the two approaches. But the prime distinction between graph theoretical schemes and traditional QSAR is that the former is 'structure-explicit' while the latter is 'structure-cryptic'.[30]

In the present study, both classification and correlation approaches have been successfully employed for development of models for the prediction of CDK4 inhibitory activity of 4-aminomethylene isoquinoline-1,3-(2*H*,4*H*)-dione derivatives.[31]

2. Experimental

2.1 *Data set*

All the 52 4-aminomethylene isoquinoline-1,3-(2*H*, 4*H*)-dione derivatives reported by Tsou *et al.* as CDK4 inhibitors were selected as a data set for the purpose of present study.[31] The basic structures for the said derivatives are shown in figure 1 and various substituents enlisted in table S1.

2.2 *Molecular descriptors (MDs)*

MDs of diverse nature were used in the current study. These included physico-chemical descriptors, path count, path cluster, estate contribution descriptors, polar surface area descriptors, element counts, topological descriptors and a variety of alignment independent descriptors. All computational work was performed on Apple workstation (8-core processor) using V-life MDS QSAR plus developed by V life sciences technologies Pvt. Ltd, Pune, India. The values of other MDs which are not the part of V-life MDS QSAR plus were computed using an in-house computer program.

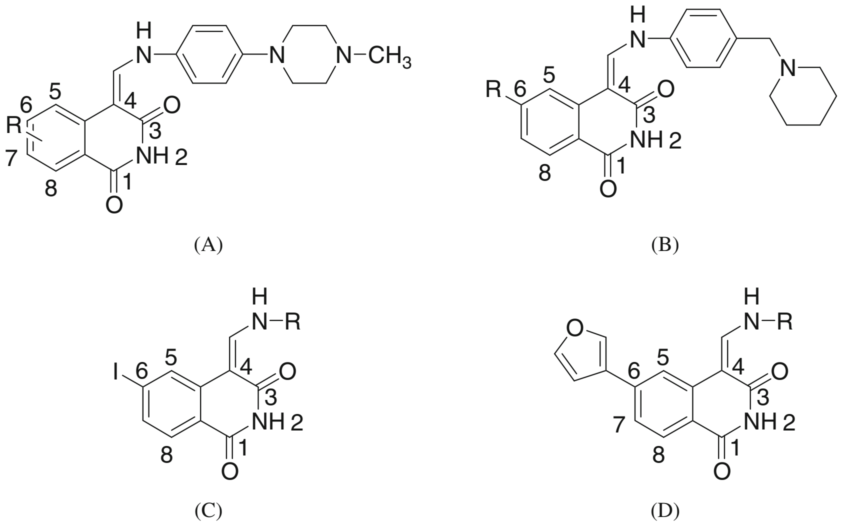


Figure 1. Basic structures and arbitrary atom numbering scheme for the 4-aminomethylene isoquinoline-1,3-dione.

MDs with significant degenerate values were omitted from a large pool of descriptors initially calculated both through V-life MDS QSAR plus software and an in-house computer program. For the remaining descriptors, a pair wise correlation analysis was carried out (one of any two indices with r $\geq$ 0.97 was excluded to reduce redundant information). The said exclusion method was used to reduce the collinearity and correlation between descriptors.

Finally, 46 descriptors were shortlisted on the basis of non-correlating nature and classification ability and subsequently employed for present study are enlisted in table S2.[32–65]

3. Classification techniques

3.1 *Decision tree (DT)*

DT provides a useful solution for many problems of classification where the information contained in the datasets is relatively complex.[66] In the present study, decision tree was grown to identify the importance of molecular descriptors. In DT, the molecules at each parent node are categorized or classified, based upon the descriptor value, into two child nodes. The prediction for molecule reaching a given terminal node is obtained by majority vote of molecules reaching the same terminal node in a training set.[67] In the present study, R program (version 2.1.0) along with the RPART library was utilized to grow DT. The active compounds were labelled as ‘A’ (n = 15) and the inactive compounds were similarly labelled as ‘B’ (n = 37). Each analogue was assigned a biological activity which was subsequently compared with the reported CDK4 inhibitory activity.

3.2 *Random forest (RF)*

Random forests (RF) were grown for CDK4 inhibitory activity. RF grows numerous classification trees. RF is an ensemble of unpruned decision trees created by using bootstrap samples of the training data and random subset of variables to define the best split at each node (tree fork).[68] Besides preserving most of the appealing features of DT, RF performs a type of cross-validation in parallel with training step by using so called Out-Of-Bag (OOB). OOB data is used to calculate prediction accuracy. In the present study, the RFs were grown separately for CDK4 inhibitory activity with the R program (version 2.1.0) using the random forest library.

3.3 *Moving average analysis (MAA)*

MAA of correctly predicted compounds is the basis of development of single molecular descriptor based model.[40,69] For the selection and evaluation of range specific features, exclusive activity ranges were discovered from the frequency distribution of response level and subsequently identify the active range by analysing the resulting data by maximization of the moving average with respect to active compounds ($<35\%$ = inactive, 35–65% = transitional, $>65\%$ = active). For the purpose of MAA-based models the compounds having reported IC_{50} values of ≤ 0.25 μM were considered to be active (and labelled as ‘A’ ($n = 18$)) while those possessing IC_{50} values > 0.25 μM were treated to be inactive (and labelled as ‘B’ ($n = 38$)). The CDK4 inhibitory activity assigned to each compound was subsequently compared with the reported biological activity.[31] The average IC_{50}(μM) values for each range were also calculated.

4. Correlation techniques

4.1 *Multiple linear regression (MLR)*

MLR is also commonly referred to as the linear free-energy relationship (LFER). This method represents an extension of the simple regression analysis to more than one dimension.[70] MLR normally generates QSAR equation by performing standard multivariable regression calculations to facilitate identification of the dependence of a drug property on any or all of the descriptors under investigation.[71]

4.2 *Partial least square regression (PLSR)*

PLSR is an iterative regression procedure that produces its solutions based on linear transformation of a large number of original MDs to a small number of new orthogonal terms called latent variables.[72] PLSR gives statistically robust solution even if the independent variables are highly interrelated among themselves or when the independent variables exceed the number of observations.

4.3 *Principal component regression (PCR)*

Principal component analysis (PCA) is a substitute for MLR when explanatory variables are correlated. It is another data reduction technique that generates a new set of orthogonal descriptors referred to as principal

Table 1. Confusion matrix for CDK4 inhibitory activity using models based on decision tree and random forest.

			Number of compound predicted					
Model	Description	Ranges	Active	Inactive	Specificity (%)	Sensitivity (%)	MCC	OOB (%)
Decision tree	Training set	Active	13	2	100	86.6	0.9	–
		Inactive	0	37				
	Cross validated set	Active	9	6	72.9	60	0.3	–
		Inactive	10	27				
Random forest		Active	12	3	97.2	80	0.7	8
		Inactive	1	36				

Table 2. Proposed MAA based models for the prediction of CDK4 inhibitory activity.

Index	Nature of range	Index value	Total compounds in the range	Number of compounds predicted correctly	Overall accuracy of prediction	Average IC_{50} (μM)
SssOE	Lower inactive	<24.868	17	17	93.3	5.17
	Active	24.868 to 25.02	11	10		0.13
	Transitional	>25.02–25.1	8	NA		NA
	Upper inactive	>25.1 to ≥46.56	16	14		18.38
$^{Ac}\xi^{C}$	Inactive	<19.28	21	21	90.6	5.71
	Transitional	19.28 to <20.46	20	NA		NA
	Active	20.46 to <22.04	11	8		0.18
χ^{A}	Lower inactive	<14.47	18	18	94.4	5.19
	Lower transitional	14.47 to <14.89	8	NA		NA
	Active	14.895 to <15.26	8	7		0.12
	Upper inactive	15.26 to <16.07	9	8		7.33
	Upper transitional	≥16.07	9	NA		NA
$C^{\xi}c$	Inactive	<5.42	20	20	96.2	5.17
	Transitional	5.42 to <5.64	25	NA		NA
	Active	5.64 to <5.85	7	6		0.10

NA: Not applicable
*Values in brackets are based upon correctly predicted analogues in the particular range

components (PCs) which describe most of the information contained in the independent variables in order of decreasing variance. Consequently, PCA reduces dimensionality of a multivariate data set of descriptors to the actual amount of data available. When PCs are employed as the independent variables to perform a linear regression, the method is termed as PCR.[73]

In the present study, the dataset was divided into training and test set by random selection method for MLR, PLSR and PCR methods using pIC_{50} [$pIC_{50} = -\log(IC_{50} * 10^{-6})$ as dependent variable and various descriptors as independent variables. These models were generated using a training set of 36 molecules. Predictive power of the resulting models was evaluated by test set of 16 molecules. The biological activities of all the compounds had uniform distribution ranging from 0.027 to 50 μM.

5. Data analysis and validation

The validation of the DT based models and self-consistency test were performed by 10-fold cross validation (CV) method. For classification models the sensitivity and specificity values were calculated which represent the classification accuracies for the active and inactive compounds, respectively. The randomness of

Table 3. Intercorrelation matrix for MDs used in MAA.

	SssOE	$^{Ac}\xi^{C}$	χ^{A}	$C^{\xi}c$
SssOE	1	0.15	0.04	−0.2
$^{Ac}\xi^{C}$		1	0.46	0.025
χ^{A}			1	0.36
$C^{\xi}c$				1

model was also determined by calculating Mathew's correlation coefficient (MCC). The MCC values ranging between -1 and $+1$ indicate the prediction potential of model. MCC takes into consideration both the sensitivity and specificity and is generally used as a balanced measure in dealing with data imbalance situation.[74] The intercorrelation between estate contribution index (SssOE), augmented eccentric connectivity topochemical index $\left({}^{A}\xi_{C}^{C}\right)$, molecular connectivity index (χ^{A}) and connective eccentricity topochemical index $\left(C^{\xi_C}\right)$ was also investigated. The degree of correlation can be appraised by correlation coefficient 'r'. Pairs of MDs with $r \geq 0.97$ are normally considered highly intercorrelated, those with $0.90 \leq r \leq 0.97$ are appreciably correlated, those with $0.50 \leq r \leq 0.89$ are weakly correlated and finally the pairs of indices with $r < 0.50$ are not intercorrelated.[75,76] Results are summarized in tables 1–3 and figures 2 and 3.

The following statistical measures were used to correlate biological activity and molecular descriptors for correlation models; n, number of molecules; k, number of descriptors in a model; df degree of freedom; r^2, coefficient of correlation; q^2, cross validated r^2; pred_r^2, r^2 for external test set; pred_r^2Se, coefficient of correlation of predicted dataset; Z score, Z score calculated by the randomization test; best _ran_ r^2; best _ran_q^2, highest q^2 value in randomization test; α, statistical significance parameter obtained by randomization test. Validation was done to study the internal stability and predictive ability of the correlation models. Internal

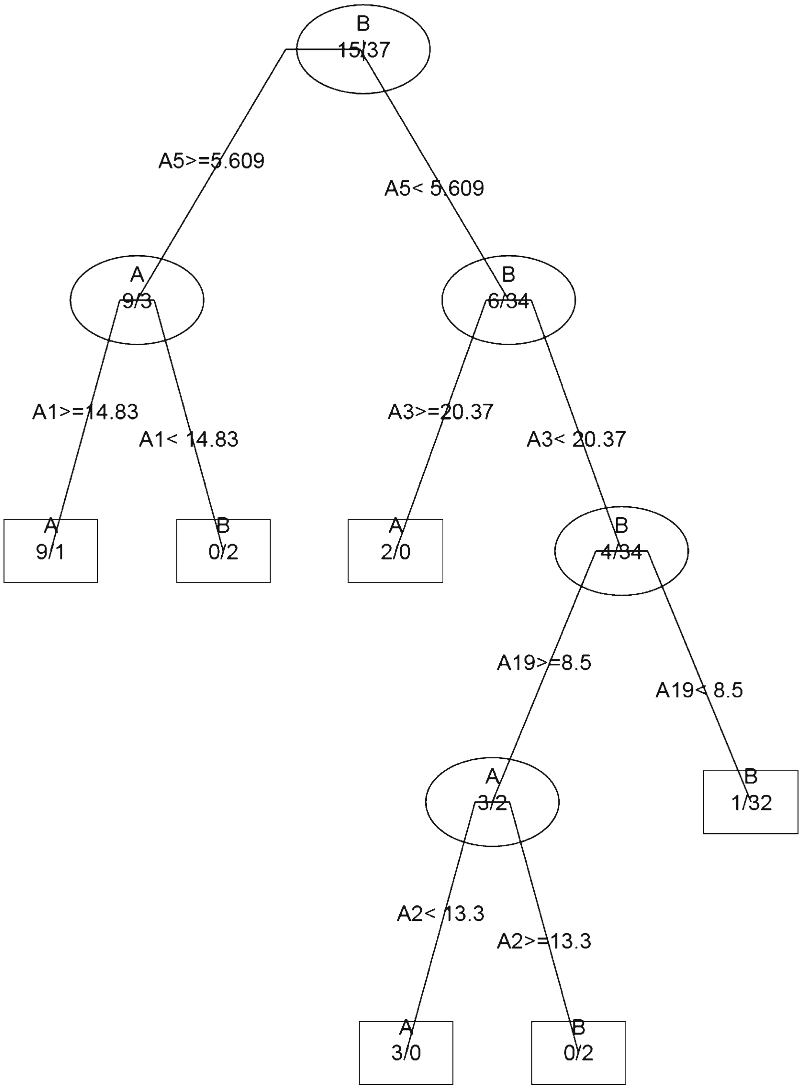


Figure 2. The decision tree for distinguishing active analogue (A) from inactive analogue; (B) A5-connective eccentricity topochemical index, A3-augmented eccentric connectivity topochemical index, A1-molecular connectivity index, A19-alignment independent descriptor T_C_O_7, A2-eccentric adjacency topochemical index.

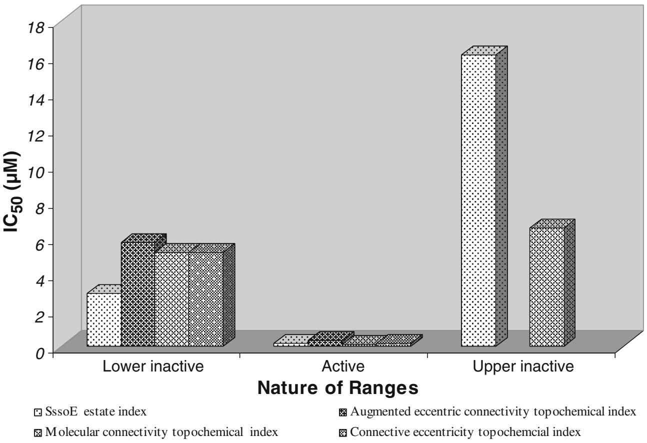


Figure 3. Average IC_{50} (nM) values (based on correctly predicted analogues) of 4-aminomethylene isoquinoline-1, 3-dione derivatives for CDK4 inhibitory activity in various ranges of MAA-based models.

validation of correlation models was carried out using leave-one-out (q^2, LOO) method which described the internal stability of a model.[77] For external validation, the activity of each molecule in test set was predicted using the model developed by training set.[78] The pred_r^2 value is indicative of the predictive power of the current model for external test set. The robustness of the models for training sets was analysed by comparing these models to those derived for random datasets. Random sets were generated rearranging the activities of molecules in the training set. The statistical model was derived using various randomly rearranged activities (random sets) with the selected descriptors and the corresponding values of q^2 were calculated. The significance of the models hence obtained was derived based on a calculated Z score.[79,80] The probability (α) of significant of randomization test is derived by comparing Z score value with Z score critical value as reported.[81] Results are summarized in tables 4, 5 and S3 and figures 4–7.

Table 4. Statistical parameters of MLR, PLS and PCR.

Parameters	MLR	PLS	PCR
N	36	36	36
df	30	32	31
r^2	0.72	0.72	0.72
q^2	0.63	0.60	0.60
F Test	15.59	27.72	20.24
r^2 se	0.51	0.49	0.52
q^2 se	0.61	0.59	0.63
pred_ r^2	0.59	0.53	0.52
pred_ r^2Se	0.86	0.87	0.78
best_ran_ r^2	0.31	0.38	0.34
best_ran_ q^2	0.11	0.21	0.13
Z score_ran_ r^2	7.83	7.41	9.63
Z score_ran_ q^2	4.61	4.57	5.86
α_ ran_ r^2	0.00	0.00	0.00
α_ ran_ q^2	0.001	0.001	0.000
α_ ran_ pred_r^2	0.050	0.100	0.000

MLR = Multiple linear regression, PLS = partial least square, PCR = principal component regression, n = number of molecules of training set, df = degree of freedom, r^2 = coefficient of correlation, q^2 = cross validated r^2, pred r^2 = r^2 for external test set, pred_ r^2Se = coefficient of correlation of predicted data set

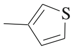
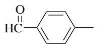
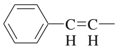
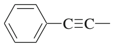
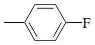
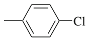
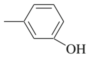
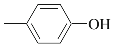
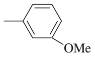
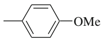
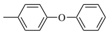
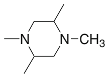
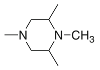
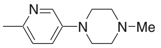
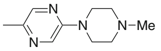
6. Results and discussion

In the present study, decision tree was built from a set of 46 descriptors (table S2).[32–64] The descriptor at root node is most important and the importance of descriptor decreases as the length of tree increases. The classification of 4-aminomethylene isoquinoline-1,3-dione analogues (figure 1) as inactive and active using a single tree, based on connective eccentricity topochemical index A5, augmented eccentric connectivity topochemical index A3, molecular connectivity index A1, alignment independent descriptor T_C_O_7 A19 and eccentric adjacency topochemical index A2 is shown in figure 2. The decision tree identified the connective eccentricity topochemical index A5 as the most important index. The decision tree has classified the analogues with an accuracy of 96%. The specificity and sensitivity of the training set was found to be in the order of 100% and 86.6%, respectively (table 1). In ten-fold cross-validation, 69% of aminomethylene isoquinoline-1, 3-diones analogues were correctly classified with regard to biological activity. The specificity and sensitivity of cross validated set was found to be 72.9% and 60%, respectively (table 1).

The random forests were grown with 46 descriptors enlisted in table S2. The importance of node was determined by mean decrease in accuracy. The RF classified aminomethylene isoquinoline-1,3-diones analogues as inactive and active with an accuracy of 96.4% with respect to CDK4 inhibitory activity. The out-of-bag (OOB) estimate of error was found to be only 8%. The specificity and sensitivity were of the order of 97.2% and 80%, respectively and the value of MCC was found to be 0.7 as given in table 1. High values of MCC simply indicate robustness of the proposed DT and RF based models for CDK4 inhibitory activities.

Using a single descriptor at a time, four independent MAA-based models using E-state contribution index (SssOE), augmented eccentric connectivity topochemical index $\left({}^{A}\xi_{C}^{C}\right)$, molecular connectivity index $\left(\chi^{A}\right)$ and connective eccentricity topochemical index $\left(C^{\xi_{c}}\right)$

Table 5. Reported and predicted activity of 4-aminomethylene isoquinoline-1, 3-dione derivatives used in test set for CDK4 inhibitory activity by MLR.

				$pIC^{\#}_{50}$		
S. No.	Index	R	IC_{50} (μM)*	Reported	Predicted	Residual
1.	13a		12.1	4.9172	6.1262	−1.209
2.	14a		0.48	6.3188	6.1262	0.1926
3.	15a		41	4.3872	4.2038	0.1834
4.	16a		1.62	5.7905	6.3819	−0.5914
5.	18a		0.1	7	5.9013	1.0987
6.	18b		29	4.5376	5.5510	−1.0134
7.	19b		3.3	5.4815	6.1262	−0.6447
8.	20b		3.5	5.4559	5.6047	−0.1488
9.	21b		0.92	6.0362	6.3819	−0.3457
10.	22b		34.7	4.4597	5.9013	−1.4416
11.	23b		0.037	7.4318	6.9340	0.4978
12.	24b	-CN	27.8	4.5560	4.9132	−0.3572
13.	25b		0.32	6.4948	6.1262	0.3686
14.	5c		2.8	5.5528	5.1650	0.3878
15.	5d		0.13	6.8862	4.9292	1.957
16.	6d		1.25	5.9031	6.75	−0.8469

* = Compound concentration in micro mole required to inhibit CDK4 activity by 50%
# = –Log (IC_{50} * 10^{-6})

were developed (table S1). The proposed models have been illustrated in table 2. The overall accuracy of prediction varied from 90.6% for augmented eccentric connectivity topochemical index $\left({}^{A}\xi^{C}_{C}\right)$ to 96.2% for connective eccentricity topochemical index $\left(C^{\xi_C}\right)$. Transitional ranges were observed in all the models indicating a gradual change in CDK4 inhibitory activity. The average IC_{50} (table 2 and figure 3) for active range in all the models varied from 0.15 μM to 0.32 μM. The observation of extremely low average IC_{50} values indicates high potency of the active ranges in the proposed models. Consequently, these models offer vast potential for development of potent CDK4 inhibitors.

Intercorrelation analysis (table 3) revealed that E-state contribution index (SssOE), augmented eccentric connectivity topochemical index $({}^{A}\xi^{C}_{C})$ molecular connectivity index (χ^{A}) and connective eccentricity topochemical index $\left(C^{\xi_C}\right)$ are not correlated with each other.

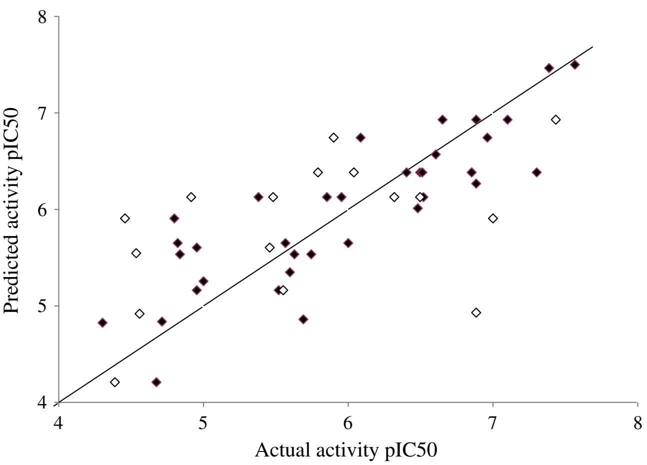


Figure 4. Graph of reported vs. predicted activities for training and test set molecules by multiple linear regression (MLR) model. Training set (solid squares) and test set (hollow squares).

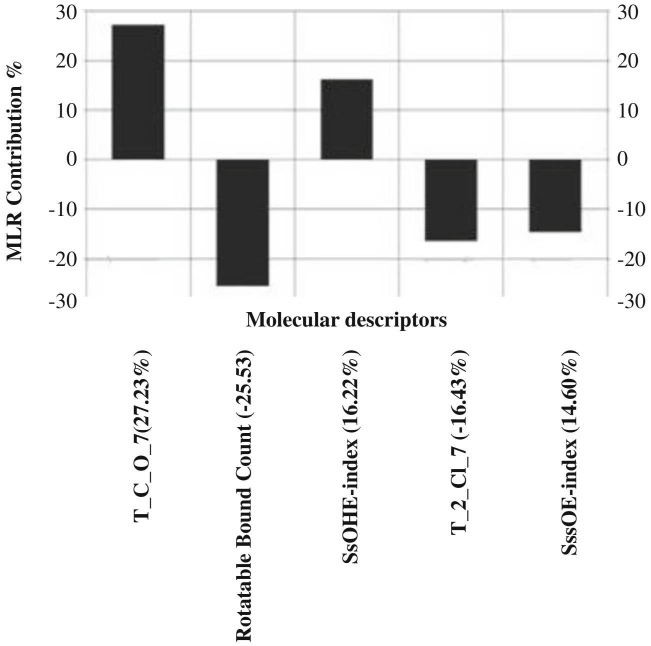


Figure 5. Plot of percentage contribution of each descriptor in developed MLR model explaining variation in the activity.

After QSAR study by MLR using forward–backward step-wise variable selection method, the final equation developed and the statistical data observed are illustrated below.

$$\begin{aligned}\mathrm{pIC}_{50} = {} & 0.1841\ (\mathrm{T\_C\_O\_7}) \\ & - 0.4806\ (\text{rotatable bond count}) \\ & + 0.1247\ (\text{SsOHE-index}) \\ & - 0.7727\ (\mathrm{T\_2\_Cl\_7}) \\ & - 0.1639\ (\text{SssOE-index}) + 7.4965.\end{aligned}$$

The QSAR model had a correlation coefficient (r^2) of 0.72, significant cross validated correlation coefficient (q^2) of 0.63, F test of 15.59, r^2 for external test set (pred_r^2) 0.59, and degree of freedom 30. The model developed predicts 63% of variance and is validated by an external set of compounds with a predictive correlation coefficient of 0.59. The model is validated by α_ran_$r^2 = 0.00$, α_ran_$q^2 = 0.001$, α_ran_pred_$r^2 = 0.05$, best_ran_$r^2 = 0.31$, best_ran_$q^2 = 0.11$, Z score_ran_$r^2 = 7.83$, Z score_ran_$q^2 = 4.61$ (table 4). The randomization test suggests that the developed model has a probability of less than 1% and that the model is generated by chance. The predictability of model was evaluated by test set of compounds.

The reported and predicted pIC_{50} for training set along with residual values are presented in table S3. The predictive ability of model evaluated using test set is presented in table 5. The plot of reported vs predicted activity and contribution of descriptors for the CDK4 inhibitory activity is shown in figures 4 and 5, respectively. The major group of contributing descriptors involved subgroups like rotatable bond count, SsOHE-index, SssOE-index and alignment independent descriptors. These descriptors help in understanding the effect of substituent at different position of 4-aminomethylene isoquinoline-1,3-(2*H*,4*H*)-dione.

The direct relationship of descriptor T_C_O_7 (27.23%) suggested that the presence of oxo-group at position one in the basic ring 4-aminomethylene isoquinoline-1,3-(2*H*,4*H*)-dione should necessarily be separated from the substituent at position six by seven bonds.

The presence of T_2_Cl_7 (having negative MLR coefficient −16.43%) in the model revealed that the

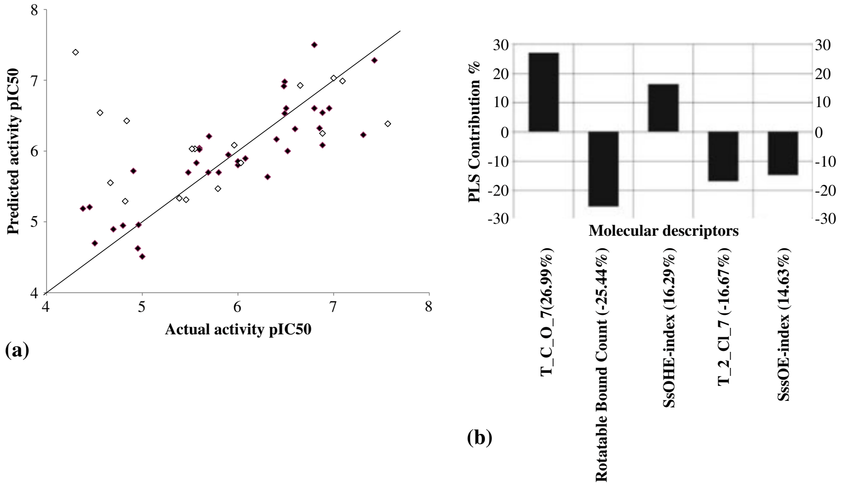


Figure 6. (**a**) Graph of reported vs. predicted activities for training and test set molecules by partial least square (PLS) regression model. Training set (solid squares) and test set (hollow squares). (**b**) Plot of percentage contribution of each descriptor in developed PLS model explaining variation in the activity.

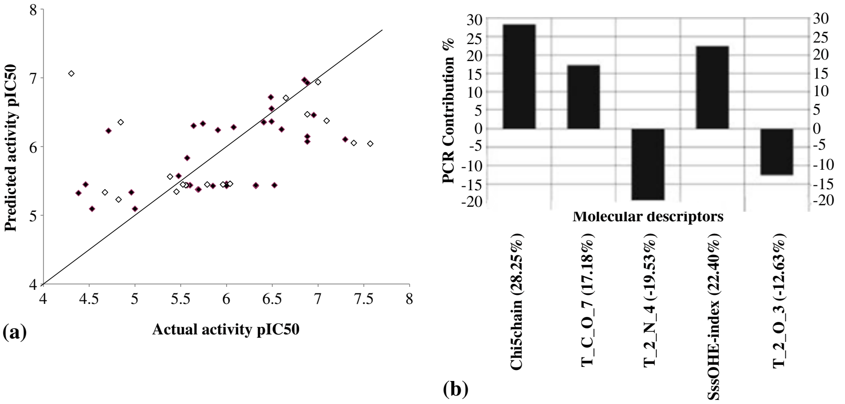


Figure 7. (**a**) Graph of reported vs. predicted activities for training and test set molecules by principal component (PCR) regression model. Training set (solid squares) and test set (hollow squares). (**b**) Plot of percentage contribution of each descriptor in developed PCR model explaining variation in the activity.

presence of halides at position seven i.e., chloro at position seven separated by seven bonds from doubly bonded oxo at position three have negative effect on the activity.

The directly related descriptor SsOHE-index (16.22%) indicated that the presence of hydroxyl group in six phenyl substitued basic ring have positive effect on the activity. SssOE-index (−14.6%) is an estate contribution descriptor which represent electrotopological estate indices for the number of oxygen group connect with two single bonds. This term is negatively correlated and indicated that compound with higher SssOE-index values show less activity and vice-versa.

The presence of descriptor rotatable bond count (having negative MLR coefficient −25.53%) signifies that the unsaturated bonds, and single bonds connected to hydrogen or terminal atoms are favourable for the biological activity.

After QSAR study by PLS using forward–backward step-wise variable selection method, the final equation developed and the statistical data observed are illustrated below.

$$\begin{aligned}\mathrm{pIC}_{50} = {} & 0.1827\ (\mathrm{T\_C\_O\_7}) \\ & - 0.4797\ (\text{rotatable bond count}) \\ & + 0.1253\ (\text{SsOHE-index}) \\ & - 0.7850\ (\mathrm{T\_2\_Cl\_7}) \\ & - 0.1645\ (\text{SssOE-index}) + 7.5007.\end{aligned}$$

The QSAR model had a correlation coefficient (r^2) of 0.72, significant cross validated correlation coefficient (q^2) of 0.6, F test of 27.72, r^2 for external test set (pred_r^2) 0.53, and degree of freedom 32. The model was validated by α_ran_r^2 = 0.00, α_ran_q^2 = 0.00, α_ran_pred_r^2 = 0.01, best_ran_r^2 = 0.38, best_ran_q^2 = 0.21, Z score_ran_r^2 = 7.41, Z score_ran_q^2 = 4.57 (table 4).

The plot of reported vs predicted activity and contribution of descriptors for the CDK4 inhibitory activity is shown in tables S4–S5 and figure 6. The major groups of descriptors involved in developing the equation by PLSR are subgroups like rotatable bond count, SsOHE-index, SssOE-index and alignment independent descriptors. The descriptors are common between MLR and PLSR. These only differ from each other in their percentage of contribution.

After QSAR study by PCR using forward–backward step-wise variable selection method, the final equation developed and the statistical data observed are illustrated below.

$$\begin{aligned}\mathrm{pIC}_{50} = {} & 7.2205\ (\mathrm{chi5chain}) \\ & + 0.1097\ (\mathrm{T\_C\_O\_7}) \\ & - 0.3324\ (\mathrm{T\_2\_N\_4}) \\ & + 0.1682\ (\text{SsOHE-index}) \\ & - 0.2061\ (\mathrm{T\_2\_O\_3}) + 9.2238.\end{aligned}$$

The QSAR model had a correlation coefficient (r^2) of 0.72, significant cross validated correlation coefficient (q^2) of 0.6, F test of 20.24, r^2 for external test set (pred_r^2) 0.52, and degree of freedom 30. The model is validated by α_ran_r^2 = 0.00, α_ran_q^2 = 0.00, α_ran_pred_r^2 = 0.001, best_ran_r^2 = 0.34, best_ran_q^2 = 0.2 (table 4).

The plot of reported vs predicted activity and contribution of descriptors for the CDK4 inhibitory activity is shown in tables S6–S7 and figure 7.

The major groups of descriptors involved in developing the equation by PCR are subgroups like chi5chain, SsOHE-index and alignment independent descriptors.

The QSAR model by PCR reveals that the descriptors SsOHE-index and T_C_O_7 are common in MLR and PCR. These only differ from each other in their percentage of contribution. The other contributing descriptors are the chi5chain, T_2_N_4 and T_2_O_3.

The chi5chain (28.25%) is directly proportional to the biological activity. The descriptor T_2_N_4 (−19.53) is negatively correlated with activity shows that increasing the distance between 4-amino phenylmethylene from oxo by increasing number of carbon atoms in the basic ring have negative effect on the activity. The descriptor T_2_O_3 (−12.63) which is also negatively correlated with activity shows that the increase and decrease of distance between two oxo-group have negative effect on the activity.

This study has helped to understand the molecular properties/features that play an important role in governing the variation in the activities. In addition, this study allows us investigate the influence of very simple and easy-to-compute descriptors in determining biological activities, which could shed light on the key factors that may aid in design of potent molecules.

Combined approaches using molecular properties and well selected MDs are not only likely to produce superior correlations but are expected to do so in a most efficient way. Structure–activity studies are highly complex and various methodologies, even if addressing limited aspects of the QSAR problem, ought to be exhaustively explored and amalgamated if possible.

7. Conclusion

In the present study, both classification and correlation approaches have been successfully employed for development of models for prediction of CDK4 inhibitory activity of 4-aminomethylene isoquinoline-1,3-(2*H*,4*H*)-dione. The accuracy of classification of single descriptor-based models using MAA varied from 90% to 96%. High accuracy of prediction offers the proposed models a vast potential for providing lead structures for the development of potent therapeutic agents for CDK4 inhibition.

Supporting information

The electronic supporting information (tables S1–S7) can be seen at www.ias.ac.in/chemsci.

References

1. Sherr C J 2000 *Cancer Res.* **60** 3689
2. Yeudall W A and Jakus J 1995 *Eur. J. Cancer* **31** 291
3. Pavletich N P 1999 *J. Mol. Biol.* **287** 821
4. Malumbres M and Barbacid M 2005 *Trends Biochem. Sci.* **30** 630
5. Hengstschläger M, Braun K, Soucek T, Miloloza A and Hengstschläger-Ottnad E 1999 *Mutat. Res.* **436** 1
6. Sridhar J, Akula N and Pattabiraman N 2006 *AAPS J.* **8** E204
7. Zieske J D 2000 *Prog. Retinal Eye Res.* **19** 257
8. Morgan D O 1997 *Annu. Rev. Cell Dev. Biol.* **13** 261
9. Lee W H, Bookstein R, Hong F, Young L J, Shew J Y and Lee E Y 1987 *Science* **235** 1394
10. Julie P I, Welburn I and Jane A 2005 *Semin. Cell & Dev. Biol.* **16** 369
11. Dai Y and Grant S 2003 *Curr. Opin. Pharmacol.* **3** 362
12. Shapiro G I 2006 *J. Clin. Oncol.* **24** 1770
13. Thomas H C, Dunlop M G and Stark L A 2007 *Cell Cycle* **6** 1293
14. Graf F, Koehler L, Kniess T, Wuest F, Mosch B and Pietzsch J 2009 *J. Oncol.* **106378** 1
15. Honma T, Yoshizumi T, Hashimoto N, Hayashi K, Kawanishi N, Fukasawa K, Takaki T, Ikeura C, Ikuta M, Suzuki-Takahashi I, Hayama S, Nishimura S and Morishima H 2001 *J. Med. Chem.* **44** 4628
16. Toogood P L, Harvey P J, Repine J T, Sheehan D J, VanderWel S N, Zhou H, Keller P R, McNamara D J, Sherry D, Zhu T, Brodfuehrer J, Choi C, Barvian M R and Fry D W 2005 *J. Med. Chem.* **48** 2388
17. Zhu G, Conner S E, Zhou X, Shih C, Li T, Anderson B D, Brooks H B, Campbell R M, Considine E, Dempsey J A, Faul M M, Ogg C, Patel B, Schultz R M, Spencer C D, Teicher B and Watkins S A 2003 *J. Med. Chem.* **46** 2027
18. Engler T A, Furness K, Malhotra S, Sanchez-Martinez C, Shih C, Xie W, Zhu G, Zhou X, Conner S, Faul M M, Sullivan K A, Kolis S P, Brooks H B, Patel B, Schultz R M, DeHahn T B, Kirmani K, Spencer C D, Watkins S A, Considine E L, Dempsey J A, Ogg C A, Stamm N B, Anderson B D, Campbell R M, Vasudevan V and Lytle M L 2003 *Biorg. Med. Chem. Lett.* **13** 2261
19. Sanchez-Martinez C, Shih C, Faul M M, Zhu G, Paal M, Somoza C, Li T, Kumrich C A, Winneroski L L, Xun Z, Brooks H B, Patel B K R, Schultz R M, DeHahn T B, Spencer C D, Watkins S A, Considine E, Dempsey J A, Ogg C A, Campbell R M, Anderson B A and Wagner J 2003 *Bioorg. Med. Chem. Lett.* **13** 3835
20. Sharma P S, Sharma R and Tyagi R 2008 *Curr. Cancer Drug Targets* **8** 53
21. Grant S and Roberts J D 2003 *Drug Resist. Update* **6** 15
22. Chu X J, DePinto E, Bartkovitz D, So S S, Vu B T, Packman K, Lukacs C, Ding Q, Jiang N, Wang K, Goelzer P, Yin X, Smith M A, Higgins B X, Chen Y, Xiang Q, Moliterni J, Kaplan G, Graves B, Lovey A and Fotouhi N 2006 *J. Med. Chem.* **49** 6549
23. Wyatt P G, Woodhead A J, Berdini V, Boulstridge J A, Carr M G, Cross D M, Davis D J, Devine L A, Early T R, Feltell R E, Lewis E J, McMenamin R L, Navarro E F, O'Brien M A, O'Reilly M, Reule M, Saxty G, Seavers L C A, Smith D M, Squires M S, Trewartha G, Walker M T and Woolford A J A 2008 *J. Med. Chem.* **51** 4986
24. McInnes C 2008 *Drug Discov. Today* **139** 875
25. Lapenna S and Giordano A 2009 *Nat. Rev. Drug Discov.* **8** 547

26. Vilar S, Estrada E, Uriate E, Santana L and Gutirrez Y 2005 *J. Chem. Inf. Model* **45** 502
27. Li H, Zheng M, Luo X, Zhu W and Jiang H 2008 In *Wiley encyclopedia of chemical biology* (Hoboken, NJ, USA: John Wiley & Sons, Inc.) p. 1. doi: 10.1002/9780470048672.wecb098
28. Katritzky A R and Gordeeva E V 1993 *J. Chem. Inf. Comut. Sci.* **33** 835
29. Devillers A and Balaban A T 1999 In *QSAR and QSPR* (The Netherlands: Gordon and Breach Science Publishers) p. 563
30. Randic M 1991 *Chemom. Intell. Lab. Syst.* **10** 213
31. Tsou H, Otteng M, Tran T, Foyd M B, Reich M, Birnberg G, Kutterer K, Ayral-Kaloustian S, Ravi M, Nilakantan R, Grillo M, McGinnis J P and Rabindran S K 2008 *J. Med. Chem.* **51** 3507
32. Goel A and Madan A K 1995 *J. Chem. Inf. Comput. Sci.* **35** 510. doi: 10.1021/ci00025a019
33. Dureja H and Madan A K 2005 *J. Mol. Mod.* **11** 525. doi: 10.1007/s00894-005-0276-3
34. Gupta S, Singh M and Madan A K 2003 *Indian J. Chem.* **42A** 1414
35. Madan A K and Dureja H 2010 In *Novel molecular structure descriptors* (eds) I Gutman and B Furtula (Serbia). University of Kragujevac and Faculty of Science Kragujevac, p. 91
36. Kumar V, Sardana S and Madan A K 2004 *J. Mol. Mod.* **10** 399. doi: 10.1007/s00894-004-0215-8
37. Bajaj S, Sambhi S S and Madan A K 2005 *Croat. Chem. Acta* **78** 165
38. Bajaj S, Sambhi S S and Madan A K 2004 *J. Mol. Struct. (Theochem)* **684** 197. doi: 10.1016/j.theochem.2004.01.052
39. Randic M 1975 *J. Am. Chem. Soc.* **97** 6609. doi: 10.1021/ja00856a001
40. Gupta S, Singh M and Madan A K 2001 *J. Comput. Aided Mol. Des.* **15** 671. doi: 10.1023/A:1011964003474
41. Bajaj S, Sambhi S S and Madan A K 2006 *QSAR Comb. Sci.* **25** 813. doi: 10.1002/qsar.200430918
42. Sharma V, Goswami R and Madan A K 1997 *J. Chem. Inf. Comput. Sci.* **37** 273. doi: 10.1021/ci960049h
43. Gupta S, Singh M and Madan A K 2008 *J. Mol. Graph Mod.* **18** 18. doi: 10.1016/S1093-3263(00)00027-9
44. Gutman I, Ruscic B, Trinajstic N and Wicox C F 1975 *J. Chem. Phys.* **62** 3399. doi: 10.1063/1.430994
45. Gutman I and Randic M 1977 *Chem*, *Chem. Phys. Lett.* **47** 15. doi: 10.1016/0009-2614(77)85296-2
46. Wiener H 1947 *J. Am. Chem. Soc.* **69** 2636. doi: 10.1021/ja01203a022
47. Wiener H 1947 *J. Chem. Phys.* **15** 766. doi: 10.1063/1.1746328
48. Balaban A T 1983 *Pure Appl. Chem.* **55** 199. doi: 10.1351/pac198855020199
49. Gupta S, Singh M and Madan A K 1999 *J. Chem. Inf. Comput. Sci.* **39** 272. doi: 10.1021/ci980073q
50. Todeschini R and Consonni V 2009 *Molecular descriptors for cheminformatics* Vol. 1 *Alphabetical listing* 2nd revised edition (Wienhein: Wiley-VCH Verlag GmbH & Co.) p. 1249
51. Wildman S A and Crippen G M 1999 *J. Chem. Inf. Comput. Sci.* **39** 868
52. Baumann K 2002 *J. Chem. Inf. Comput. Sci.* **42** 26
53. Hall L H, Moheny B K and Kier L B 1991 *J. Chem. Inf. Comput. Sci.* **31** 76
54. Hall L H and Kier L B 1995 *J. Chem. Inf. Comput. Sci.* **35** 502
55. Hall L H and Kier L B 1991 The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling In *Reviews of computational chemistry* Vol. 2 (eds) D B Boyd and K Lipkowitz (Wienhein: VCH Verlag GmbH & Co.) p. 367
56. Kier L B 1985 *Quant. Struct. Act. Relat.* **4** 109
57. Kier L B 1986 *Quant. Struct. Act. Relat.* **5** 7
58. Kier L B 1986 *Acta Pharm. Jugosl.* **36** 171
59. Miller K J 1990 *J. Am. Chem. Soc.* **112** 8533
60. Wang R, Fu Y and Lai L 1997 *J. Chem. Inf. Comput. Sci.* **37** 615
61. Palm K, Luthman K, Ungell A-L, Strandlund G, Beigi F, Lundahl P and Artursson P 1998 *J. Med. Chem.* **41** 5382
62. Winiwarter S, Bonham N M, Ax F, Hallberg A, Lennernas H and KarlCn A 1998 *J. Med. Chem.* **41** 4939
63. Dureja H and Madan A K 2007 *Med. Chem. Res.* **16** 331
64. Gupta M, Gupta S, Dureja H and Madan A K 2012 *Chem. Biol. Drug Design* **79(1)** 38. doi: 10.1111/j.1747-0285.2011.01264.x
65. Bajaj S, Sambhi and Madan A K 2004 *QSAR Comb. Sci.* **23(7)** 506. doi: 10.1002/qsar.200439999A
66. Kim H and Koehler G J 1995 *Omega Int. J. Mgmt. Sci.* **23** 637
67. Myles A J, Feudale R N, Liu Y, Woody N A and Brown S D 2004 *J. Chemomet.* **18** 275. doi: 10.1002/cem.873
68. Breiman L 2001 *Machine Learning* **45(1)** 5
69. Dureja H, Gupta S and Madan A K 2008 *Sci. Pharm.* **76** 377. doi: 10.3797/scipharm.0803-30
70. Berk R A 2003 *Regression analysis: A constructive critique* (London: SAGE Publications Ltd.) p. 103
71. Schultz T W, Cronin M T D, Walker J D and Aptula A O 2003 *J. Mol. Struct.* **622** 1
72. Wold S and Dunn W J 1998 *J. Chem. Inf. Comput. Sci.* **23** 6
73. Dunteman G H 1989 *Principal component analysis* (ed.) D H Dunteman (London: SAGE Publications Ltd.) pp. 15–22
74. Han L, Wang Y and Bryant S H 2008 *BMC Bioinformatics* **9** 401. doi: 10.1186/1471-2105-9-401
75. Nikolic S, Kavacevic G, Milicevic A and Trinanjstic N 2003 *Croat. Chem. Acta* **76** 113
76. Trinajsti N, Nikolic S, Basak S C and Lukovits I 2001 *SAR QSAR Environ. Res.* **12** 31
77. Zheng W and Tropsha A 2000 *J. Chem. Inf. Comput. Sci.* **40** 185
78. Gedeck P, Rohde B and Bartels C 2006 *J. Chem. Inf. Comput. Sci.* **46** 1924
79. Gilbert N 1976 *Statistics* (Philadelphia PA: W B Saunders)
80. Golbraikh A and Tropsha A 2003 *J. Chem. Inf. Comput. Sci.* **43** 144
81. Shen M, Xiao Y, Golbraikh A, Gombar V K and Tropsha A 2003 *J. Med. Chem.* **46** 3013