Article

Journal of Translational Toxicology
Vol. 1, 21–27, 2014
www.aspbs.com/jtt

AMERICAN SCIENTIFIC PUBLISHERS
Copyright © 2014 American Scientific Publishers
All rights reserved
Printed in the United States of America

ToxiPred: A Server for Prediction of Aqueous Toxicity of Small Chemical Molecules in *T. Pyriformis*

Nitish Kumar Mishra[1], Deepak Singla[1], Sandhya Agarwal[1], Open Source Drug Discovery Consortium[2], and Gajendra P. S. Raghava[1,*]

[1] *Bioinformatics Centre, Institute of Microbial Technology (CSIR), Chandigarh 160036, India*
[2] *The Open Source Drug Discovery (OSDD) Consortium, Council of Scientific and Industrial Research, Anusandhan Bhavan, 2 Rafi Marg, Delhi 110001, India*

Background: Toxicity Prediction is one of the crucial issues as various industrial chemicals are linked with acute and chronic human diseases like carcinogenicity, mutagenecity. Thus, there is a growing need to risk assessment of these chemicals. *Tetrahymena pyriformis* is used as a model organism to accessed the environmental fate of a chemical to address the toxicity potential of organic chemicals. Our study is based on large diverse dataset of 1208 compounds taken from an international open competition ICANN09 was organized for aqueous toxicity prediction of chemical molecules against *Tetrahymena pyriformis*. *Results*: This study described the development of Quantitative Structure Toxicity Relationship (QSTR) model for the prediction of aqueous toxicity against *T. pyriformis*. Firstly, model developed on 1002 V-life calculated molecular descriptors shows a R/R^2 0.874/0.76 with RMSE 0.523. Further, selection of relevant descriptors leads to only 9 descriptors, which shows a performance R/R^2 0.846/0.71 with RMSE 0.574 while on blind dataset 0.756/0.570 with RMSE 0.570 respectively. Second, model developed on CDK based 178 descriptors shows correlation (R) 0.876/0.85, R^2 0.77/0.72 with RMSE 0.518/0.556 on training and blind dataset respectively. Next, model developed on selected 6 descriptors from CDK shows nearly equal performance with R 0.866/0.823, R^2 0.75/0.66 with RMSE 0.541/0.609 on training and blind dataset respectively. Finally, a hybrid model based on selected 17 descriptors from both V-life and CDK shows significant improvement in performance on both training and blind dataset with R 0.89/0.85, R^2 0.79/0.72 with RMSE 0.491/0.557 respectively. It was also observed that Molecular mass (M.W.), and XLogP have very high correlation with toxicity of chemical molecules, it suggests that size and solubility of chemical molecules play major role in toxicity. Our results suggest that it is possible to develop web service for computing toxicity of chemicals using non-commercial software. *Conclusions*: Our present study demonstrates that performance of a QSTR model depends on the quality/quantity of descriptors as well as on used techniques. Based on these observations, we developed a web server ToxiPred (http://crdd.osdd.net/raghava/toxipred), for environmental risk assessment of small chemical compounds.

KEYWORDS: *Toxicity, QSAR/QSATR, Prediction, Machine Learning Techniques, ToxiPred, SVM.*

INTRODUCTION

Toxicity assessment for a given compound using toxicological experiment is a mammoth task due to cost and time.[1,2] A number of biological interactions and different environment with the living organisms are responsible for accurate determination of toxicity, but data that quite often are not available.[3] A generally accepted strategy for overcoming the shortage of experimental measurements is the analysis based on Quantitative Structure–Activity Relationships (QSAR).[4,5] Computational tools fasten the environmental assessment process by significantly reduce the cost of experimental.

In past several QSAR/QSTR based models have been developed for toxicity prediction.[6–14] In year 2011, also a method for classification of toxic and non-toxic compounds was developed on a large diverse dataset.[15] The most critical limitation of existing QSAR studies is there low performance on blind/independent dataset despite its high accuracy on training data set. This is due to over optimization or over training on dataset used for training. Therefore, it is important to use

*Author to whom correspondence should be addressed.
Email: raghava@imtech.res.in
Received: 14 July 2012
Revised/Accepted: 22 September 2012

standard cross-validation criteria for testing, training and independent validation. Thus, there is a need to develop fast and robust *in-silico* model for predicting toxicity of chemicals. In order to address this problem, the highly diverse and large dataset taken from International Conference on Artificial Neural Networks (ICANN09) benchmarking the performance of toxicity prediction methods (http://www.cadaster.eu/node/65) was used.

The purpose of this study is, firstly to develop a highly robust and accurate QSAR model for the prediction of aqueous toxicity of small chemical molecules, and secondly to develop software/server for public use. For this purpose, we have used CDK[16] and Vlife software's for calculating descriptors of chemicals and Weka/RapidMiner for feature selection. This study will be useful for experimental biologist for predicting the toxicity of a compound against *T. pyriformis.*

MATERIAL AND METHODS

Dataset

In this study, total 1213 molecules have been used for developing QSAR models for aqueous toxicity prediction, obtained from http://www.cadaster.eu/node/67 (ICANN09 competition web site). MOPAC based optimized 3D structure of these molecules are available in mol2 format along with pIGC50 values (logarithm of 50% growth inhibitory concentration). There was an error in 5 molecules for descriptors calculations therefore the overall study is based on 1208 compounds rather than 1213 compounds. The dataset was divided randomly into training and blind set with 1108 and 100 molecules respectively. Further, In order to explore the diversity of toxicity dataset, their toxicity value and four molecular descriptors molecular weight (Mol. wt.), XlogP, nHBA (no. of hydrogen bond acceptors), nHBD (no. of hydrogen bond donors) calculated using CDK libraries for each compounds were analyzed by radar chart (Fig. 1).

Molecular Descriptors

In this study, QSTR models were developed using descriptors computed from two software's namely CDK, and Vlife. Vlife allows computing ~1002 type of molecular descriptors that can be broadly categorized into structural, thermodynamic, electronic, molecular surface based descriptors. We also computed 178 molecular descriptors using Chemistry Development Kit (CDK) library that includes topological descriptor, geometric descriptor, molecular properties and Eigen values based indices, physiochemical and electronic descriptors. The CDK is a Java based open source library for structural chemo- and bioinformatics projects. Our group had earlier implemented this library in the form of a standalone as well as a web server WebCDK, which is available at http://crdd.osdd.net:8081/WebCDK/.

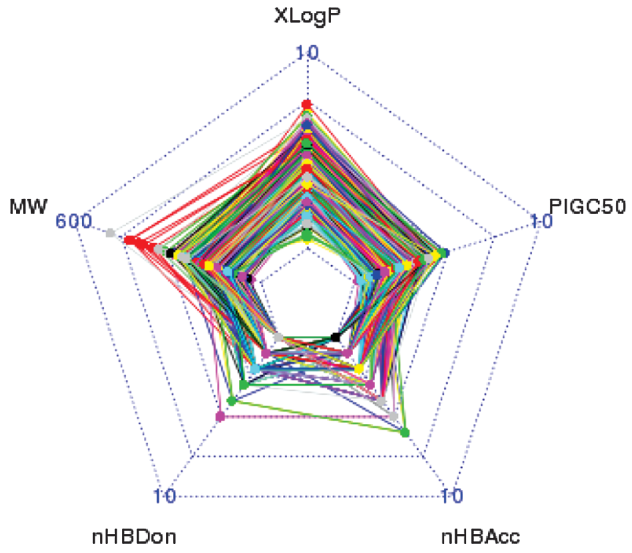

Figure 1. Depict the radar chart of four simple molecular descriptors: molecular weight (MW), nHBDon (no. of hydrogen bond donars). nHBAcc (no. of hydrogen bond acceptors), XLogP and toxicity (pIGC50) on entire dataset with each color line represent a compound.

Descriptor Selection

In a QSAR study, selection of relevant molecular descriptors from descriptor space is most important and tricky step to build an efficient predictable model. Most of molecular descriptor programs compute a large set of molecular descriptors that include highly correlated and irrelevant descriptor. There are number of techniques which allow selecting appropriate descriptors. Among them only a small subset of descriptors are statistically significant for QSAR based model development. To search a set of significant descriptors, we adapted different feature selection methods such as CfsubsetEval module with best-fit algorithm, F-stepping approach, and removal of highly correlated descriptors.

Cross-Validation Techniques

The performance of the QSAR model was evaluated using five-fold cross-validation technique. In five-fold CV, the data set is randomly divided in five partitions of similar size. Out of these five sets four sets are used for training and the remaining fifth set for testing. The model was rebuilt five times, once for each fold ensuring that all compounds were used for testing once. In order to check the general ability of model, an independent test set is most commonly used. Therefore, in this study we have also used an independent dataset of 100 compounds to evaluate the performance of our models.

QSAR Model Construction

QSAR methodology quantitatively correlates the structural molecular properties (descriptors) with functions

Delivered by Publishing Technology to: Guest
IP: 14.139.224.82 On: Sun, 11 May 2014 11:13:34
Copyright: American Scientific Publishers

(biological activities) for a set of compounds by means of linear or non-linear statistical methods.[17–20] In the present study, we have used both linear (MLR) and non-linear (SMO) statistical methods for prediction of aqueous toxicity of small chemical molecules in *T. pyriformis*. Brief description of these methods is given below.

Non-Linear Statistical Method

WEKA-3.6.0 Based Method

The machine-learning package WEKA 3.6.0[22] is a collection of machine-learning algorithms, which supports in several standard data mining tasks, data preprocessing, clustering, classification, regression, visualization, and feature selection. Here, we used SMOreg (Sequential Minimization Optimization)[23] algorithms implemented in WEKA to predict the end point toxicity.

Linear Statistical Method

Multiple Linear Regression (MLR)

MLR tries to model the relationship between two or more independent descriptors and dependent variable such as y, by fitting a linear regression equation to observed data with corresponding parameters (constants) and an error term. In MLR, every value of the independent variable is associated with a value of the dependent variable. The multiple linear relations between y and the $\{x_p\}$ is defined by following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon(x)$$

Where y is a dependent variable, $\{x_1, x_{2\ldots\ldots}x_p\}$ are the independent variables, $\{\beta_1, \beta_{2\ldots\ldots}\beta_p)$ is the slop (beta coefficient) for particular independent variable and $\varepsilon(x)$ is a random noise (e.g., measurement errors). In current study, MLR equation has obtained through R package, was used for QSAR modeling.

Evaluation Parameter

Once a regression model was constructed, statistical significance of models was assessed using the following statistical parameters:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ToX^{\mathrm{act}} - ToX^{\mathrm{pred}})^2}$$

$$R = \frac{\left(n\sum TOX^{\mathrm{act}}ToX^{\mathrm{pred}} - \sum TOX^{\mathrm{act}}\sum ToX^{\mathrm{pred}}\right)}{\times\left(\sqrt{n\sum(TOX^{\mathrm{act}})^2 - (\sum TOX^{\mathrm{act}})^2}\right.}$$
$$\left.\sqrt{n\sum(ToX^{\mathrm{pred}})^2 - (\sum TOX^{\mathrm{act}})^2}\right)^{1/2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(ToX^{\mathrm{act}} - ToX^{\mathrm{pred}})^2}{\sum_{i=1}^{n}(ToX^{\mathrm{act}} - \overline{\mathrm{ToX}})}$$

Where n is the size of test set, $\mathrm{Tox}^{\mathrm{pred}}$ is the predicted pIGC50 and $\mathrm{Tox}^{\mathrm{act}}$ is the actual pIGC50, is the average of the toxicity of test set, RMSE is the root mean squared error between actual and predicted pIGC50 of compounds, R is the Pearson's correlation coefficient between actual and predicted value, R^2 (Coefficient of determination) is the statistical parameter for proportion of variability in model.

Delivered by Publishing Technology to: Guest User
IP: 14.139.224.5 On: Sun, 11 May 2014 11:12:34
Copyright: American Scientific Publishers

RESULTS

Diversity Analysis of Data Set

The diversity in the dataset is very important for robust QSTR model development. In order to explore the chemical domain of total dataset, a radar chart analysis was performed on total 1209 compounds. The radar chart shows that Mol. wt. Varies from 32.03 to 483.59; XlogP from 2.41 to 6.5; no. of hydrogen bond donar range from 0 to 5, no. of hydrogen bond acceptor ranged from 0 to 6; and toxicity value from 2.67 to 3.34. As shown Figure 1,

Table I. Performance of models developed on different set of descriptors calculated using various software packages.

			Train			Blind		
Software packages	Descriptors	Methods	R	R^2	RMSE	R	R^2	RMSE
CDK	**178**	**PLS**	**0.876**	**0.77**	**0.518**	**0.850**	**0.72**	**0.556**
	11	SMOPuk	0.867	0.75	0.537	0.816	0.65	0.620
	11	MLR	0.853	0.73	0.559	0.799	0.64	0.633
	6	SMOPuk	0.866	0.75	0.541	0.823	0.66	0.609
	6	MLR	0.851	0.72	0.563	0.802	0.64	0.628
V-Life	**1002**	**PLS**	**0.874**	**0.760**	**0.523**	**0.867**	**0.750**	**0.530**
	20	SMOPuk	0.849	0.720	0.570	0.781	0.6	0.665
	9	SMOPuk	0.846	0.71	0.574	0.756	0.57	0.693
	9	MLR	0.831	0.69	0.596	0.785	0.61	0.655
Hybrid-1	31	SMOPuk	0.88	0.77	0.516	0.83	0.69	0.586
	17	SMOPuk	0.89	0.79	0.491	0.851	0.72	0.557
	17	MLR	0.867	0.75	0.534	0.826	0.68	0.594

the entire dataset highly diverse and cover large chemical space.[24] The radar chart are highly similar to Cheng et al. 2011 showing the applicability of dataset.[15]

Performance of Linear Statistical Method (MLR) and Non-Linear Statistical Method (SMO)

In order to evaluate the performance of different software's, we have developed QSAR model on different sets of descriptors calculated by various software's (open source as well as commercial). Performance of various models, trained and blind dataset is described below.

V-Life Descriptors Based Model

In order to develop model, we removed the molecules from V-life for which CDK was unable to calculate the descriptors for making the data composition uniform throughout the study. In this study a PLS model was developed using V-life calculated 1002 descriptors

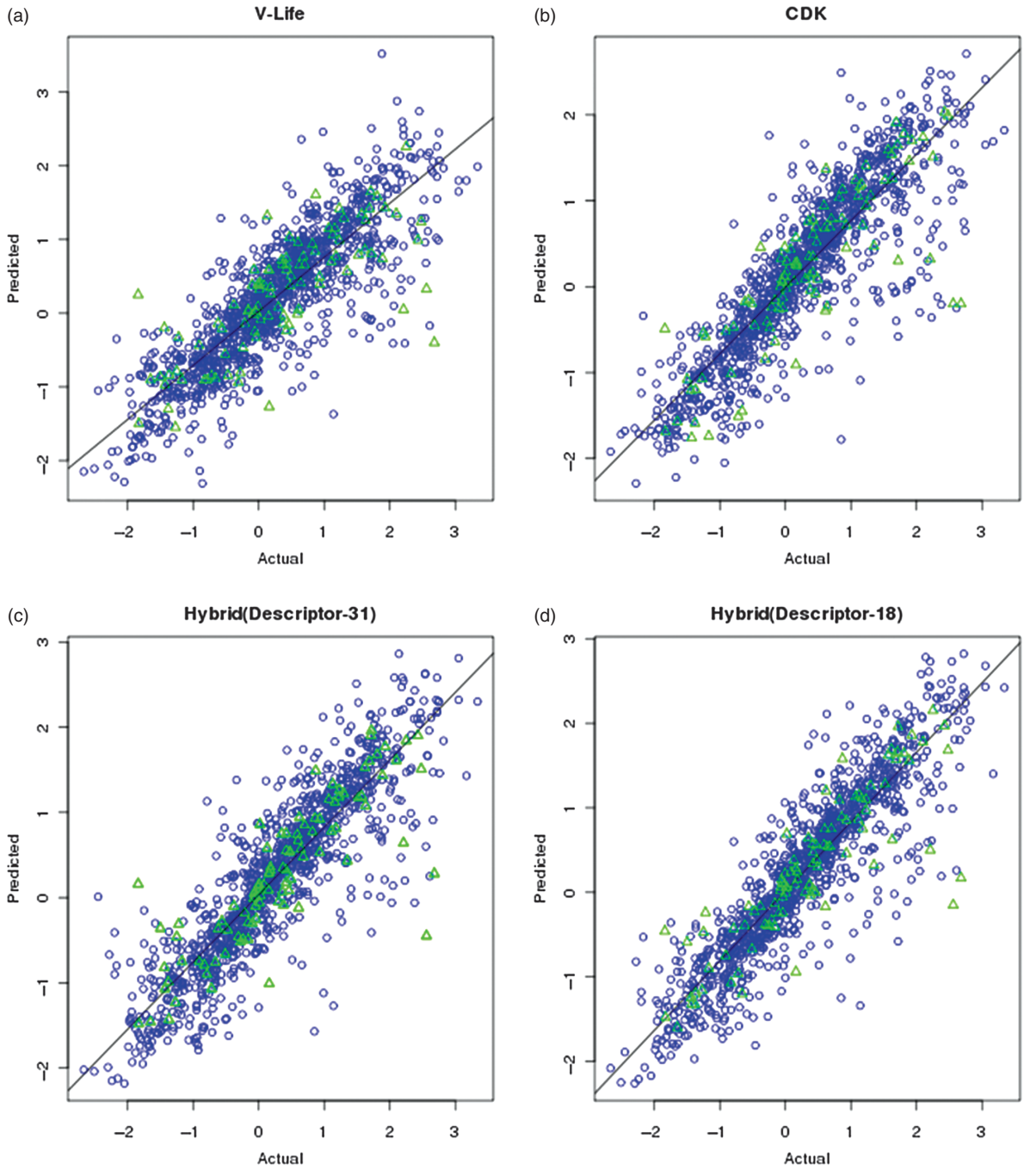

Figure 2. **Showing the scatter plots of four different models with actual value on *X*-axis and predicted value on *Y*-axis. The blue color circle represents the training set and green color triangle for blind dataset.**

Table II. Correlation between selected Vlife descriptors calculated using Vlife software.

Descriptors	Mol.Wt	slogp	chiV5chain	SdSE-index	Ipc	MMFF_21	MMFF_46	MMFF_51	T_2_N_0	T_N_Br_3	T_N_Br_6	MDEN-23
Mol.Wt	1	0.63	−0.04	0.01	−0.07	−0.28	0	0.03	0.32	0.16	0.04	0.08
slogp		1	0.01	0.12	−0.06	−0.36	0.02	0.02	0.1	0.05	0.02	0.05
chiV5chain			1	0.03	0	−0.04	−0.01	0	−0.03	−0.01	−0.01	−0.01
SdSE-index				1	0	−0.07	−0.01	0	0.24	−0.02	−0.01	0.05
Ipc					1	0.05	0	0	−0.01	0	0	0
MMFF_21						1	−0.02	−0.01	−0.15	−0.05	−0.02	−0.03
MMFF_46							1	0	0.08	−0.01	0	0.41
MMFF_51								1	−0.01	0	0	0
T_2_N_0									1	0.08	0.04	0.15
T_N_Br_3										1	−0.01	0.09
T_N_Br_6											1	0
MDEN-23												1

in WEKA and achieve nearly equal performance (R^2) 0.76/0.75 on training and blind dataset receptively (http://crdd.osdd.net/raghava/toxipred/supple.php). A second model on 20 descriptors selected using CfsubsetEval module of WEKA shows R/R^2 0.849/0.72 and 0.781/0.6 with RMSE 0.570/0.665 on training and blind dataset receptively. A third model using SMO with Puk-kernel on 9 best-selected descriptors on training and blind dataset shows correlation (R) 0.846/0.756, R^2 0.71/0.57 with RMSE 0.574/0.693 (Table I, Fig. 2(A)) respectively. The MLR based model on selected 9 descriptors shows better performance as compared to non-linear WEKA based model (http://crdd.osdd.net/raghava/toxipred/supple.php).

Delivered by Publishing Technology to: Guest User
IP: 14.139.234.82 On: Sun, 11 May 2014 ...
Copyright: American Scientific Publishers

CDK Descriptors Based Model

The second model developed on train dataset and tested on blind dataset (randomly created) shows $R_{\text{train}}/R^2_{\text{train}}$ 0.876/0.77, $R_{\text{blind}}/R^2_{\text{blind}}$ 0.85/0.72 on training and blind dataset using PLS in WEKA. We have developed a QSAR model on 11 significant descriptors selected using CfsubsetEval achieved $R_{\text{Train}}/R_{\text{Blind}}$ 0.867/0.816, $R^2_{\text{Train}}/R^2_{\text{Blind}}$ 0.75/0.65 with RMSE 0.537/0.620 on training and blind dataset receptively (Table I). We further number of descriptors from 11 to 6 and developed a third model, shows nearly same performance (in term of R^2) as shown in (http://crdd.osdd.net/raghava/toxipred/supple.php), and Figure 2(B).

Hybrid Model

The hybrid model developed on randomly selected blind dataset was also shows better performance on minimum of only 17 relevant descriptors (for detail see (http://crdd.osdd.net/raghava/toxipred/supple.php)). First, hybrid model on 31 descriptors shows R/R^2 0.88/0.77 on training and 0.83/0.69 on blind dataset (Fig. 2(C)). A second model on 17 descriptors shows a better performance with R 0.89/0.85, R^2 0.79/0.72 with RMSE 0.491/0.557 (Table I, Fig. 2(D)) on training and blind dataset respectively.

Table III. Correlation between selected CDK descriptors.

	ATSc3	BCUTc-11	FNSA-1	RPSA	GRAV-3	XLogP
ATSc3	1	0.13	−0.11	0.18	−0.05	−0.07
BCUTc-1l		1	0.18	−0.33	0.26	0.3
FNSA-1			1	0.12	0.45	0.09
RPSA				1	−0.11	−0.48
GRAV-3					1	0.58
XLogP						1

Interpreting Best QSAR Model

As shown in (http://crdd.osdd.net/raghava/toxipred/supple.php), models developed using all three techniques MLR, PLS and SMO performed equally well. It is also observed that hybrid model developed using descriptors obtained from two or more than two software perfom better than model developed using descriptors obtained an individual software. It is clear from results that free software of descriptor calculation are sufficiently powerful for developing QSTR models. The descriptor selection is important for developing QSTR model and for improving the performance of model. In this study, we also used free software even for feature selection. This study shows that linear model is performing more or less equal to non-linear models. From the bunch of thousands of descriptors, only few statistically relevant descriptors were selected and used for the model development. Our study also suggests that among the vast of molecular descriptors molecular weight, and xlogP plays significant role in toxicity prediction of chemical compounds. From the Table III, Table IV, Table V it was found that descriptors selected in this study shows very little correlation with each other and more with toxicity value. The positive correlation with mol. wt. and solubility (xLogP) are following the two of the four parameter famous lipinski rule of five. The XlogP descriptor identified in our study also identified in past by Schuurmann et al. 2003 as important in predicting the toxic molecules. The possible reason may be that with increasing molecular weight of a compound its solubility decrease and therefore, that particular compound may persist in the body of *T. pyriformis* and become toxic.

Table IV. Correlation between selected V-life and CDK descriptors.

		CDK					
	Descriptors	ATSc3	BCUTc-11	FNSA-1	RPSA	GRAV-3	XLogP
V-life	Mol.Wt.	−0.08	0.23	0.47	−0.19	0.86	0.64
	chiV5chain	0.03	0.09	−0.04	0.01	−0.03	−0.03
	SdSE-index	0.03	0.17	0.02	−0.12	0.07	0.15
	Ipc	−0.02	−0.05	−0.04	0.04	−0.1	−0.06
	MMFF_21	−0.22	−0.49	−0.22	0.09	−0.37	−0.3
	MMFF_46	−0.01	0.02	0.01	−0.04	0.02	−0.01
	MMFF_51	−0.03	−0.05	−0.01	−0.03	0.05	0.06
	T_2_N_0	−0.08	0.29	0.55	−0.24	0.4	0.16
	T_N_Br_3	−0.02	0.1	0.16	−0.04	0.07	0.03
	T_N_Br_6	−0.02	0.1	0.03	−0.04	0.02	0.01
	MDEN-23	−0.02	0.04	0.05	−0.06	0.09	0.03

Web Server

One of the major challenges for researchers working in the field of toxicology is to predict the toxicity of a chemical compound. Best of our knowledge, two free softwares namely T.E.S.T (Toxicity Estimation Software Tool) and OpenTox are available for toxicity prediction.[25,26] In order to complement existing effort for providing service to community, we developed a web server "ToxiPred" (http://www.crdd.osdd.net/raghava/toxipred), for predicting toxicity of molecules. We integrate JME molecular editor[27] in ToxiPred that allows user to draw their molecules of choice. This server is launch using Apache under Linux (Red Hat) environment. The common gateway interface (CGI) script of ToxiPred is written using PERL version 5.03. This is a user friendly web server that allows to predict pIGC50 of a small chemical against *Tetrahymena pyriformis.*

Open Science Online Technology to: Guest User
IP: 14.139.234.82 On: Sun, 18 May 2014 11:13:34
Copyright: American Scientific Publishers

Table V. Correlation between descriptors and PIGC50 values.

No.	Descriptor name	Correlation (PIGC50)	*P*-value
1	Mol.Wt.	0.7	0.011
2	chiV5chain	−0.03	0.93
3	SdSE-index	0.22	1.72E-13
4	Ipc	−0.08	0.30
5	MMFF_21	−0.39	1.47e-0.9
6	MMFF_46	0.08	1.96E-05
7	MMFF_51	0.05	0.67
8	T_2_N_0	0.38	0.008
9	T_N_Br_3	0.12	0.003
10	T_N_Br_6	0.06	0.020
11	MDEN-23	0.1	0.969
12	ATSc3	−0.14	8.78E-11
13	BCUTc-1l	0.32	0.755
14	FNSA-1	0.39	7.32E-12
15	RPSA	−0.31	0.222
16	GRAV-3	0.7	7.97e-0.7
17	XLogP	0.75	<2e-16

DISCUSSION

In this study, we have developed several models for the prediction of pIGC50 of small organic chemical molecules in *Tetrahymena pyriformis*. Our present study suggests that descriptors are keys for any QSAR modeling but it is not always possible that all relevant descriptors were implemented in single software. Thus it is important to use descriptors obtained from different software for developing a model. As shown in our result section our hybrid model based on selected set of descriptors obtain from different software performed better than models developed using descriptors of individual software. After analyzing selected descriptors from different software, we found that molecular weight, and xlogP shows very high correlation (>= 0.70) with pIGC50 value. In order to provide the facility to scientific community, we developed a webserver "ToxiPred" to predict toxicity of chemical compounds. We hope that present model will aid in the area of drug designing.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

NKM, DS and SA perform all the analysis of data and developed QSAR models. DS, SA and NKM have developed server ToxiPred. Manuscript have been written by DS and NKM, OSDD members provides suggestions to this work. GPSR conceived and gave overall supervision to the project, helped in interpretation of data and refined the drafted manuscript. All authors read and approved the final manuscript.

Acknowledgments: The authors are thankful to Open Source for Drug Discovery (OSDD) foundation and Council for Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), New Delhi, India for financial support. The manuscript has Institute of Microbial Technology (IMTECH) communication no. 032/2011.

REFERENCES

1. D. L. Hill, The Biochemistry and Physiology of Tetrahymena, edn., Academic Press, New York **(1972)**, pp. 230.
2. J. G. Hengstler, H. Foth, R. Kahl, P. J. Kramer, W. Lilienblum, T. Schulz, and H. Schweinfurth, The reach concept and its impact on toxicological sciences. *Toxicology* 220, 232 **(2006)**.
3. P. R. Duchowicz, A. G. Mercader, F. M. Fernandez, and E. A. Castro, Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR. *Chemometr. Intell. Lab. Syst.* 90, 97 **(2008)**.
4. C. Hansch and A. Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York **(1979)**.
5. M. T. D. Cronin and J. C. Dearden, Review, QSAR in toxicology, 1. Prediction of Aquatic Toxicity. *Quant. Struct. Act. Relat.* 14, 1 **(1995)**.
6. K. Roy and G. Ghosh, QSTR with extended topochemical atom (ETA) indices. 12. QSAR for the toxicity of diverse aromatic

compounds to Tetrahymena pyriformis using chemometric tools. *Chemosphere* 77, 999 **(2009)**.
7. A. M. Richard, Structure-based methods for predicting mutagenicity and carcinogenicity: Are we there yet? *Mutat. Res.* 400, 493 **(1998)**.
8. F. Cheng, J. Shen, Y. Yu, W. Li, G. Liu, P. W. Lee, and Y. Tang, *In-silico* prediction of tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere* 82, 1636 **(2011)**.
9. G. Klopman, S. K. Chakravarti, H. Zhu, J. M. Ivanov, and R. D. Saiakhov, ESP: A method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases. *J. Chem. Inf. Comput. Sci.* 44, 704 **(2004a)**.
10. G. Klopman, H. Zhu, M. A. Fuller, and R. D. Saiakhov, Searching for an enhanced predictive tool for mutagenicity. *SAR QSAR Environ. Res.* 15, 251 **(2004b)**.
11. P. Mazzatorta, E. Benfenati, D. Neagu, and G. Gini, The importance of scaling in data mining for toxicity prediction. *J. Chem. Inf. Comput. Sci.* 42, 1250 **(2002)**.
12. P. Mazzatorta, E. Benfenati, C. D. Neagu, and G. Gini, Tuning neural and fuzzy-neural networks for toxicity modeling. *J. Chem. Inf. Comput. Sci.* 43, 513 **(2003)**.
13. A. M. Richard, Commercial toxicology prediction systems: A regulatory perspective. *Toxicol. Lett.* 102, 611 **(1998a)**.
14. A. M. Richard, Future of toxicology–predictive toxicology: An expanded view of chemical toxicity. *Chem. Res. Toxicol.* 19, 1257 **(2006)**.
15. F. Cheng, J. Shen, Y. Yu, W. Li, G. Liu, P. W. Lee, and Y. Tang, *In silico* prediction of tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere* 82, 1636 **(2011)**.
16. C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen, Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* 12, 2111 **(2006)**.
17. A. M. Richard, C. Yang, and R. S. Judson, Toxicity data informatics: Supporting a new paradigm for toxicity prediction. *Toxicol. Mech. Methods* 18, 103 **(2008)**.
18. A. Garg, R. Tewari, and G. P. Raghava, KiDoQ: Using docking based energy scores to develop ligand based model for predicting antibacterials. *BMC Bioinformatics* 11, 125 **(2010)**.
19. N. K. Mishra, S. Agarwal, and G. P. Raghava, Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol.* 10, 8 **(2010)**.
20. N. K. Mishra and G. P. S. Raghava, Prediction of specificity and cross-reactivity of kinase inhibitors. *Letters in Drug Design and Discovery* 8, 223 **(2011)**.
21. D. Singla, M. Anurag, D. Dash, G. P. S. Raghava, A web server for predicting inhibitors against bacterial target GlmU protein. *BMC Pharmacol* 6, 11 **(2011)**.
22. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and H. I. Witten, The WEKA data mining software: An update. *SIGKDD Explorations* 11, 10 **(2009)**.
23. T. Knebel, S. Hochreiter, and K. Obermayer, An SMO algorithm for the potential support vector machine. *Neural Comput.* 20, 271 **(2008)**.
24. T. I. Netzeva, A. Gallego Saliner, and A. P. Worth, Comparision of applicability of domain of a quantitative structure-activity relationship for estrogenecity with a large chemical inventry. *Environ. Toxicol. Chem.* 25, 1223 **(2006)**.
25. T. M. Martin, P. Harten, R. Venkatapathy, S. Das, D. M. Young, A hierarchical clustering methodology for the estimation of toxicity. *Toxicol. Mech. Methods.* 18, 251 **(2008)**.
26. B. Hardy, N. Douglas, C. Helma, M. Rautenberg, N. Jeliazkova, V. Jeliazkov, I. Nikolova, R. Benigni, O. Tcheremenskaia, S. Kramer, T. Girschick, F. Buchwald, J. Wicker, A. Karwath, M. Gütlein, A. Maunz, H. Sarimveis, G. Melagraki, A. Afantitis, P. Sopasakis, D. Gallagher, V. Poroikov, D. Filimonov, A. Zakharov, A. Lagunin, T. Gloriozova, S. Novikov, N. Skvortsova, D. Druzhilovsky, S. Chawla, I. Ghosh, S. Ray, H. Patel, and S. Escher, Collaborative development of predictive toxicology applications. *J. Cheminform.* 31 2, 7 **(2010)**.
27. Java Molecular Editor [http://www.molinspiration.com/jme/].

Provided by Indian Institute of Technology Kanpur (IITK)
IP: 14.139.234.82 On: Sun, 11 Oct 2015 10:10:17
Copyright: American Scientific Publishers