Available online on www.ijtpr.com
International Journal of Toxicological and Pharmacological Research 2015; 7(6); 303-308

ISSN: 0975-5160

Research Article

Machine Learning Classifier Algorithms to Predict Endocrine Toxicity of Chemicals

Renjith P[1]*, Jegatheesan K[2]

[1]*Cognizant Technology Solutions, 3rd Floor, Athulya, Infopark SEZ, Kakkanad, Kochi 682030, Kerala, India.*
[2]*Center for Research and PG Studies in Botany and Department of Biotechnology, Thiagarajar College (Autonomous), Madurai - 625 009, Tamilnadu, India.*

Available Online:30th November, 2015

ABSTRACT

Bayesian Logistic Regression (BLR), Bayesian Network (BN), Naïve Bayes (NB), Logistic Regression (LR), Artificial Neural Networks (ANN), Simple Logistic (SL), Lazy Learning (LL), Random Forest (RF), Rotation Forest (Rot-F), and C4.5 (J48) machine learning classifier algorithms were examined to build predictive models for endocrine disrupting chemicals. Datasets of Estrogen Receptor (ER) and Androgen Receptor (AR) disrupting chemicals along with their Binding Affinity (BA) values were used as knowledgebase for building the predictive models. Substructure fingerprints (fragment counts) of knowledgebase chemicals were generated using Kier Hall Smarts topological descriptor that utilizes electrotopological state (e-state) indices. LL, Rot-F, and RF algorithms tested on ER training set (200 molecules) gave superior prediction models with Kappa statistic values of 0.89, 0.85, and 0.89. Whereas in case of AR training set (170 molecules), LL, and RF algorithms gave promising results with Kappa statistic values of 0.95, and 0.93. Each model built using classifier algorithms were tested on both AR and ER test datasets (32 and 24 molecules each) and prediction accuracies of classifying chemicals as endocrine disruptor or non-endocrine disruptor had been calculated. BN, NB, Rot-F, RF, and J48 algorithms on ER test dataset showed prediction accuracy above 80% with RMSE values 0.36, 0.37, 0.38, 0.37, and 0.41 respectively. Whereas in case of AR test dataset, BLR, SL, and LL algorithms gave better results with prediction accuracy above 75% with RMSE values of 0.5, 0.49, and 0.47 respectively. Finally, the significance of chemical substructures towards endocrine disruption activity was characterized and ranked, on the basis of molecular descriptors identified by the RF method.

Keywords: Endocrine Disruptor Chemical, Toxicity Prediction, Classifier Algorithm, Machine Learning, QSAR

INTRODUCTION

In the last decade, we have witnessed a perfect storm in terms of the impact on our ability to predict chemical toxicity. Compared to a decade ago, compute power is far cheaper, resulting in faster predictions. Hardware continues to shrink at a rate that much of the available modelling software can now easily run on a laptop or notebook computer. The move to cloud-based servers means that compute power is available for intensive calculations with results served up easily via web-based interfaces. The number of chemical compounds now available via public databases is in the tens of millions. The data associated with the structures is provided in a manner that allows it to be downloaded and used to build computational models. There are also increasing amounts of data collated for toxicity endpoints, such as for drug-induced liver injury (DILI)[1,2]. the human Ether-à-go-go Related Gene (hERG) ion channel[3,4]. Pregnane X receptor (PXR)[5] and for transporters involved in toxicities[6-8]. There are far more examples of computational models and software that are now freely accessible that can be used for toxicity prediction, and as the community becomes aware of these tools they will be used increasingly often. Computational toxicity prediction is certainly starting to reach more researchers and is becoming more widely accepted9. Based on the discourse in so many venues like online, in journals, and in popular media, it is clear that the pharmaceutical industry is seen as having reached a productivity tipping point, and the cost of new drug development is extremely high[10]. Initiatives such as REACH and other new legislations are likely to require even more testing of compounds for which there is insufficient information on the hazards that they pose to human health and the environment. As a response to this need, initiatives such as ToxCast intend to build up ways of predicting potential toxicity and developing a commercial approach for prioritizing the thousands of chemicals that require toxicity testing. Ultimately, we need to understand the toxicity of far more compounds than is reasonable or ethical to screen in vivo in animals or even in vitro. There has to be a first tier prioritization to filter molecules of interest to pharmaceutical, consumer products, and environmental researchers. The predictive models that have been built in recent years, and those that are becoming available with efforts such as eTox and OpenTox, may be of some utility. So a perfect storm of many factors could

**Author for Correspondence*

Table 1: Results of different classifier algorithms applied on Estrogen dataset.

Entry	Classifier Algorithm	Training Set – 200 Molecules				Test Set – 32 Molecules			
		Kappa statistic	Root Mean Squared Error (RMSE)	Correctly Classified Molecules	Incorrectly Classified Molecules	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Correctly Classified Molecules	Incorrectly Classified Molecules
1	Bayesian Logistic Regression	0.59	0.45	160	40	0.41	0.41	19	13
2	Bayes Network	0.61	0.4	162	38	0.28	0.36	27	5
3	Compliment Naïve Bayes	0.51	0.48	153	47	0.19	0.43	26	6
4	Naïve Bayes	0.51	0.42	153	47	0.25	0.38	25	7
5	Naïve Bayes Multinomial	0.52	0.43	154	46	0.28	0.37	26	6
6	Logistic Regression	0.6	0.36	160	40	0.4	0.49	18	14
7	Neural Network – Multi Layer Perceptron	0.75	0.31	176	24	0.3	0.48	23	9
8	Simple Logistic	0.6	0.37	161	39	0.4	0.47	17	15
9	Lazy – IB1	0.89	0.26	187	13	0.31	0.56	22	10
10	Lazy - IBk	0.89	0.18	189	11	0.32	0.56	22	10
11	Lazy - KStar	0.89	0.18	189	11	0.37	0.52	21	11
12	Rotation Forest	0.85	0.26	185	15	0.3	0.38	26	6
13	Random Forest	0.89	0.21	189	11	0.25	0.37	28	4
14	C4.5 - J48	0.76	0.31	177	23	0.25	0.41	26	6

position computational models well for predicting human toxicities for the future. Endocrine disruptors are compounds occurring in nature or synthetic substances that may mimic or interfere with the function of hormones in the body. Endocrine disruptors may turn on, shut off, or modify signals that hormones carry, which may affect the normal functions of tissues and organs. Many of these substances have been associated with immune, developmental, neural, reproductive and other problems in laboratory and wildlife animals. These chemicals are also referred as endocrine active compounds, environmental hormones and endocrine modulators. Though the best studied is the environmental chemicals with estrogenic activity, chemicals with progesterone, androgen, anti-estrogen, anti-androgen have also been identified.A wide and varied range of substances including DDT, diethylstilbestrol or the synthetic estrogen DES, polychlorinated biphenyls (PCBs), dioxin and dioxin-like compounds and some other pesticides are thought to cause endocrine disruption. Endocrine disruptors may cause male fertility reductions declining the numbers of males born, male reproductive organ abnormalities, Female reproductive health issues including early puberty, early reproductive senescence and fertility problems, Increases in ovarian, prostate and mammary cancers, some neurodegenerative diseases and immune and autoimmune diseases increase. Effects on obesity and diabetes are observed when Endocrine disrupting chemicals with estrogenic activity like BPA have been exposed. Computational predictive toxicology models are important alternative methods to decrease dependency on slow and costly animal experiments. Machine learning methodology studies automatically learn to make accurate predictions based on past observations. They are often much more accurate than human-crafted rules without involving a human expert or programmer. It is a cheap and flexible automatic method to search for hypotheses explaining data and can be applied to any learning task.Application of Computational approaches in toxicity prediction and drug discovery have need of molecular descriptors that imitate chemicals' physicochemical properties and structural information. Quantitative structure-property relationship (QSPR) and quantitative structure-activity relationship (QSAR) prognostic models are established by comparing

Table 2: Results of different classifier algorithms applied on Androgen dataset.

Entry	Classifier Algorithm	Training Set – 170 Molecules				Test Set – 24 Molecules			
		Kappa statistic	Root Mean Squared Error (RMSE)	Correctly Classified Molecules	Incorrectly Classified Molecules	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Correctly Classified Molecules	Incorrectly Classified Molecules
1	Bayesian Logistic Regression	0.46	0.38	145	25	0.25	0.5	18	6
2	Bayes Network	0	0.41	133	37	0.48	0.55	13	11
3	Compliment Naïve Bayes	0.39	0.43	139	31	0.42	0.65	14	10
4	Naïve Bayes	0.39	0.4	139	31	0.35	0.56	16	8
5	Naïve Bayes Multinomial	0.37	0.39	140	30	0.39	0.56	14	10
6	Logistic Regression	0.52	0.32	146	24	0.41	0.53	16	8
7	Neural Network – Multi Layer Perceptron	0.66	0.29	153	17	0.29	0.5	17	7
8	Simple Logistic	0.46	0.33	145	25	0.36	0.49	18	6
9	Lazy – IB1	0.95	0.13	167	3	0.33	0.58	16	8
10	Lazy - IBk	0.95	0.09	167	3	0.31	0.55	16	8
11	Lazy - KStar	0.95	0.1	167	3	0.29	0.47	18	6
12	Rotation Forest	0.66	0.26	153	17	0.38	0.49	16	8
13	Random Forest	0.93	0.13	166	4	0.35	0.45	17	7
14	C4.5 - J48	0.54	0.33	148	22	0.44	0.58	14	10

the structural information with properties or activities. The representation of structural information in the models is a problem as the structures are complicated for use themselves in the models. The structural information is extracted in digital representation or numerical form suitable for model development by Molecular descriptors that serve as the link between the Chemicals' physicochemical properties or biological activities and its molecular structures.An effective molecular descriptor represents chemical properties or chemical structure features varying across a set of chemicals in the same way to a biological end point related with the chemicals. Insights of a basic mechanism of action, increasing drug's efficacy or ways of reducing its toxicity is provided by descriptors open to physical construal in a biological perspective. Regression and classification models derived from enormous descriptor sets of impenetrable physical meaning repeatedly yielded relatively greater precision and reliability than the physically interpretable descriptors. The possibility that 3D descriptors might genuinely yield worse predictions due to an inability to locate the bioactive conformer is possible[11]. Alternatively, 2D methods could correlate strongly with simple physicochemical properties such as logP[12]. that may correlate with nonspecific contributions to binding affinity, rather than the specific contributions that are required for, desirable, compound selectivity[3] and may be better captured by 3D approaches. Hence, 2D descriptors may outperform/perform as well as 3D descriptors for "biased" datasets in which, say, the set of actives was disproportionately made up of topologically similar molecules and/or active and inactive were well separated in terms of simple physicochemical properties[13]. even when the 3D descriptors would be valuable under circumstances where such conditions did not apply.A rapid and thriving technique for the prediction of different endocrine disrupting chemicals was developed in the current research. A simple molecular similarity computation blended with machine learning algorithms was the basis of this method. Predictive models were built based on the binding affinity values of Estrogen Receptor (ER) and Androgen Receptor (AR) disrupting chemical datasets. The significance of chemical substructures towards endocrine disruption activity was characterized and ranked, based on RF method.

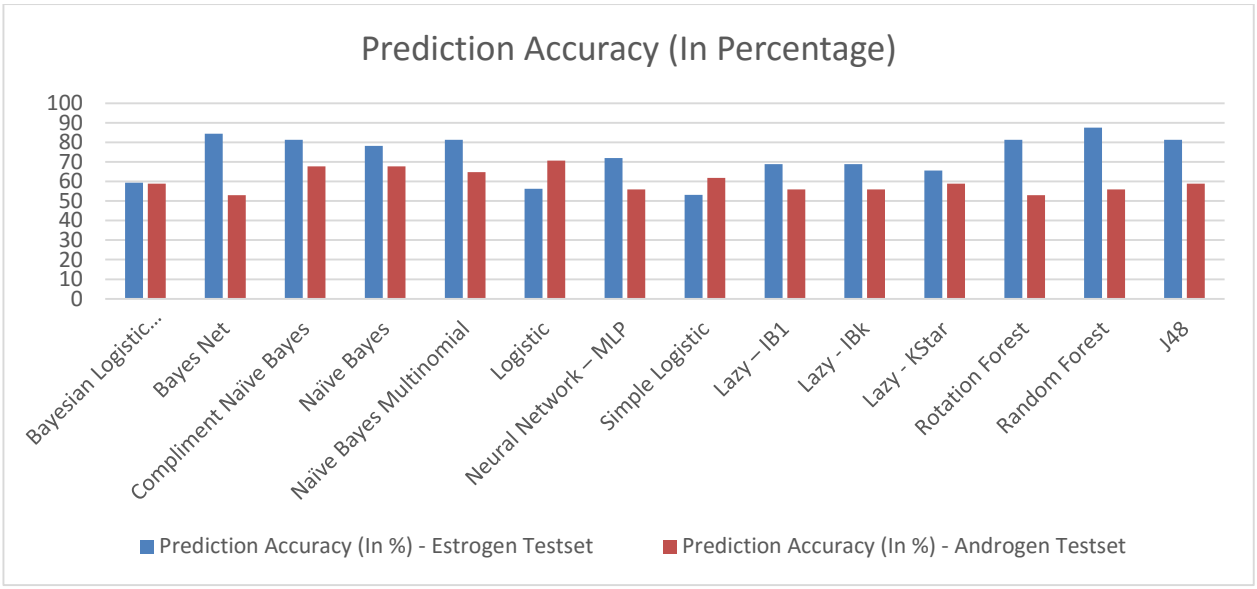
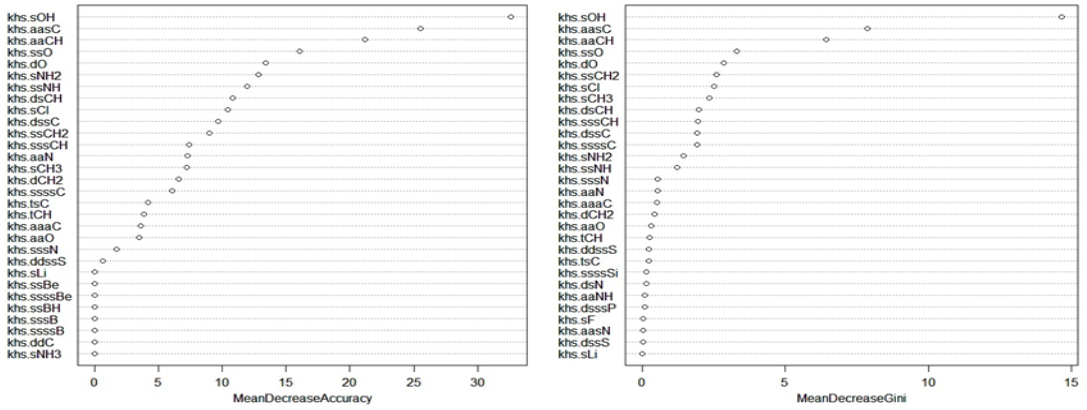


Figure 1: Prediction accuracy of multiple classifier algorithms on Estrogen and Androgen datasets.

A. Estrogen dataset



B. Androgen dataset

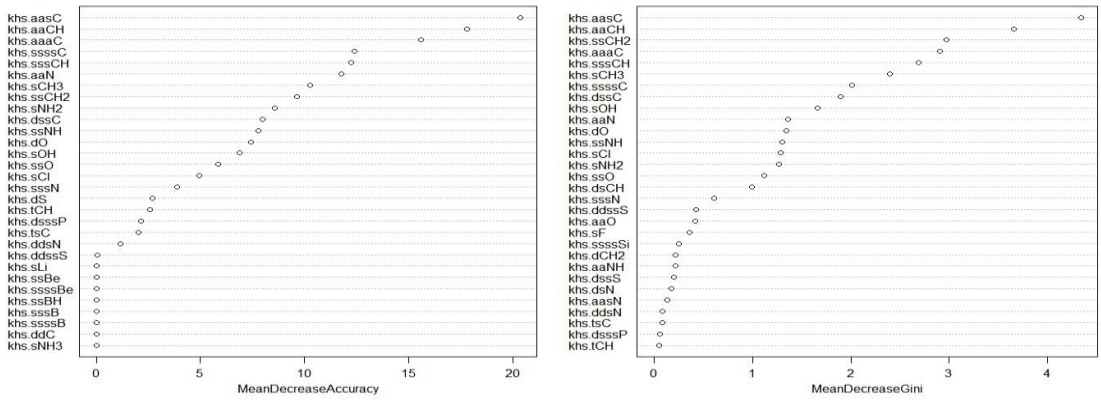


Figure 2: Substructure toxicological significance: Dot chart of variable importance as measured by a Random Forest.

MATERIALS AND METHODS

Data and Software Availability

ER and AR binding datasets used in this study were downloaded as Structure Data Files (SDF) from FDA Endocrine Disruptor Knowledgebase website. All molecular structures were standardized and compounds with ambiguous activity values as well as the duplicates were removed in the preprocess step. The open source pipeline generation platform KNIME v.2.10.0 software was used for the data preparation and screening tasks. Rcdk package of R-Programming had been used to generate Kier Hall Smarts topological descriptors of dataset molecules. Weka v.3.6 data mining software was

utilized to apply machine learning classifier algorithms and build predictive models. R-Programming randomForest package had been used to apply RF algorithm and examine substructure toxicological significance of molecules.

Data Preparation

ER and AR binding dataset molecules were preprocessed for structural consistency errors and rectified. Total dataset consisted of 232 ER and 194 AR tested molecules with their Relative Binding Affinity (logRBA) values. A Spreadsheet was prepared with molecule IDs and corresponding logRBA values. Each molecule record was categorized and labeled as "disruptor" or "non-disruptor" based on the logRBA value ranges. logRBA from -4.5 to 2.6 was categorized as disruptor and value -10000 as non-disruptor. A new a new spreadsheet was developed with these conversions for next step of descriptor calculation.

Molecular Fingerprint Calculations

Hall Smarts topological descriptor set which incorporates 79 electrotopological state (e-state) key indices values was generated for molecule datasets using Rcdk package in R programming.

Training and Test Datasets

Datasets were randomly separated into training and test sets to build and validate predictive models. ER dataset of 232 molecules was divided into 200 molecules of training and 32 molecules of test sets. Similarly AR dataset of 194 molecules was divided into 160 molecules of training and 34 molecules of test sets.

Machine Learning Classifier Algorithms

Bayesian Logistic Regression (BLR), Bayesian Network (BN), Naïve Bayes (NB), Logistic Regression (LR), Artificial Neural Networks (ANN), Simple Logistic (SL), Lazy Learning (LL), Random Forest (RF), Rotation Forest (Rot-F), and C4.5 (J48) Machine Learning classifier algorithms were experimented in this study. Weka data mining software was used to build and test different machine learning algorithms into the datasets. The predictor algorithm was supplied with spreadsheets of active and inactive molecules with their fingerprints for statistical model building. Predictive models were built using the ER and AR training datasets for every classifier algorithms tested and prediction accuracies evaluated using test datasets subsequently.

RESULTS AND DISCUSSION

Training Set and Model Building

Fourteen different classifier algorithms were applied and experimented using training datasets. The results of different classifier algorithms tested on ER and AR training datasets are given in Table 1 and Table 2 respectively. Best prediction model was selected based on the Kappa statistic and Root Mean Squared Error (RMSE) values. Models with Kappa values 0.61 – 0.80 and 0.81 – 0.99 ranges were considered to be substantial agreement and almost perfect agreement with the hypothesis respectively. Models were also checked for the lowest RMSE value, less number of incorrectly classified Molecules and highest number of correctly classified molecules. Based on these criteria, LL, Rot-F, and RF algorithms were chosen as the best models for ER training dataset and LL, and RF algorithms for AR dataset.

Test Set Validations

The results of different predictive models applied on ER and AR test datasets are given in Table 1 and Table 2 respectively. Prediction accuracies are measured based on the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) values and numbers of correctly and incorrectly classified molecules. BN, NB, Rot-F, RF, and J48 models gave significant prediction accuracies (>80%) on ER test dataset with RMSE values 0.36, 0.37, 0.38, 0.37, and 0.41 respectively. Whereas in case of AR test dataset, BLR, SL, and LL models gave promising results with prediction accuracy above 75% with RMSE values of 0.5, 0.49, and 0.47 respectively. The present work was compared with previously published validations of these models with different datasets[14,15].

Prediction Accuracy of Classification Algorithms - A Quality Chart

The power of various classifier algorithms on predicting endocrine disruption property of chemicals was calculated and a comparison chart was prepared. Prediction accuracy was calculated using the following equation. Prediction accuracy (In Percentage) = (Number of correctly classified molecules / Total number of molecules in test dataset) * 100 The comparison of prediction accuracies for different classifier algorithms on ER and AR test datasets is depicted in Figure 1. The individual models accuracies ranged for AR disruptor between 52% - 70% and for ER disruptor between 52% to 88%. This was nearer to the range of 70% to 82% as observed for cross validation of models for toxicity prediction[16].

Substructure Toxicological Significance

Random Forest

Random Forest classifier algorithm in R programming was used to identify and rank the substructures of molecules based on their contribution on Estrogen and Androgen disruption toxicity. Dot chart of variable importance as measured by a Random Forest give substructures contributing to toxicity is given in Fig. 2.

Hydroxyl group (sOH), aCa- (aasC), aCHa (aaCH), -O- (ssO) and =O (dO) were identified as the top five substructures that lead to Estrogen toxicity as given in Figure 2a. The top five substructures that lead to Androgen toxicity were identified to be aCa- (aasC), aCHa (aaCH), aaaC, >C< (ssssC), >CH- (sssCH) as shown in Fig. 2b. This substructure importance evaluation using random forest classification method gives insights into toxicity analysis and lead optimization[16].

CONCLUSION

In this study, we compared 14 classification algorithms for predicting the endocrine disruption ability of compounds. All supervised classifier models were built with 200 ER and 160 AR disruptor molecules as training datasets. Fingerprints of training as well as test datasets were generated with Kier Hall Smarts topological descriptor. Accurate models were built using LL, Rot-F, and RF algorithms on ER and LL, and RF algorithms for AR

training datasets. Evaluations of models demonstrate that, BN, NB, Rot-F, RF, and J48 models for ER dataset and BLR, SL, and LL models for AR dataset gave best results respectively. Random forest classifier provided the top five substructures that lead to ER and AR toxicity making way for predicting the endocrine disrupting chemicals.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

REFERENCES

1. Cruz-Monteagudo M, Cordeiro MN, Borges F. Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. J. Comput. Chem. 2007; 29:533-549.
2. Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ, Tropsha A. Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. Chem. Res. Toxicol. 2010; 23:171-183.
3. Shamovsky I, Connolly S, David L, Ivanova, S, Nordén B, Springthorpe, B, Urbahns K. Overcoming undesirable HERG potency of chemokine receptor antagonists using baseline lipophilicity relationships. J. Med. Chem. 2008; 51:1162-1178.
4. Hansen K, Rathke F, Schroeter T, Rast G, Fox T, Kriegl JM, Mika S. Bias-correction of regression models: a case study on hERG inhibition. J. Chem. Inf. Model. 2009; 49:1486-1496
5. Pan Y, Li L, Kim G, Ekins S, Wang, H, Swaan PW. Identification and validation of novel human pregnane Xreceptor activators amongst prescribed drugs via ligand-based virtual screening. Drug Metab. Dispos. 2010; 39:337-344.
6. Diao L, Ekins S, Polli JE. Novel inhibitors of human organic cation/carnitine transporter (hOCTN2) via computational modeling and in vitro testing. Pharm. Res. 2009; 26:1890-1900.
7. Diao L, Ekins S, Polli JE. Quantitative structure activity relationship for inhibition of human organic cation/carnitine transporter. Mol. Pharm. 2010; 7:2120-2130.
8. Zheng X, Ekins S, Raufman JP, Polli JE. Computational models for drug inhibition of the human apical sodium-dependent bile acid transporter. Mol. Pharm. 2009; 6:1591-1603.
9. Ekins S. Computational Toxicology: risk assessment for pharmaceutical and environmental chemicals. John Wiley and Sons: Hoboken, NJ, USA, 2007.
10. Munos B. Lessons from 60 years of pharmaceutical innovation. Nat. Rev. Drug Discov. 2009; 8:959-968.
11. Sheridan RP, Kearsley S.K. Why do we need so many chemical similarity search methods? Drug Discov. Today 2002; 7:903–911.
12. Brown RD, Martin YC. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding, J. Chem. Inf. Comput. Sci. 1997; 37:1–9.
13. Rohrer SG, Baumann K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data, J. Chem. Inf. Model 2009; 49:169–184.
14. Dubus E, Ijjaali I, Petitet F, Michel A. In Silico Classification of hERG Channel Blockers: a Knowledge-Based Strategy. Chem. Med. Chem. 2006; 1:622–630.
15. Thai KM, Ecker GF. A binary QSAR model for hERG potassium channel blocker. Bioorg. Med. Chem. 2008; 16:4107–4119.
16. Drwal MN, Siramshetty VB, Banerjee P, Goede A, Preissner R, Dunkel M. Molecular similarity-based predictions of the Tox21 screening outcome, Frontiers in Environmental science 2015; 3:article 54.