QSAR Classification of ToxCast and Tox21 Chemicals on the Basis of Estrogen Receptor Assays

Qingda Zang (1), Daniel Rotroff (2, 3), Richard Judson (2)

(1) ORISE Postdoctoral Fellow at the U. S. Environmental Protection Agency (EPA), Research Triangle Park, NC, USA , (2) National Center for Computational Toxicology, U. S. EPA, Research Triangle Park NC, USA, (3) Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, NC, USA

Richard Judson | judson.richard@epa.gov | 919-541-3085

www.epa.gov

Abstract

The ToxCast and Tox21 programs have tested ~8,200 chemicals in a broad screening panel of *in vitro* high-throughput screening (HTS) assays for estrogen receptor (ER) agonist and antagonist activity. The present work uses this large *in vitro* data set to develop *in silico* QSAR models using machine learning (ML) methods and a novel approach to manage the imbalanced data sets seen in all targets we have tested. Training compounds from the ToxCast project were classified as active or inactive based on a composite ER Interaction Score derived from a collection of 13 ER *in vitro* assays. A total of 1,537 chemicals from ToxCast were used to derive and optimize the binary classification models while 5,073 additional chemicals from the Tox21 project were used to externally validate the model performance. QSAR classification models were built to relate the molecular structures of chemicals to their ER activities using LDA, CART, and SVM with 51 molecular descriptors from QikProp and 4328 structural fingerprints as explanatory variables. A random forest (RF) feature selection method was used to extract the structural features most relevant to ER activity. The performance was evaluated using various metrics, including overall accuracy, sensitivity, specificity, G-mean, as well as area under the ROC curve (AUC).

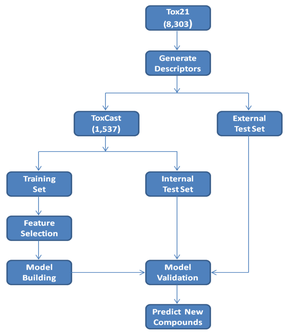

Figure 1. The flowchart for the whole classification process.

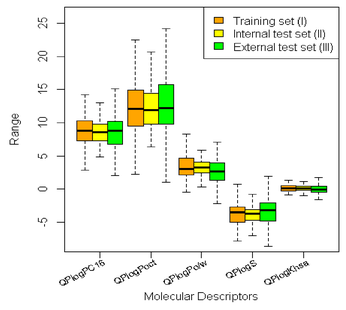

Figure 2. Data distribution of molecular descriptors for training and test sets.

Table 1. Data Sets of Chemicals Used for Classification Study

Data set	Total	Active	Inactive	Active/Inactive
Tox21	6610	435	6175	1:14.2
ToxCast	1537	264	1273	1:4.82
Training set (I)	1025	176	849	1:4.82
Internal test set (II)	512	88	424	1:4.82
External test set (III)	5073	171	4902	1:28.7

Methods

All data processing, multivariate analysis, and model building were implemented using the R statistical analysis software for Windows (Version 2.15.1). The packages *pheatmap*, *varSelRF*, *MASS*, *rpart*, *e1071* as well as *ROCR* in R were used to perform hierarchical clustering, feature selection, linear discriminant analysis (LDA), classification and regression tree (CART), support vector machine (SVM) and the receiver operating characteristic (ROC) analysis.

Classification Analysis – LDA, CART and SVM

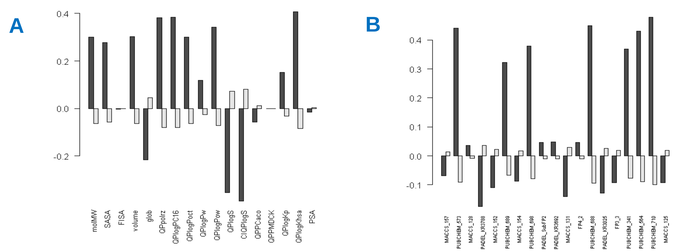

Figure 3. Group mean of molecular descriptors (A) and structural fingerprints (B). Black color: active class; Grey color: inactive class.

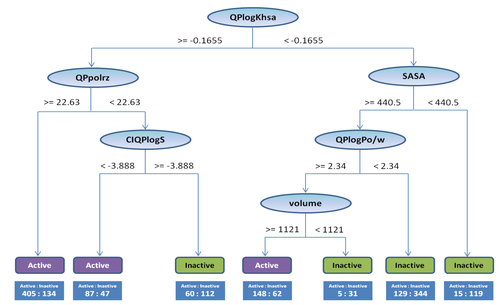

Figure 4. Classification trees for the model of molecular descriptors.

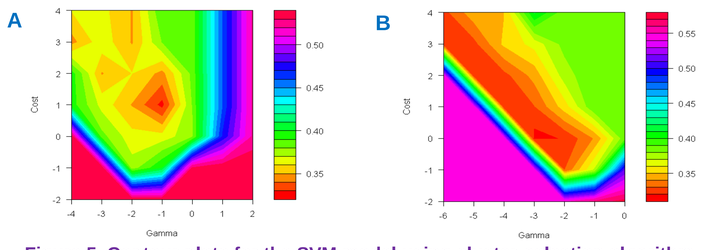

Figure 5. Contour plots for the SVM model using cluster-selection algorithm. (A) The molecular descriptor model; (B) The structural fingerprint model.

Feature Selection - Random Forests

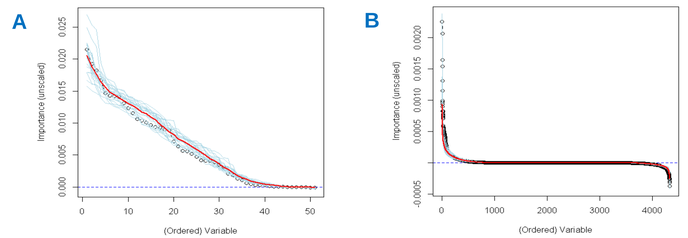

Figure 6. Feature importance computed from random permutations of RF. (A) 51 molecular descriptors from QikProp; (B) 4328 structural fingerprints.

Results

The best model was obtained using SVM in combination with a set of descriptors Identified from a large set via the RF algorithm, which recognized the active and inactive compounds with accuracies of 76.1% and 82.8%, and with a total accuracy of 81.6% on the internal test set and 70.8% on the external test set.

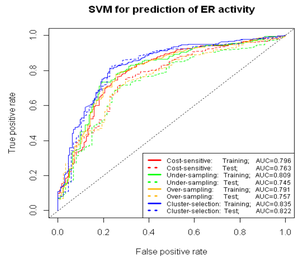

Figure 7. The ROC curves of the classifiers for descriptor model.

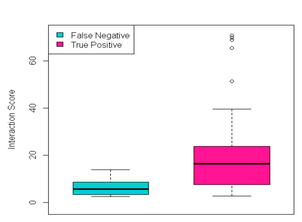

Figure 8. The boxplots from 88 active chemicals of the internal test set.

Conclusions

The present study demonstrates that a combination of high-quality experimental data and ML Methods can lead to robust models that achieve excellent predictive accuracy, which are potentially useful for facilitating the virtual screening of chemicals for environmental risk assessment.

References

Qingda Zang, Daniel M. Rotroff, Richard S. Judson, Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure−Activity Relationship and Machine Learning Methods. *J. Chem. Inf. Model.* 2013, *53(12)*, 3244–3261.