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).

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

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 82.6%, and with a total accuracy of 81.6% on the internal test set and 70.8% on the external 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