Predicting ToxCast™ and Tox21 Bioactivity Using Toxprint Chemotypes

Jeremy Fitzpatrick[1], Prachi Pradeep[1,2], and Grace Patlewicz[1]

1. US EPA, National Center for Computational Toxicology, Research Triangle Park, NC

2. Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

ORCID ID:0000-0002-5401-9706

Jeremy Fitzpatrick 1 Fitzpatrick.Jeremy@epa.gov

Abstract

The EPA ToxCast™ and Tox21 programs have generated bioactivity data for nearly 9076 chemicals across ~1192 assay endpoints; however, for over 70% of these chemical-assay endpoint pairs, there is no data. This creates certain challenges when trying to exploit the data to build predictive models. To fill the chemical-assay data gaps, we constructed random forest models for each assay endpoint, using chemotypes which are fragment based structural descriptors. For each model, the assay endpoint data was split into a training set containing 80% of the active chemicals and an equal number of inactives, with the remainder used as the test set. Many assay endpoints still lacked sufficient data to build robust models. However, 272 models with at least 200 chemicals in their training sets were successfully derived. 250 models of these were able to generate bioactivity predictions with greater than 60% balanced accuracy. Our models were able to predict ToxCast™ and Tox21 bioactivity values in the absence of experimental data which will facilitate the development of other predictive toxicity models.

Chemicals vs. Assay Endpoints: The white space represents areas where the chemical assay end point was not tested or where a hitcall could not be made.

Workflow

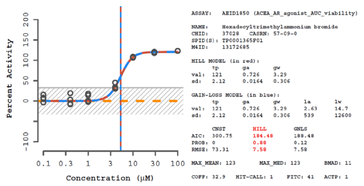
Gather Data — Retrieve all structural data and hitcall information that is available for ToxCast™ and Tox21 chemicals

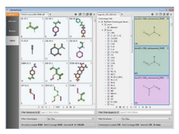
Build Models — Using the toxprint chemotypes as descriptors, build random forest models for each assay endpoint to predict the hitcall

Assess Accuracy — Determine which random forest models are capable of predicting their assay endpoints.

Hit Call and Chemotypes

An active hit call is defined as a signal three fold above bmad (baseline median absolute deviation) or a 20% change. The majority of chemicals with hit call information ~8500 also had structural information available from the Comptox Dashboard. The majority of those without available structures were mixtures.

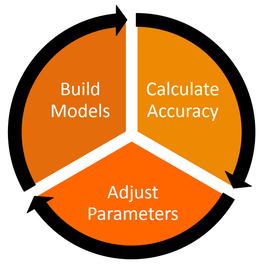




The Toxprint Chemotypes (toxprint.org) are a set of 729 chemical structural fragments that have been openly defined. We have chosen to use these as descriptors because they are freely available, and they have been used successfully to predict the hit calls of individual assays, within random forest models by other researchers within NCCT.

Random Forest Modeling

We constructed random forest models to predict each assay endpoint using the chemotypes as descriptors. A random forest is a collection of decision trees that vote for a given outcome using a majority rule. In our case, thousands of random forest models were built for each endpoint to determine the optimum set of parameters to use.



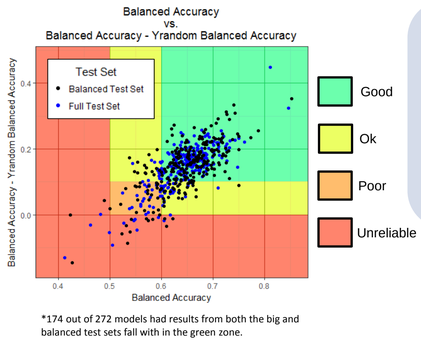
Build Model
- Each model was built using the gbm package in R
- Chemotypes were used as the descriptors
- 5-fold cross validation was carried out using the caret package in R

Calculate Accuracy
- Accuracy was calculated and averaged over all cross validations.

Adjust Parameters
- Results were used to adjust parameters for best fit
- Adjusted parameters were: number of trees, interaction depth, minimum number of samples to split, impact of each tree

Each random forest model was optimized in R using the caret package, with the gbm package being used to build the forest. The gbm package created a forest of gradient boosted decision trees. The caret package allows for easy optimization of the forest by optimizing the number of trees in each forest, the total number of splits, minimum number of samples for a split, and impact of each tree on the outcome of the model.

Modeling Results

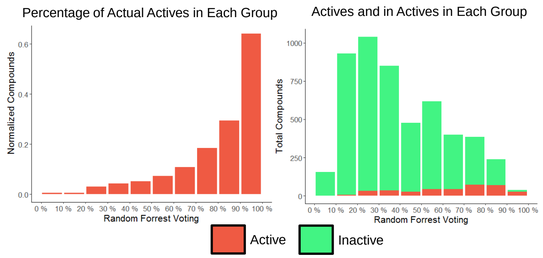


*174 out of 272 models had results from both the big and balanced test sets fall with in the green zone.

In order for a model to be considered successful the follow criteria had to be met:

- The training set had to contain at least 100 active and inactive compounds since this helped prevent over fitting
- A model needed to have a balanced accuracy of at least 60%
- The balanced accuracy of the model needed to be at least 10 points higher than for a Y-randomized model.

Total Number in Each Section	
382	49
85	28



The confidence of an individual prediction can be assessed based on the overall number of trees voting for or against an outcome. For example, a chemical which was predicted to be active is more likely to be a true active if 90% of trees vote active instead of only 60% of the trees voting active.

Top Five Models

Model Endpoint*	Balanced Accuracy on Large Test Set	Difference with Y-randomized model	Balanced Accuracy on Balanced Test Set	Difference with Y-randomized model
OT_ER_ERaERa_0480	0.810	0.449	0.889	0.481
OT_ER_ERaERb_0480	0.846	0.324	0.852	0.352
ATG_PPARa_TRANS_up	0.688	0.242	0.740	0.333
OT_ER_ERaERb_1440	0.716	0.255	0.725	0.313
BSK_CASM3C_TissueFactor_down	0.678	0.257	0.625	0.273

*Assay and Endpoint Descriptions Are Available From: https://actor.epa.gov/dashboard/

The alcohol aromatic phenol chemotype was the most influential in all three ER assay predictions. It's normalized influence ranged from 12 to 29%. Other five and six membered rings also had a large influence on the prediction outcome.

In contrast, the two non-ER assay endpoints modeled in the top five assays required a more diverse set of chemotypes in order to make a prediction. No single chemotype had a normalized influence above 8%.

Next steps

- Use physicochemical properties such as LogKow as descriptors and compare to current results
- Experiment with alternative descriptors such as those used in Mansouri et al[1] for the prediction of GPCR assays in ToxCast™
- Compare the balanced accuracies of the model predictions with the reproducibility of the experimental assays themselves

Conclusions

- We were able to successfully predict ToxCast™ and Tox21 assay outcomes in the majority of cases (174 predictions rated good out of 272 assays) where there were at least 200 chemicals available to make up a training set.
- The chemotypes conveyed a significant amount of information over a broad selection of the ToxCast™ and Tox21 assays

References

ToxCast ™ and Tox21 latest data releases available from: https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data

Structural information available from the CompTox Dashboard: https://comptox.epa.gov/dashboard

gbm package for R: Ridgeway, Greg. Generalized Boosted Regression Models. March 21, 2017. Documentation available at: https://cran.r-project.org/web/packages/gbm/gbm.pdf

Caret package for R: Kuhn, Max, et al. Classification and Regression Training. April 18, 2017. Documentation available at: https://cran.r-project.org/web/packages/caret/caret.pdf

1. Mansouri, Kamel, et al. In silico study of in vitro GPCR assays by QSAR modeling. In Silico Methods for Predicting Drug Toxicity. June 17, 2016, pp. 361-381.

U.S. Environmental Protection Agency
Office of Research and Development