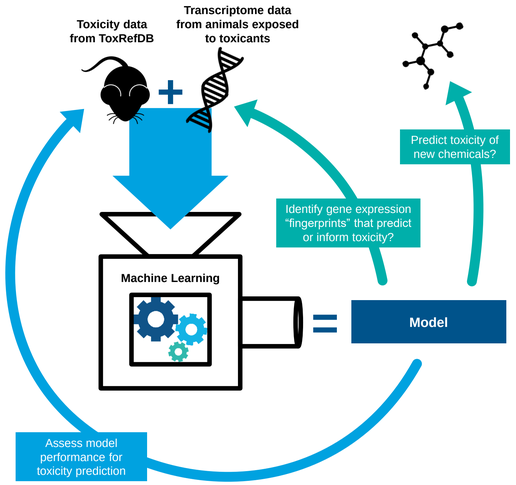
Empirical Prediction of In Vivo Toxicity Using Transcriptomic Data and Machine Learning

Joanne Trgovcich, Revathi Muralidharan, Alex Kliminsky, Sangeetha Kumar, Raga Avanasi, and Arun Varghese | ICF

Abstract

The past decade has seen significant research and regulatory initiatives devoted to the ultimate objective of supplanting in vivo toxicity studies with more humane and less expensive in vitro assays. The same period has witnessed considerable interest in toxicity prediction using quantitative structure activity relationship (QSAR) data. Numerous studies have shown both the potential and limitations of toxicity predictions based on both in vitro and QSAR data. Transcriptomic data generated using microarray data and next generation sequencing platforms offer a potentially rich set of predictors of cellular responses including toxic responses, but its potential for empirical predictions is less well studied. We propose and demonstrate a framework by which complex transcriptomic datasets can be transformed to a structure that may be combined with available data on animal testing to apply machine learning-based toxicity prediction. Specifically, we acquired transcriptome data from the NCBI Geo database for chemicals for which animal toxicity data were also reported in EPA's ToxRefDB database. We converted the toxicogenomics data to a set of predictors corresponding to gene groups with values reflecting levels of gene up-regulation or down-regulation. This approach considers each gene expression change as a unique feature, and therefore does not require any a priori knowledge of gene ontologies or pathways to infer relationships. Using this training dataset, we used a range of appropriate machine learning algorithms to predict toxicity at alternative levels of exposure concentrations, exposure duration, and cell types. We assessed model performance using 5-fold cross validation and found F1-scores ranging from 55% to 90% for alternative models. We discuss our findings and contrast the potential of the toxicogenomics and machine learning-based empirical prediction with traditional methods of assessing toxicity based on genomics data.

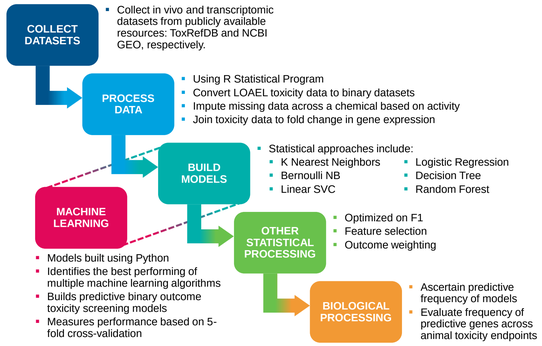
Approach to Prediction of In Vivo Toxicity Using Transcriptomic Data and Machine Learning



1 Background and Hypothesis

- Previously, we assessed the value of both in vitro assay and quantitative structure activity relationship (QSAR) data in predicting in vivo toxicity using numerous statistical models. (The Role of Feature Selection and Statistical Weighting in Predicting In Vivo Toxicity Using In Vitro Assay and QSAR Data, Wignall, J., Martin, M., Varghese, A., Trgovcich, J. Society of Toxicology Meeting, 2016).
- Incorporation of feature selection into our earlier models shows that some types of biological data are more useful than others for toxicity prediction.
- We therefore set out to test whether high throughput datasets such as genomic, transcriptomic, proteomic, lipidomic or metabolomic data could be applied to toxicity prediction.
- **The goal of this study was to develop a framework by which complex transcriptomic datasets can be transformed to a structure that may be combined with available data on animal testing to apply machine learning-based toxicity prediction.**
- **In addition, we set out to test the hypothesis that ToxML models can identify gene expression "fingerprints" of toxicity within the dataset of global gene expression changes in response to exposure to toxicants.**

2 Approach



- Collect in vivo and transcriptomic datasets from publicly available resources: ToxRefDB and NCBI GEO, respectively.
- Using R Statistical Program
- Convert LOAEL toxicity data to binary datasets
- Impute missing data across a chemical based on activity
- Join toxicity data to fold change in gene expression
- Statistical approaches include:
  - K Nearest Neighbors
  - Bernoulli NB
  - Linear SVC
  - Logistic Regression
  - Decision Tree
  - Random Forest
- Optimized on F1
- Feature selection
- Outcome weighting
- Models built using Python
- Identifies the best performing of multiple machine learning algorithms
- Builds predictive binary outcome toxicity screening models
- Measures performance based on 5-fold cross-validation
- Ascertain predictive frequency of models
- Evaluate frequency of predictive genes across animal toxicity endpoints

3 Datasets

- Transcriptome data obtained from NCBI GEO, accession GSE59927 (deposited by NIEHS).
- Rat tissues tested: bone marrow, brain, heart, intestine kidney, liver, spleen, and muscle. Rat hepatocytes were also tested.
- Gene expression changes were measured using GE Healthcare/Amersham Biosciences CodeLink™ UniSet Rat I Bioarray (platforms included GPL5424, GPL5425, and GPL5426). Each platform harbors over 10,000 gene probes.
- This dataset includes gene expression changes in rats exposed to various toxicants including Acetaminophen, Aspirin, Busulfan, Diethylstilbestrol, Progesterone, Methotrexate, Clofibrate, N,N-Dimethylformamide, Mifepristone, Theophylline, Coumarin, 4-Nonylphenol, 1,4-Dichlorobenzene, 4-Chloroaniline, Diethanolamine, 4-Octylphenol, Chlorpyrifos, Amitraz, Dichlorvos, Abamectin, Troglitazone, Nitrobenzene, 4-Nitrotoluene.
- CASRN numbers were used to crosswalk gene expression data with binarized animal toxicity data for 1045 endpoints obtained from ToxRefDB (October 2014 release); ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data
- For this preliminary analysis, the data crosswalk was joined by chemical (dose and exposure duration were not considered in crosswalk). This preliminary approach enabled analysis of dose- and exposure duration-independent gene expression features with predictive value in this model.

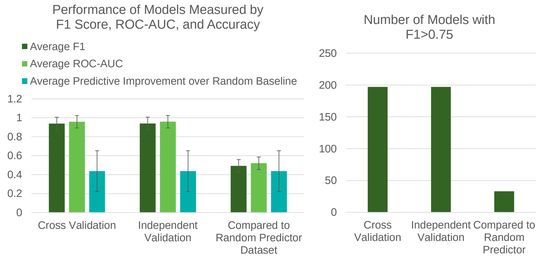
4 Model Validation

Cross Validation
Animal toxicity data was joined to gene expression data by chemical for each array platform. Each row represented a chemical at various doses and exposure durations. Binarized LOAELs from ToxRefDB and fold change of each gene were joined to toxicity data in columns. The model performance results are obtained by means of five fold cross validation.

Independent Validation
Results were derived by fitting the model on one dataset (e.g., 5424 platform crosswalk) to predict toxicity based on differential gene expression in the other datasets (e.g., crosswalk using 5425 platform dataset). This approach provides a more stringent analysis of the toxicity prediction models relative to the cross validation approach. Delta F1 values were <0.025 for 444 toxicity endpoints.

Validation against Random Predictors
In this approach, a random number vector the same size as the predictor transcriptomics dataset **for the cross-validation approach** was used to build a model on the 5424 dataset.

5 Model Performance



- Models successfully predict toxicity based on gene-expression values
- F1 and AUC scores demonstrate the model successfully predicts animal toxicity compared to a random predictor dataset.

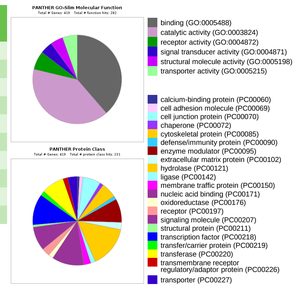
6 Genes Highly Correlated with Toxicity

Table 1. Top 10 Differentially Expressed Genes Correlating with Toxicity* (T Test > 0.95)

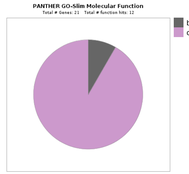
Orm1	orosomucoid 1
Frag1	FGF receptor activating protein 1
Fndc1	fibronectin type III domain containing 1
Ndrg1	N-myc downstream regulated gene 1
Klhdc8a	kelch domain containing 8A
Ldhb	lactate dehydrogenase B
Dusp5	dual specificity phosphatase 5
Ik	IK cytokine
Col1a1	procollagen, type 1, alpha 1
Tpmt	thiopurine methyltransferase

* Excludes ESTs, hypothetical, and predicted genes

T-Tests for each gene for each outcome if the mean gene expression values were different across outcomes (0/1, toxic/non-toxic). 7673 genes had T values >0.5. These genes were examined for Molecular Functions and Protein Classes using Panther Classification System http://www.pantherdb.org/geneListAnalysis.do



7 Genes Most Frequently Identified in Multiple Models



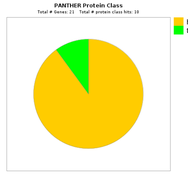


Table 2. Most Frequently Identified Genes Represented in Multiple Models

DNA (cytosine-5-)-methyltransferase 1
nerve growth factor receptor (TNFRSF16) associated protein 1
uromodulin
paraoxonase 1
ubiquinol cytochrome c reductase core protein 2
BCL2-associated transcription factor 1
calbindin 1
oligodendrocyte transcription factor 1
Homeobox protein Hox-D8 (Hox-4.3) (Hox-5.4) (DBSS)
defensin beta 1
histamine N-methyltransferase
solute carrier family 22 (organic cation transporter), member 2
transmembrane protein 8 (five membrane-spanning domains) (DBSS)
phospholipase D2
cubilin (intrinsic factor-cobalamin receptor)
klotho (Mm.) (DBSS)
aquaporin 4
serine dehydratase
apolipoprotein A-II
SH3 and cysteine rich domain 3 (DBSS)
hydroxysteroid 11-beta dehydrogenase 2

Frequency Distribution of the Most Influential Genes
60 genes were identified that correlated with toxicity in multiple datasets. A subset is shown in Table 2. Many genes represented at high frequency are known to be associated with toxicity to chemical exposure. In addition, many interesting genes were identified that may shed new light on toxicological pathways.

These 60 genes were also analyzed using PANTHER Classification System. Molecular functions including binding and catalytic activity, and protein classes including hydrolases and carrier proteins were highly represented in the gene set. The PANTHER Overrepresentation Test also identified fatty acid metabolism and lipid metabolism as processes highly represented by this gene set.

Conclusions

We successfully developed models to predict animal toxicity data using transcriptomic datasets

- These models performed extremely well in validation tests using multiple methods
- When applying a stringent independent validation test, the model correctly correlated changes in gene expression to 444 animal toxicity endpoints.

This approach also enabled identification of differentially expressed genes that correlate with toxicity

- Preliminary findings suggest this approach may be useful in identifying chemical-specific gene expression fingerprints that could be applied to developing novel toxicity prediction tools
- Identification of expected (Paraoxonase 1) and unexpected (Orosomucoid 1) genes associated with toxicity may shed new light on pathways of toxicity

Future Directions

1. Increase the size of the training and test transcriptomic datasets
2. Test whether model performance can be enhanced by incorporation of artificial intelligence computational methods
3. Conduct analysis of effects of dose, route, and exposure duration on models
4. Ascertain model performance using transcriptome data from different cell and tissue types (e.g., cell culture, PBMC)
5. Expand models to incorporate lipidomic, proteomic, and metabolomic datasets

Dream big. **Then call ICF.**

Contact: Joanne Trgovcich Joanne.Trgovcich@icf.com