Development of QSAR Models to Predict Systemic Toxicity Points of Departure

Prachi Pradeep[1,2] and Richard Judson[2]

[1]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee
[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Prachi Pradeep | pradeep.prachi@epa.gov | ORCID iD: 0000-0002-9219-4249 | Phone: 919-541-5150

www.epa.gov

INTRODUCTION

Human health risk assessment associated with environmental chemical exposure is limited by the tens of thousands of chemicals little or no experimental in vivo toxicity data. Data gap filling techniques, such as quantitative structure activity relationship (QSAR) models based on chemical structure information, are commonly used to predict hazard in the absence of experimental data. This study presents a set of QSAR models developed using chemical structural and physicochemical properties for chronic or sub-chronic in vivo points of departure (POD, the point on the dose-response that marks the beginning of a low-dose extrapolation). The in vivo data is taken from the EPA's ToxValDB, a compilation of information on ~3000 unique chemicals from a variety of public data sources. These models will inform chemical screening and prioritization efforts.

DATA PREPARATION

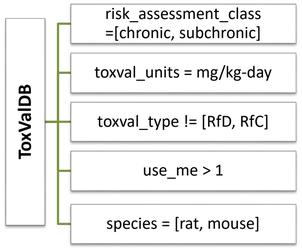


Figure 1: Schematic of data selection for modeling purposes from the ToxValDB. The final dataset for rat data is N = 1691 and mouse data is N = 668.

MOLECULAR FEATURES

- PubChem fingerprints (881 bits)
- Chemistry development kit (CDK) descriptors (18)

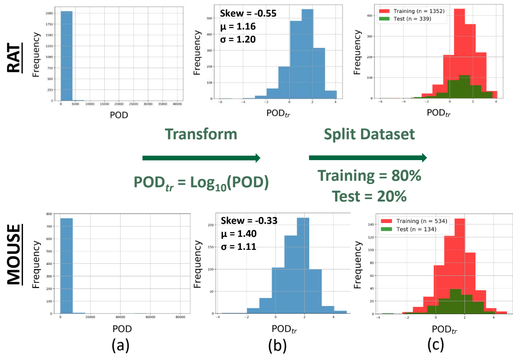


Figure 2: The POD values were log-transformed for both rat and mouse datasets. (a) Histogram of untransformed POD data, (b) Histogram of transformed POD (POD_{tr}) data, and (c) Histogram of training and test data relative to each other.

CHALLENGES

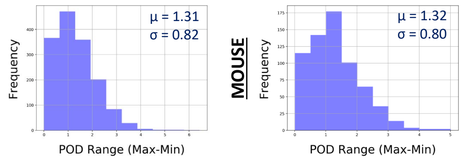
1. Experimental Variability

- Data from different labs (sources) running the "same" experiment may get different answers
- Sources of variability: Species, strain, dose range, dose spacing, length of study etc.

2. Model Uncertainty

- A model gives a result (a POD), but this is an estimate of the "true" POD. We are uncertain about what the true POD is
- Uncertainty in the evaluation data will lead to uncertainty in the model and our estimate of its quality

Figure 3: Distribution of the range of POD values for the rat and mouse dataset. Variability in experimental data leads to uncertainty in the model predictions. Roughly, the root mean squared error (RMSE) in the models can be estimated to be around 1.15 (√1.31) for rat and 1.15 (√1.32) for mouse models.



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

TYPES of MODELS

Given the variability and skewness in the training dataset, 3 types of models were developed:

1. Point-estimate models

- Regression models were built to predict a POD value for each chemical
- Experimental POD = Minimum POD value

2. Point-estimate, balanced-dataset models

- The training data was re-constructed to reduce skewness by adding 10% duplicate data from the long tail
- Regression models were built to predict a POD value for each chemical using the re-constructed data
- Experimental POD = Minimum POD value

3. Point-estimate with confidence interval models

- A POD distribution was constructed for each chemical using mean = experimental POD value (= Median POD value) and standard deviation = 0.5 log-units, based on the typical lab to lab variability
- Bootstrap models were built with random sampling of POD values for each chemical from the pre-generated POD distribution

METHODS

FEATURE SELECTION

Fingerprints
1. Unsupervised 80% variance threshold
2. Collinearity 80% threshold
Rat: 881 -> 39 bits
Mouse: 881 -> 44 bits

Descriptors
1. Supervised recursive feature elimination using linear regression
18 -> 5 descriptors

ALGORITHM

1. k nearest neighbors (kNN)
2. Support vector regression (SVR)
3. Random forest (RF)
4. Gradient boosting regression (GBR)

HYPER-PARAMETER TUNING

5-fold GridSearchCV

1. kNN: k, weights, algorithm
2. SVR: Epsilon, C, gamma, kernel
3. RF: Max features, N trees
4. GBR: N trees, max depth, loss function, learning rate

MODEL VALIDATION

Internal validation
5-fold internal CV on 80% training set

External validation
20% test set

RESULTS

1. Point-estimate models

1. Point-estimate models	Random Forest Hyper-parameters	5-fold Internal Cross-validation			External Validation		
		RMSE	RMSE/σ	R^2	RMSE	RMSE/σ	R^2
Rat	Max features = sqrt Number of trees = 1000	0.98	0.84	0.29	1.13	0.76	0.25
Mouse	Max features = sqrt Number of trees = 1000	0.97	0.88	0.23	0.97	0.83	0.30

Table 1: Performance metrics for the point-estimate models for 5-fold internal cross validation and external validation.

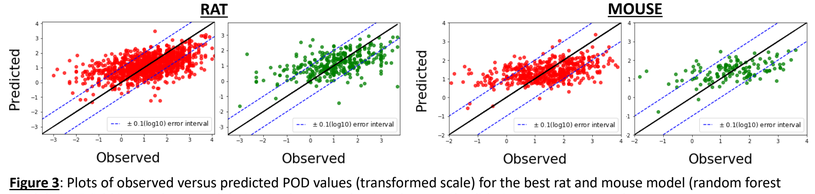
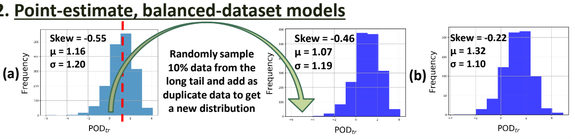


Figure 3: Plots of observed versus predicted POD values (transformed scale) for the best rat and mouse model (random forest model) for 5-fold internal cross validation (red scatter plots) and external validation (green scatter plots).

RESULTS

2. Point-estimate, balanced-dataset models

Figure 5: Sample histogram of re-constructed data for (a) rat and (b) mouse. Total number of samples in the re-constructed training dataset for rat data is N = 1860 and mouse data is N = 734.



2. Point-estimate, balanced-dataset models	Random Forest Hyper-parameters	5-fold Internal Cross-validation			External Validation		
		RMSE	RMSE/σ	R^2	RMSE	RMSE/σ	R^2
Rat	Max features = sqrt Number of trees = 1000	0.98	0.84	0.29	1.13	0.76	0.25
Mouse	Max features = sqrt Number of trees = 1000	0.93	0.84	0.29	0.86	0.79	0.39

Table 2: Performance metrics for the point-estimate balanced-dataset models for 5-fold internal cross validation and external validation.

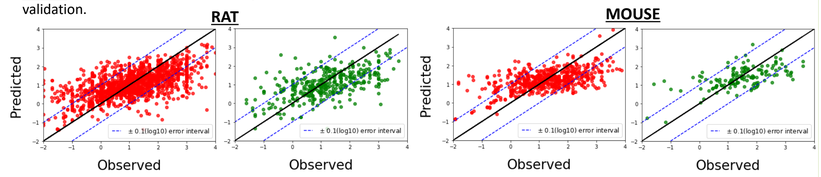


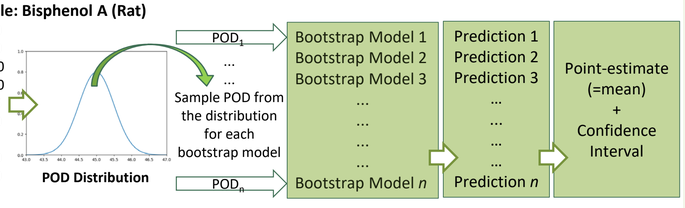
Figure 3: Plots of observed versus predicted POD values (transformed scale) for the best rat and mouse model (random forest model) for 5-fold internal cross validation (red scatter plots) and external validation (green scatter plots).

3. Point-estimate with confidence interval models

Illustrative example: Bisphenol A (Rat)

DTXCID30182 40.0
DTXCID30182 40.0
DTXCID30182 200.0
DTXCID30182 200.0
DTXCID30182 37.5
DTXCID30182 50.0
DTXCID30182 5.0
DTXCID30182 50.0
DTXCID30182 50.0
DTXCID30182 5.0

Median POD = 45



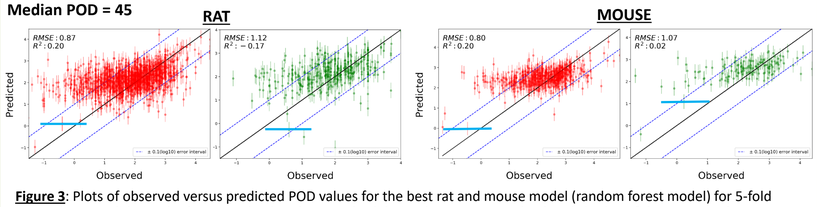


Figure 3: Plots of observed versus predicted POD values for the best rat and mouse model (random forest model) for 5-fold internal cross validation (red scatter plots) and external validation (green scatter plots) with 95% confidence intervals.

CONCLUSIONS

Summary: Three types of rat and mouse predictive models were developed using chemical structural and physicochemical properties for chronic and sub-chronic in vivo points of departure.

Future work: (1) Add study parameters as model descriptors to account for additional lab-to-lab data variability (Species, strain, study duration, exposure route), and (2) Re-construct datasets to reduce sampling by sampling equally from both tails (and not duplicate data).