ToxCast Revisited: A Comprehensive Statistical Analysis of the In Vitro-to-In Vivo Predictive Capacity of High-Throughput Toxicity Screens

Russ Wolfinger
Director of Scientific Discovery and Genomics
SAS Institute Inc., Cary, NC

SRP Meeting
October 24, 2012
A Comprehensive Statistical Analysis of Predicting In Vivo Hazard Using High-Throughput In Vitro Screening

Russell S. Thomas,*,† Michael B. Black,* Lili Li,† Eric Healy,* Tzu-Ming Chu,† Wenjun Bao,† Melvin E. Andersen,* and Russell D. Wolfinger†

*The Hamner Institutes for Health Sciences, Research Triangle Park, North Carolina 27709; and †SAS Institute Inc., Cary, North Carolina 27513

To whom correspondence should be addressed at The Hamner Institutes for Health Sciences, 6 Davis Drive, Research Triangle Park, NC 27709. Fax: (919) 558-1300. E-mail: rthomas@thehamner.org.

Received February 24, 2012; accepted April 23, 2012

Over the past 5 years, increased attention has been focused on using high-throughput in vitro screening for identifying chemical hazards and prioritizing chemicals for additional in vivo testing. The U.S. Environmental Protection Agency’s ToxCast program has generated a significant amount of high-throughput screening data allowing a broad-based assessment of the utility of these assays for predicting in vivo responses. In this study, a comprehensive cross-validation model comparison was performed to evaluate the predictive performance of the more than 600 in vitro assays from the ToxCast phase I screening effort across 60 in vivo endpoints using 84 different statistical classification methods. The predictive performance of the in vitro assays was compared and combined hazards. In the United States, the Environmental Protection Agency’s (EPA) ToxCast project and the Tox21 consortium have used high-throughput screening to characterize the in vitro biological activity of chemicals across multiple cellular pathways and biochemical targets with the intent to prioritize compounds for conventional toxicity testing (Dix et al., 2007; Kavlock et al., 2009). In Europe, multiple research consortia such as Sens-it-iv, ACuteTox, and carcinoGENOMICS have also examined the ability of in vitro systems for predicting in vivo responses (dos Santos et al., 2009; Jennen et al., 2010; Sjostrom et al., 2008). To date, the research efforts have produced mixed results.
ToxCast Data from the EPA

~200 Cellular Assays

Activity of the Chemical Based on Concentration in the Well

Terminal
Node 1
N = 3

A_51_P161890 <= 0.01

Terminal
Node 2
N = 6

A_51_P108645 <= -0.80

Node 1
A_51_P108645 <= -0.80

A_51_P108645 > -0.80

Node 2
A_51_P161890 <= 0.01

Predictive Combinations of Assays

In Vivo Hazard Prediction and Prioritization

In Vitro High Throughput Screens

~300 Biochemical Assays

~300 Cellular Assays

309 Phase I Chemicals
(Pesticides/HPV)

In Vivo Endpoints

9 Phase I Chemicals
(Pesticides/HPV)
Currently Published Work on Predictive Toxicity Signatures in ToxCast

- Reproductive toxicity signature
- 74% Balanced Accuracy
- Pre-filtered assays and lumped subset into 6 classes based on genes and functional grouping
- Only study with external validation set

- Rat liver tumor signature
- No formal classification statistical analysis (cross-validation)

- Developmental toxicity signature
- 71% Balanced Accuracy
- Pre-filtered assays and aggregated assays based on genes and GO categories

- Vascular development signature
- 80% Accuracy
Why is a Software Company Getting Involved with ToxCast?

• Our life sciences team has collaborated with the EPA, Hamner Institute, NIEHS, UNC, and NC State for many years, and has its roots in toxicology-based microarray data analysis.

• The ToxCast data is highly valuable and presents numerous analytical challenges, several of which JMP Genomics software can help address.

• We wanted to test and stretch the software in new directions and participate in the project by providing, as much as possible, a “neutral third party” assessment of the predictive performance of the assays.

• Previous work in collaboration with Fred Wright at UNC; current work in collaboration with Rusty Thomas at Hamner.
1,224 Chemical Structure Descriptors
>600 In Vitro High Throughput Screens
60 In Vivo Endpoints from ToxRef Database

5-Fold Cross Validation

8 Classification Algorithms,
~12 Feature Selection Approaches, 84 Classification Model Combinations

No Aggregation
Aggregate Based on Genes
Aggregate Based on GO Category
No Pre Filter
Pre Filter

Model #1
... Model #84

Partition Data into 5 Equal Sets

Set 1 Set Aside
Repeat 5X
Build Model
Select Features
Predict Hold Out Set

Repeat 10X

Range and Central Tendency of In Vivo Predictive Performance Regardless of Statistical Model

Aggregate Based on GO Category
Aggregate Based on Genes
No Pre Filter
Pre Filter
Workflow Details

- 60 Endpoints x 84 Models x 10 Iterations x 5 Folds = 252,000 separate model fits; computationally intensive.

- Predictive models include discriminant, distance scoring, k-nearest neighbors, logistic regression, general linear model, partial least squares, partition trees, and radial basis machine.

- Various Sets of Predictor Variables:
  1. *In Vitro* Assays, optionally aggregated
  2. Chemical Structural Descriptors, computed by Dragon

- All endpoints are binary, so performance criteria include: Accuracy, Balanced Accuracy, Sensitivity, Specificity, Negative Predictive Value (NPV), Positive Predictive Value (PPV), and Area Under the Curve (AUC)
Predictive Performance Criteria for Binary Endpoints

### Table: In Vivo Animal Response

<table>
<thead>
<tr>
<th>Prediction Based on In Vitro Assays or Chemical Structure</th>
<th>In Vivo Animal Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Positive</td>
<td>TP</td>
</tr>
<tr>
<td>Positive Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

- **Sensitivity** = TP / (TP + FN)
- **Specificity** = TN / (FP + TN)
- **PPV** = TP / (TP + FP)
- **NPV** = TN / (FN + TN)

**Accuracy** = (TP + TN) / (TP + FP + FN + TN)

**Balanced Accuracy** = (Sensitivity + Specificity) / 2, adjusts for prevalence

**RMSE** = square root (average (true value − predicted probability)²)

**AUC** is area under the Receiver Operating Characteristic (ROC) curve; it measures sorting efficiency and is directly related to the Mann-Whitney rank-sum statistic; a value of 1 indicates perfect sorting.
Prevalence of Positive Chemicals Among Endpoints Should be Kept in Mind
In Vitro Assays

Chemical Structure
Chronic Mouse
Developmental Rabbit
Chronic Rat
Multi-generational Rat
Developmental Rat
Fig. 12
Predictability

Mean \(\text{(Log}_2(1/\text{RMSE})\)

Positive Prediction

Negative Prediction
Mean \( \log_2(1/RMSE) \)

Positive Prediction

Negative Prediction
Learning Curves

• An appropriate way to assess adequacy of sample size in a predictive modeling context

• Completely different from classic power and sample size calculations, which are designed for hypothesis testing, not predictive modeling

• Constructed by taking subsets of different sizes and performing cross-validation model comparison on each, then plot results with sample size as the x-axis

• An upward trend indicates that results would likely improve with more samples; a flat curve indicates otherwise.
Best model from CVMC = PT_075 (AUC = 0.75583)
Worst model from CVMC = KNN_036 (AUC = 0.61631)
MGR_Rat_ReproductiveOutcome
Best model from CVMC = GLM_025 (AUC = 0.54170)
Worst model from CVMC = PLS_057 (AUC = 0.49296)
MGR_Rat_Testis

Best model from CVMC = PT_074_025 (AUC = 0.50620)
Worst model from CVMC = GLM_020 (AUC = 0.44754)
Summary

- The current ToxCast *in vitro* high-throughput screening assays provide somewhat limited ability to predict *in vivo* toxic responses.

- Sensitivity and specificity of the *in vitro* assays is related to the balance of positive and negative chemicals, but even for balanced endpoints, the overall predictive performance is relatively low.

- The *in vitro* assays provide somewhat lower predictive performance than chemical structure.

- Aggregating the assays based on gene and biological processes did not appear to improve predictive performance.

- Pre-filtering the in vitro assay data, as has been done in previous studies, can significantly bias estimates of cross-validation performance in an optimistic direction.

- Analysis of predictability and learning curves can help in prioritizing chemicals for further study and in assessing adequacy of sample sizes.