**Data and text mining**

**tcpl: the ToxCast pipeline for high-throughput screening data**

Dayne L. Filer¹, Parth Kothiya¹, R. Woodrow Setzer², Richard S. Judson² and Matthew T. Martin²,*

¹Oak Ridge Institute for Science Education Fellow at National Center for Computational Toxicology and ²National Center for Computational Toxicology, US EPA, Research Triangle Park, NC, USA

*To whom correspondence should be addressed.

Abstract

**Motivation:** Large high-throughput screening (HTS) efforts are widely used in drug development and chemical toxicity screening. Wide use and integration of these data can benefit from an efficient, transparent and reproducible data pipeline. Summary: The tcpl R package and its associated MySQL database provide a generalized platform for efficiently storing, normalizing and dose–response modeling of large high-throughput and high-content chemical screening data. The novel dose-response modeling algorithm has been tested against millions of diverse dose-response series, and robustly fits data with outliers and cytotoxicity-related signal loss.

**Availability and Implementation:** tcpl is freely available on the Comprehensive R Archive Network under the GPL-2 license.

**Contact:** martin.matt@epa.gov

1 Introduction

Advances in molecular biology and robotics have propelled high-throughput and high-content screening techniques into the forefront of drug discovery and toxicology. US NIH-funded efforts such as the Molecular Libraries Screening Initiative (Austin et al., 2004) and the US EPA ToxCast (Dix et al., 2007) and cross-agency Tox21 (Collins et al., 2008; Kavlock et al., 2012) efforts are making publicly available in vitro screening data on tens to hundreds of thousands of chemicals and thousands of unique assays. In particular, the ToxCast program includes over 1600 different cell-based and biochemical assay endpoints with heterogeneous producers, experimental designs, readouts and input file formats, spanning up to 8200 unique chemicals.

Successful implementation of any high-throughput screening (HTS) initiative requires substantial data science efforts to store and process the large quantities of data generated. The diversity of the ToxCast data necessitated the development of a flexible HTS data pipeline capable of efficiently processing and storing large volumes of heterogeneous data. Here we introduce the ToxCast pipeline (tcpl), a novel R extension (R Core Team, 2016), to provide storage, normalization, dose–response modeling and visualization solutions for HTS screening efforts. tcpl provides functionality to identify potentially active compounds (positive hit-calls) for both single- and multiple-concentration testing paradigms. Multiple-concentration processing also includes dose-response modeling to give potency and efficacy estimates, and categorization of each concentration series to better identify potential false positive and false negative results.

Other efforts have similarly worked towards providing HTS data solutions, but are limited by some combination of lacking documentation, expensive licensing, input format, scalability, operating system and reproducibility (Fourches et al. (2014); Frommolt and Thomas (2008); SigmaPlot, https://systatsoftware.com/; Genedata Screener, https://www.genedata.com/products/screener/). The open-source tcpl package improves on previous efforts in ways including addressing generalizability of curve-fitting, the ability to handle heterogeneous data, incorporation of chemical-tracking information, and comprehensive database support. Most significantly, the tcpl package provides functionality for dose-response modeling that draws from the Student’s* t-distribution, facilitating robust fitting that eliminates the need for outlier detection and removal. In addition to the novel fitting...
algorithm, tcpl provides a categorization algorithm that segregates results into categories for rapid identification of potential false-positive and false-negative results.

The constrained fitting algorithm has been validated against millions of dose-response curves through manual curation both internally and externally. Additionally, the tcpl package has undergone contracted code review. The strength and visibility of the package has already led to it being used outside the EPA in all three sectors.

### 2 Implementation

The tcpl package provides functionality for two screening paradigms: (i) single-concentration screening, intended to only identify potentially active compounds and (ii) multiple-concentration screening intended to identify potentially active compounds and estimate the efficacy and potency through dose-response modeling. Although data processing differs, the tcpl database stores the same raw information (level 0) for each screening paradigm. In addition to the brief summaries provided below, detailed information about all processing steps is included in the package vignette.

Single-concentration processing includes two processing levels. Level 1 processing normalizes the data, either to fold-induction or percent-of-control units, and converts the concentrations to a logarithm scale. Level 2 processing collapses the replicates by taking the median for each sample-assay pair. Using an activity cutoff defined by the user, level 2 processing also determines the activity call for each sample-assay pair.

Multiple-concentration processing includes six processing levels. Briefly, level 1 processing defines concentration and replicate indices, giving integer values 1 … N to increasing concentrations and technical replicates, where 1 represents the lowest concentration or first technical replicate. Level 2 processing allows for basic transformations of the raw data, e.g. logarithmic conversion, and removes data deemed poor quality by the user. Similar to level 1 in single-concentration processing, level 3 normalizes data to fold-change or percent-of-control and converts concentrations to a logarithm scale. At level 4 data are modeled (described below), before level 5 processing defines the winning model and the activity call. Level 5 processing also separates each data series into categories to facilitate easy triaging of the results. Level 6 processing identifies potential false positive and false negative results, giving problematic data series a flag.

The tcpl package heavily utilizes the popular ‘data.table’ (Dowle et al., 2015) and ‘parallel’ R packages to address performance when working with millions of samples.

To address both data storage and reproducibility issues the tcpl package interacts with a MySQL relational database (included in the package distribution) to store all data and processing decisions made by the user. In addition to storing the data at every level of processing, the accompanying database stores chemical and assay annotations to facilitate further analyses and disseminating results.

### 3 Dose-response modeling

We designed the dose-response modeling procedure to handle the outliers and cytotoxicity-related loss of signal common to HTS efforts. The modeling methodology evaluates each data series with three models: (i) a constant model at zero, (ii) a constrained three-parameter Hill model and (iii) a constrained five-parameter gain-loss model. To facilitate simple cross-experiment comparisons and reduce the parameter space, dose-response modeling is constrained to a zero-centered, positive response paradigm. Therefore, negative response data requires an inverse transformation during the normalization process. To obtain robust results without removing any data, we define the log-likelihood using a Student’s t-distribution with 4 degrees of freedom (Lange et al., 1989).

Let $t(z; ν)$ be the Student’s $t$-distribution with ν degrees of freedom, $y_i$ be the observed response at the $i$th observation, and $μ_i$ be the estimated response at the $i$th observation. We calculate $z_i$ as

$$z_i = \frac{y_i - μ_i}{\exp(σ)}$$

(1)

where $σ$ is the natural logarithm of the scale term. Then the log-likelihood is

$$\sum_{i=1}^{n} \left[ \ln(t(z_i; 4)) - σ \right],$$

(2)

where $n$ is the number of observations.

Figure 1 gives an illustration of the three models, with their respective $μ$. The modeling methodology uses a one dimensional optimization to maximize the constant model, and the Nelder-Mead algorithm to maximize the non-linear models. The tcpl vignette provides information about the constraints on the Hill and gain-loss models. Utilizing the Student’s t-distribution and the gain-loss model provide robust solutions to the problems of outliers and the cytotoxicity-related signal loss that commonly occurs at high concentrations. After fitting each model, the model with the lowest Akaike information criterion value is selected as the ‘winning’ model.

### 4 Modeling demonstration and comparison

To illustrate the robust performance of the dose–response modeling algorithm in the presence of outliers, Figure 2 shows a comparison of the results provided by tcpl versus a traditional non-linear least-squares (NLLS) approach. To mimic typical HTS data, we sampled data from a normal distribution around a Hill model where the standard deviation was 10% of the asymptotic top value. Each dataset contained seven concentrations and three replicates with an AC_{50} value of 5 and an asymptotic top value of 100. Then we selected a random data point and changed the value to ± 100% of the asymptotic top value (0 and 200) to simulate an outlier.

For each outlier direction we generated ten thousand simulated datasets that were modeled using both the tcpl package and the drc package (Ritz et al., 2015), with a three-parameter logistic model (with the bottom asymptote set to 0 to match the tcpl algorithm). For the negative outlier direction, the mean and standard deviation of the AC_{50} values for tcpl and NLLS were 5.13 ± 0.37 and 5.12 ± 0.54, respectively. On average, both methods accurately estimated the AC_{50} value, but the NLLS method had significantly greater variance.
concentrations, the average increase in the AC50 value of 2.2 using the
POD estimates are calculated when the constant model is the selected
sidered active. The four POD estimates are illustrated in Figure 3. No
first reaches the user-defined cutoff value for a data-series to be con-
3
and the ACB is the concentration at which the model first reaches
5 Hit-calling and point-of-departure estimates
A dose-response series must meet two criteria to have an active hit-
call: (i) the Hill or Gain-Loss model must win and (ii) both the mod-
eled and observed maximum responses must meet an efficacy cutoff
based on an expandable list of methods assigned by the user.
In addition to the standard AC50 (activity concentration at 50%
maximal activity) provided for the Hill and Gain-Loss models, the
tcpl package provides three point-of-departure (POD) estimates for
the winning model. The AC50, or activity concentration at 10%, is
derived solely from the model parameters. Conversely, the ACB (ac-
activity concentration at baseline) and ACC (activity concentration at
cutoff) are based on levels of noise and significance, respectively.
The package estimates the noise of an assay by calculating the median
absolute deviation over all response values given by the first two con-
centrations (bmad). The baseline region is then defined as 0±3bmad,
and the ACB is the concentration at which the model first reaches
3bmad. Similarly, the ACC is the concentration at which the model
first reaches the user-defined cutoff value for a data-series to be con-
sidered active. The four POD estimates are illustrated in Figure 3. No
POD estimates are calculated when the constant model is the selected
model winner, because the POD estimates do not apply.

6 Conclusions
The tcpl package provides a resource for processing HTS data that is:
1. easily implemented through open source R and MySQL software
platforms, making it easily usable for researchers within the
HTS field;
2. generalized to encompass diverse inputs and experimental de-
signs, including numerous normalization methods;
3. transparent and reproducible with open source code and a
database-driven system to store every processing decision
throughout the pipeline;
4. proven through significant testing, including use on over 1600
assay endpoints, generating >2 million individual curves and in-
dependent use by multiple groups in the US and Europe;
5. freely available on the Comprehensive R Archive Network.

Acknowledgements
We thank the NCCT scientists for their testing of the tcpl package. We also
thank Agnes Karmaus and Kenneth Ball for their review of this manuscript.
Conflict of Interest: none declared.
References
1138–1139.
Science, 319, 906–907.
Dix,D.J. et al. (2007) The ToxCast program for prioritizing toxicity testing of
environmental chemicals. Toxicol. Sci., 95, 5–12.
Dowle,M. et al. with contributions from R Saporra, S. L. and Antonyan,
1.9.6.
Foucres,D. et al. (2014) Hts navigator: freely accessible cheminformatics
software for analyzing high-throughput screening data. Bioinformatics, 30,
588–589.
Frommolt,P. and Thomas,R.K. (2008) Standardized high-throughput evalu-
Kavlock,R. et al. (2012) Update on epa’s toxcast program: providing high
throughput decision support tools for chemical risk management. Chem.
Lange,K.L. et al. (1989) Robust statistical modeling using the t distribution.
R Core Team (2016) R: A Language and Environment for Statistical
e0146021.