US EPA TOXCAST TOX21 DATA RELEASE OCTOBER 2014

This file describes the contents of the October 2014 ToxCast Tox21 data release. The zip file contains the following summary-level file, not including this README file:

[1] "ToxCast_Tox21_Level5&6_20141022.csv"

In addition to the above listed file, the ToxCast program also released a MySQL dump file containing all data and a beta version of the R package (tcpl) that interacts with the MySQL database used to process all of the data for this release. For information/data not included in the listed summary files, users will need to download and interact with the MySQL database. We also encourage the database users to utilize the 'tcpl' R package containing numerous queries and functionality for easily loading and visualizing the data. At the bottom of this file is an R script to produce all of the listed files, utilizing the MySQL database and 'tcpl' R package.

All information in the summary file is reported at the sample level (specifically the Tox21 sample id) and only contains the Tox21 (NCATS assay library). Each row in this file contains a unique combination of sample and assay endpoint with all of the model information applied to the underlying concentration response data.

All columns starting with 'cnst' refer to parameters from the constant model.
All columns starting with 'hill' refer to parameters from the hill model.
All columns starting with 'gnls' refer to parameters from the gain-loss model (the product of 2 hill models, one with a negative hill slope, and that share a top/upper-asymptote).
For more information on the specifics of the modeling process please refer to the data analysis R package documentation.
All columns starting with 'modl' refer to the winning models parameters. Below is the list of the 75 columns exported for this dataset from Level 4 (modeling), 5 (model selection & hit calling) and 6 (flagging) data processing.

[1] l5id = unique id level 5 processing
[2] spid = sample id (id blindly provided to vendors)
[3] chid = chemical id (DSSTox GSID) 1:1 with casrn
[4] casn = CAS Registry number
[5] chnm = chemical name
[6] code = CAS Registry number (excel protected)
[7] aeid = assay endpoint id (unique id)
[8] aenm = assay endpoint name
[9] l4id = unique id for level 4 processing
[10] bmad = baseline median absolute deviation (noise around baseline)
[11] resp_max = maximal single replicate response
[12] resp_min = minimal single replicate response

[13] max_mean = maximal mean response at a given concentration
[14] max_mean_conc = corresponding concentration of max_mean
[15] max_med = maximal median response at a given concentration
[16] max_med_conc = corresponding concentration of max_med
[17] logc_max = maximum tested log concentration (log uM)
[18] logc_min = minimum tested log concentration (log uM)
[19] cnst = constant model successfully run (1 or 0)
[20] hill = hill model successfully run (1 or 0)
[21] hcov = hill model covariance
[22] gnls = gain-loss model successfully run (1 or 0)
[23] gcov = gain-loss model covariance
[24] cnst_er = constant model error term
[25] cnst_aic = constant model AIC (used to select winning model)
[26] cnst_rmse = constant model RMSE
[27] cnst_prob = constant model probability (based on AIC)
[28] hill_tp = hill model top of curve
[29] hill_tp_sd = hill model top standard deviation
[30] hill_ga = hill model logAC50 (gain logAC50)
[31] hill_ga_sd = hill model AC50 standard deviation
[32] hill_gw = hill model slope
[33] hill_gw_sd = hill model slope standard deviation
[34] hill_er = hill model error term
[35] hill_er_sd = hill model error standard deviation
[36] hill_aic = hill model AIC (used to select winning model)
[37] hill_rmse = hill model RMSE
[38] hill_prob = hill model probability (based on AIC)
[39] gnls_tp = gain-loss top of curve
[40] gnls_tp_sd = gain-loss top of curve standard deviation
[41] gnls_ga = gain-loss model gain logAC50
[42] gnls_ga_sd = gain-loss model gain logAC50 standard deviation
[43] gnls_gw = gain-loss model gain slope (positive)
[44] gnls_gw_sd = gain-loss model gain slope standard deviation
[45] gnls_la = gain-loss model loss logAC50
[46] gnls_la_sd = gain-loss model loss logAC50 standard deviation
[47] gnls_lw = gain-loss model loss slope (negative)
[48] gnls_lw_sd = gain-loss model loss slope standard deviation
[49] gnls_er = gain-loss model error term
[50] gnls_er_sd = gain-loss model error standard deviation
[51] gnls_aic = gain-loss model AIC (used to select model winner)
[52] gnls_rmse = gain-loss model RMSE
[53] gnls_prob = gain-loss model probability (based on AIC)
[54] nconc = number of tested concentrations
[55] npts = number of data points
[56] nrep = number of replicates
[57] nmed_gtbl = number of median values greater than baseline
[58] hitc = hit call (based on ‘coff’ and winning model)
[59] modl = winning model
[60] fitc = fit category (defined by many parameters)
[61] coff = response cutoff (used to define hit-call)
[62] actp = activity probability (1-cnst_prob)
[63] modl_er = winning model error term
[64] modl_tp = winning model top of curve (where applicable)

[65] modl_ga = winning model gain logAC50 (where applicable)
[66] modl_gw = winning model gain slope (where applicable)
[67] modl_la = winning model loss logAC50 (where applicable)
[68] modl_lw = winning model loss slope (where applicable)
[69] modl_rmse = winning model RMSE
[70] modl_prob = winning model probability
[71] modl_acc = winning model log concentration at 'coff'
[72] modl_acb = winning model log concentration at 'bmad'
[73] resp_unit = response units
[74] flag_id = concatenated list of flag ids
[75] flag = concatenated list of flag names

The parameters for the winning model are given regardless of hit-calling; therefore, many inactive chemicals have a gain AC50 chemical in the "modl_ga" column, for example. The "hitc" column provides the activity call (1=active, 0=inactive, -1=unable to model)

For questions or concerns, please contact Monica Linnenbrink at: linnenbrink.monica@epa.gov.

```
######################################################################
## R Script to produce October 2014 ToxCast Tox21 Data Release
######################################################################

rm(list = ls())
library(tcpl)
library(reshape2)
library(data.table)

getwd()

aeids <-
tcplLoadAeid(c('assay_source_name','export_ready'),list('Tox21',1))$ae
id

dat5  <- tcplLoadData(lvl=5L, 'aeid', aeids)
dat5  <- tcplPrepOtpt(dat5)

dat6  <- tcplLoadData(lvl=6L, 'aeid', aeids)
agg6  <- dat6[ , list(flag_id = paste(l6_mthd_id, collapse='|'),
                      flag = paste(flag, collapse='|')),
               by = l5id]
setkey(dat5, l5id)
setkey(agg6, l5id)

dat56 <- merge(x = dat5, y = agg6, by = 'l5id', all.x = TRUE)
write.csv(dat56, "ToxCast_Tox21_Level5&6_20141022.csv")

######################################################################
## End R script
######################################################################
```