ACCOUNTING FOR UNCERTAINTY IN THE APPLICATION OF HIGH THROUGHPUT DATASETS

R. Woodrow Setzer, National Center for Computational Toxicology, US Environmental Protection Agency, Research Triangle Park, NC

The use of high throughput screening (HTS) datasets will need to adequately account for uncertainties in the data generation process and propagate these uncertainties through to ultimate use. Uncertainty arises at multiple levels in the construction of predictors using *in vitro* HTS data. Many HTS assays may have an inherent level of variability deriving from the nature of the biological processes used in the assay. In addition, measurement processes used to translate a biological effect *in vitro* into a quantified datapoint introduces variability as well. Finally, predictors are typically statistical or machine learning models, informed by biology mostly through fitting to training sets. The structure of predictors is not typically informed by biology in any deeply meaningful way. Quantifying the magnitude and nature of the resulting uncertainty is important both for the use of such predictors for prioritization and for the design of studies to generate more reliable predictors. This presentation will discuss how variability introduced by biological and measurement processes leads to uncertainties in the output from high throughput screening (HTS) assays. The ideas presented are illustrated for a set of HTS assays and target toxicity data used in EPA's ToxCast™ program. Simulation results inspired by the ToxCast™ data show how the uncertainty in HTS results is transformed into uncertainty about predictions using those results, and help to quantify the relative importance of variability at different levels of the data generation process to uncertainty of ToxCast™ predictions. *This work was reviewed by EPA and approved for publication but does not necessarily reflect official Agency policy*