LETTER TO EDITOR

Incorporating Biological, Chemical, and Toxicological Knowledge Into Predictive Models of Toxicity


Office of Research & Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711

Thomas et al. (2012) recently published an evaluation of statistical models for classifying in vivo toxicity endpoints from ToxRefDB (Knudsen et al., 2009; Martin et al., 2009a, b) using ToxCast in vitro bioactivity data (Judson et al., 2010) and chemical structure descriptors. We commend the authors for a thorough assessment of statistical tools for uncovering patterns of associations among thousands of covariate features derived from in vitro measurements, chemical structure, and toxicity endpoints from animal studies. They were largely unsuccessful in accurately classifying toxicities based on in vitro bioactivity or chemical structure. However, their conclusion that the current ToxCast phase I assays and chemicals have limited applicability for predicting in vivo chemical hazards using statistical classification methods is misleading and warrants clarification.

The approach of Thomas et al. (2012) is primarily a statistical path to producing classifiers that does not incorporate knowledge of biological or adverse outcome pathways to group assays or endpoints from the in vitro or in vivo data sets. Classification accuracy has two key ingredients: relevant assays or descriptors and a sufficient number of representative chemicals, positives and negatives, for the different types of toxicity endpoints being predicted. This is a classical statistical issue, and emphasizes the difficulty in finding robust statistical associations with relatively small number of samples for the many ToxRefDB apical toxicity endpoints. Hence, Thomas et al. (2012) conclude that the current ToxCast phase I assays and chemical library have limited applicability for predicting in vivo chemical hazards. The results described in Thomas et al. (2012) are not altogether surprising and are consistent with findings from other research groups, including our own. In a modeling study of ToxCast/ToxRefDB-like data (Judson et al., 2008), we demonstrated that although most statistical or machine-learning methods perform well in the presence of a small number of causal features, most show significant degradation in performance as irrelevant features are added—a well-known characteristic of such models (Almuallim and Dietterich, 1991). We also demonstrated in Judson et al. (2008) that all statistical and machine-learning methods performed better with feature selection. This earlier study with simulated data led us to a hypothesis-driven approach for feature selection and data aggregation incorporating biological, chemical, and toxicological knowledge into modeling from the ToxCast and ToxRefDB data. Using this hypothesis-driven approach, we have developed useful predictive models from the ToxCast phase I data.

In our predictive models from the ToxCast phase I data, we have incorporated knowledge of biological and adverse outcome pathways for assay selection and as a guide to produce biologically relevant and statistically significant classifiers (Judson et al., 2010, 2011; Kleinstreuer et al., 2011; Knudsen and Kleinstreuer, 2011; Martin et al., 2011; Reif et al., 2010; Sipes et al., 2011). For example, Martin et al. (2011) focused on the subset of phase I chemicals with high-quality rat multigeneration reproductive toxicity studies in ToxRefDB and sufficient in vitro bioactivity in ToxCast. Predictive signature development of reproductive toxicity required careful consideration of analysis and model development techniques because reproductive toxicity is an aggregated multimodal and multieffect outcome. No one in vitro assay had the ability to broadly identify reproductive toxicants, and computational modeling allowed us to explore the complex relationships between in vivo observations and networks of in vitro activity. The initial inputs into the model were hundreds of ToxCast assay results collectively mapped to genes (e.g., PPARA: a suite of peroxisome proliferator-activated receptor alpha) and gene families (e.g., GPCR: broad spectrum G protein-coupled receptor inhibition assays). The aggregate activity across the assays per gene or gene family provided the quantitative inputs.
into the model. The assay-gene combinations were further filtered based on a feature selection process that evaluated the statistical association to the training set data. The filtered gene set was then weighted in a multivariate model using linear discriminate analysis and evaluated for stability using fivefold cross-validation. The modeling process resulted in a final model with 80% balanced accuracy, consisting of eight features. Many other approaches and methods could have been deployed, but complex machine-learning algorithms have a tendency to overfit the data, thereby lowering the output model’s ability to be externally predictive. External validation of the model was accomplished with data excluded from the training set for this express purpose, data that were not used in developing the initial model. Among the chemicals selected for external validation, the model provided accurate predictions for 16 of 21 chemicals. The five chemicals with inaccurate predictions provide valuable insight into potential limitations or gaps of the model. As additional assay data become available for more chemicals in phase II of ToxCast, the predictive model of reproductive toxicity, and the other existing models, will be forward validated and expanded to improve overall model performance and fill recognized biological pathway gaps critical to reproductive function. Outside of Environmental Protection Agency (EPA), other researchers have demonstrated similar success using high-throughput approaches applied to nanomaterials testing (Nel et al., forthcoming). Because of our success in developing well-performing, statistically valid predictive models, we think that Thomas et al.’s (2012) failure to develop useful models is dependent on their methodology and not indicative of a systemic weakness in the ToxCast data.

Thomas et al. endorse the utility of ToxCast phase I data as a “survey of potential molecular initiating events . . . that can be interpreted as risk factors for toxicity.” We fully agree with that opinion and are working to expand the publicly available ToxCast data set (http://actor.epa.gov/actor/faces/ToxCastDB/Home.jsp) beyond phase I chemicals and assays. We believe that our published classifiers based on the existing ToxCast data are useful for prioritizing chemicals for further testing, and the applicability and robustness of these data and models will increase as the number of chemicals and assays continues to increase in ToxCast (Dix et al., 2007; Kavlock et al., 2012) and in the Tox21 collaboration between federal agencies in the United States (http://epa.gov/nctc/Tox21/).

This Letter to the Editor does not represent EPA policy.

REFERENCES


