Supporting read-across predictions of chemical toxicity using high-throughput text-mining

Nancy C. Baker[1], Thomas Knudsen[2], Antony J. Williams[2], Kevin M. Crofton[2], Grace Patlewicz[2]

[1]Leidos, RTP, NC; [2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC

Read-across is a technique used to fill data gaps within chemical safety assessments. It is based on the premise that chemicals with similar structures are likely to have similar biological activities. Known information on the property of a chemical (source) is used to make a prediction of the same property for another chemical (target) that is considered similar. Building scientific confidence in read-across remains an ongoing challenge. Although systematic frameworks have been established to identify sources of uncertainty, practical approaches to address uncertainty remain limited. Exploiting response profiles in high-throughput (HT) and/or high-content (HC) screening data may provide a means to characterize biological similarity, therefore reducing uncertainty. A largely untapped resource for read-across to date is the biomedical literature. This information has the potential to support read-across by facilitating the identification of valid source analogues with similar biological and toxicological profiles as well as providing the mechanistic understanding for any prediction made. A key challenge in using such information is to convert and translate its unstructured form into a computable format that can be linked to chemical structure. We developed a novel text-mining strategy to represent literature information as keyword features (toxicity signatures) at the chemical level. The elements of the toxicity signatures were weighted using a rule-based algorithm that assessed the strength of the literature relationship. This weight was used to rank and visualize the signature as literature ToxPIs (LitToxPIs) for ~6,000 chemicals described in the biomedical literature for a variety of toxicity types including genetic toxicity, developmental toxicity, reproductive toxicity and thyroid toxicity. We then developed a user interface (UI) that facilitates exploration of the literature evidence underpinning the signatures. As an example, the literature evidence extracted from the 2,745 articles about bisphenol A resulted in a toxicity signature showing reproductive toxicity as the most significant toxicity type for this chemical. When the same corpus of ~6,000 chemicals was filtered for chemicals with any evidence of reproductive toxicity, 2,092 chemicals were found. In a ranking of this subset of chemicals by the strength of the reproductive toxicity signature, bisphenol A was ranked third in the list. This UI provides a tool that allows researchers to substantiate structure based read-across predictions with literature reports of in vitro and in vivo toxicity and thereby increase scientific confidence in those predictions. *This abstract does not necessarily represent U.S. EPA policy.*