United States Environmental Protection Agency

Open / Big Data opportunities and challenges at EPA

Richard Judson
U.S. EPA, National Center for Computational Toxicology
Office of Research and Development

IUTOX Open Data in a Big Data World: A Toxicology Perspective
SOT 2019
March 11, 2019

EPA Projects Using Big Data

- Key Point: All data has to be public and easily accessible
- Prioritization
  - Need to select chemicals for risk assessment out of tens of thousands
  - Require data in multiple domains:
    - In vivo (human and ecotox), in vitro, pharmacokinetics, physico-chemical, exposure, use
    - Quantitative and qualitative information
    - Data needs to be organized, made consistent
- Supplementing and replacing traditional in vivo toxicology data
  - New Approach Methodologies (NAM)
  - In vitro data on thousands of chemicals
  - Many kinds of models

Major Data Resources

- Databases
  - ToxRefDB, ToxValDB – in vivo data from multiple public sources
    - EPA, NIH, FDA, DOE, ECHA, EFSA, states, NGOs
    - PODs, effects, genetox
    - 111,000 chemicals, 800,000 data points
  - Invitrodb – In vitro data from the ToxCast program
    - ~10,000 chemicals, 1000 assays
    - Concentration-response transcriptomics on 2000 chemicals (10^9 data points)
- Data Portals
  - First Generation: https://actor.epa.gov
  - Second Generation https://comptox.epa.gov
  - Public downloads of all data as flat files, database dumps
- Data Pipelines and Model software
  - Multiple software projects to process data and run models

Challenges

- Public data is heterogeneous and “imperfect”
  - Data must be cleaned and normalized (e.g. units, nomenclature)
  - Software can help with QC and normalization, but manual QC is unavoidable
- Much data is described in text in the open literature
  - Requires either expensive manual input or sophisticated text mining
- Open is not always open
  - Regulatory agencies often provide only brief summaries of data (e.g. chemical name and POD) but no further details
  - Limited data access means limited QC review
- Data is complex
  - Making it understandable to audiences with varying experience is hard