ScrubChem: Building Bioactivity Datasets from PubChem Bioassay Data

Jason Bret Harris[1, *] and Richard Judson[2]

1 .Oak Ridge Institute of Science and Education (ORISE), Oak Ridge, TN

2. National Center for Computational Toxicology (NCCT), U.S. EPA, RTP, NC

*Contact: Harris.Jason@epa.gov (ORCID: 0000-0002-7371-0463)

Abstract

The PubChem Bioassay database is a non-curated public repository with data from 64 sources, including: ChEMBL, BindingDb, DrugBank, EPA Tox21, NIH Molecular Libraries Screening Program, and various other academic, government, and industrial contributors. Methods for extracting this public data into quality datasets, useable for analytical research, presents several big-data challenges for which we have designed manageable solutions. According to our preliminary work, there are approximately 549 million bioactivity values and related meta-data within PubChem that can be mapped to over 10,000 biological targets. However, this data is not ready for use in data-driven research, mainly due to lack of structured annotations. **We used a pragmatic approach ~~for~~that provides increasing access to bioactivity values in the PubChem Bioassay database. This included restructuring of individual PubChem Bioassay files into a relational database (ScrubChem). ScrubChem contains all primary PubChem Bioassay data that was: reparsed; error-corrected (when applicable); enriched with additional data links from other NCBI databases;~~,~~ and improved by adding key biological and assay annotations derived from logic-based language processing rules. The utility of ScrubChem and the curation process were illustrated using an example bioactivity dataset for the androgen receptor alpha protein.** This initial work serves as a trial ground for establishing the technical framework for accessing, integrating, curating, analyzing, and making use of such massive bioactivity data. *This abstract does not necessarily reflect U.S. EPA policy.*