A public database supporting evidence-based metabolomics

Risa R. Sayre[1,2], John F. Wambaugh[1], Antony Williams[1], Christopher M. Grulke[1]

1) U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, RTP NC

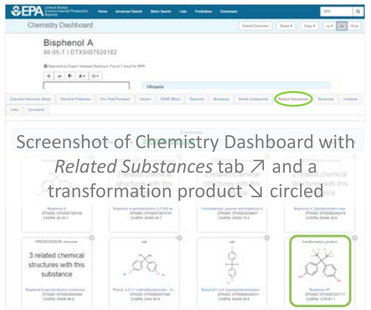
2) Oak Ridge Institute of Science and Education (ORISE) Research Participant

www.epa.gov

Risa R. Sayre | ORCID 0000-0002-6173-8020 | sayre.risa@epa.gov | 919-541-4871

Background

U.S. EPA's Rapid Exposure and Dosimetry (RED) project provides tools to rapidly generate quantitative human exposure and internal dose estimates. To **support identification of likely sources** of chemicals found in biological media through non-targeted/suspect screening analysis (SSA/NTA), our project added **substance relationships between chemicals** and their transformation products to the CompTox Chemistry Dashboard[1].



Screenshot of Chemistry Dashboard with *Related Substances* tab ↗ and a transformation product ↘ circled

Our effort is unique in that we include only **empirically-validated relationships** for a given species. Restriction to observed *in vivo* transformations allows **development of exposure estimates** based on dose levels demonstrated to yield a detectable amount of product. More thorough knowledge of exposome products and relationships can also identify **candidate substances and pathways** that lead to detectable internal doses to inform future **high-throughput assay research**.

Database content

We propose five categories of substances found in human biomonitoring samples:
1) **endogenous metabolome**,
2a) **exogenous nutrients**,
2b) **markers of exposure to exogenous nutrients**,
3a) **xenobiotics**, and
3b) **markers of exposure to xenobiotics**.

Your input requested: Are these categories sufficient?

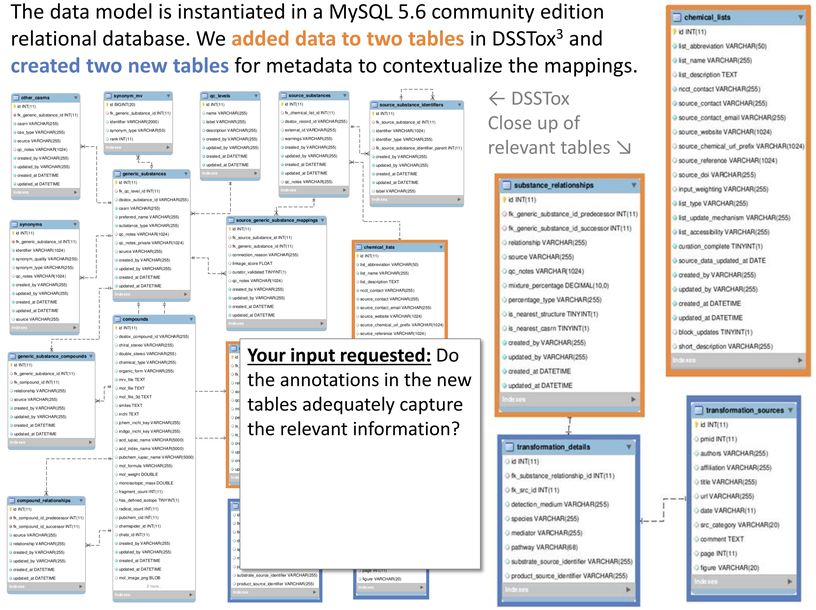
Substances are defined by their generation source, and are expected to be structurally heterogeneous. Some compounds can appear in more than one category. For example, formaldehyde is formed in amino acid production (1), can be observed internally after occupational exposures (3a), and is also formed in the body when breaking down methanol (3b). Another example is cholesterol: it is present in cellular membranes (1), from consumption of animal fat (2a), or as an effect of glucocorticoid medication (3b).

Your input requested: How do you recommend identifying the origins of NTA compounds?

Since the database seeks to support analytical chemists, we exclude compounds predicted based on pathways, which could represent intermediates that may not be detectable. We also don't assume conservation across species, due to cases like bisphenol A, where the conjugating enzyme in rats is comparable, but the product is found in different tissues due to different enterohepatic recirculation[2].

Database structure

The data model is instantiated in a MySQL 5.6 community edition relational database. We **added data to two tables** in DSSTox[3] and **created two new tables** for metadata to contextualize the mappings.



← DSSTox Close up of relevant tables ↘

Your input requested: Do the annotations in the new tables adequately capture the relevant information?

Data sources for xenobiotic mappings

Source	Structure	Chemical identifiers	Records	DSSTox Mappable
NHANES	XML from parsed PDF	name, CAS	164	75
TKKB	MySQL	name	1029	614
ChemBL	MySQL	name, InChI, SMILES	1245	101 (by name only)
T3DB	CSV	name, CAS, InChI	791	406
MeSH SCR	XML	name	2081	625
PubChem	parsed search results	name	344	117
HMDB	PMIDs from XML	name, CAS, InChI, SMILES	19362	203 (by name only)
DrugLabels	XML from parsed PDF	name	518	79
PubMed	text	name	≈109122	≈14502 potential

A record was considered a positive mapping when it contained a description of an experiment where one named compound was dosed and different compound(s) were detected in tissue(s) or excreta. The initial search effort yielded 1417 unique 3a/3b pairs where both parent and transformation product were already curated into DSSTox.

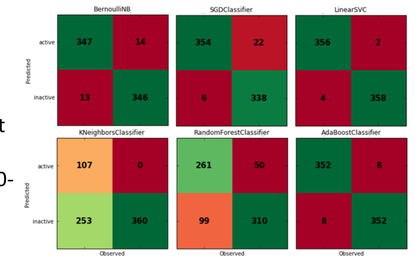
Your input requested: Any other data source ideas?

Collecting xenobiotic mappings from abstracts

In this example, we sought to find whether any chemical names referring to DSSTox identifications by mass:charge ratio and neutral monoisotopic mass of compounds detected in pooled human blood samples using LC-QTOF[4] were in category 3b. After finding a high number of false positives (>99%) in a PubMed search for "metabolite of [name]", we used the abstracts manually classified from that effort to build a natural language processing model to identify abstracts containing mappings more efficiently.

Method

- Balanced positive and negative sets of abstracts (362 each)
- Remove stop words (Stanford)
- Identify informative features from most common words (top √n of both sets)
- Train a suite of binary classifiers with 10-fold cross-validation
- Create a consensus prediction from all classifiers with >90% accuracy



↑ Confusion matrices for training performance of some tested algorithms

Results

The method was validated on 884 abstracts containing the chemical names of interest, 220 of which were known to be positive. The F1 score of consensus predictions was 98.0%, but there was no consensus for 9.5% of positives.

Discussion

We have over 10,000 rows of putative mappings of transformation products in our database, mostly xenobiotics. Future work toward developing methods to improve identification of substances measured in human blood and their sources supports specific research projects active within the agency (e.g. PFAS chemicals).

Your input requested: Please give feedback on making this product more helpful for your work.

References

Cited: 1) Williams AJ, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminform. 2017 Nov 28;9(1):61. 2) Taylor JA, et al. Similarity of Bisphenol A Pharmacokinetics in Rhesus Monkeys and Mice: Relevance for Human Exposure. Environ Health Perspect. 2011;119:422-430. 3) Richard AM, Williams CR. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutat Res. 2002 Jan 29;499(1):27-52. 4) McEachran AD, Sobus JR, Williams AJ. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal Bioanal Chem. 2017 Mar;409(7):1729-1735.
Other references: a) Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The Blood Exposome and Its Role in Discovering Causes of Disease. Environ Health Perspect. 2014 Aug; 122(8). b) Sobus JR, et al. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J Expo Sci Environ Epidemiol. 2017 Dec 29.

This project was supported in part by an appointment to the Internship/Research Participation Program at the National Center for Computational Toxicology, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.