Structure Identification Using High Resolution Mass Spectrometry Data and the EPA's CompTox Chemistry Dashboard

Antony Williams[1,*], Andrew McEachran[2], Jon Sobus[3]

[1]U.S.-EPA, Office of Research and Development, National Center for Computational Toxicology (NCCT), Research Triangle Park, NC
[2]Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, NC
[3]U.S.-EPA, Office of Research and Development, National Exposure Research Laboratory (NERL), Research Triangle Park, NC

United States Environmental Protection Agency

ORCID: 0000-0002-2668-4821
Antony Williams I williams.antony@epa.gov I 919-541-1033

EAS
Somerset, NJ
November 14-16, 2016

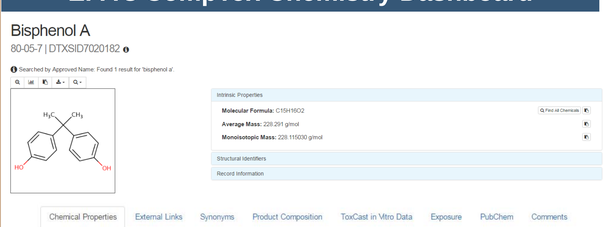
Problem Definition and Goals

Problem: Non-targeted analysis (NTA) and suspect screening workflows routinely identify <10% of chemicals observed in environmental samples.
Goals: To develop tools within the freely available and highly curated EPA CompTox Chemistry Dashboard to assist in structure identification to improve NTA and suspect screening workflows. Details regarding methodology and access will be standardized and open to mass spectrometry users.

Abstract

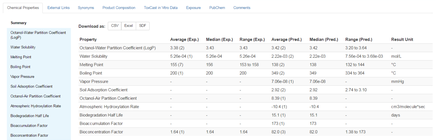
The iCSS CompTox Chemistry Dashboard (https://comptox.epa.gov) is a publicly accessible web-based application provided by the National Center for Computational Toxicology at the US-EPA. It serves a number of purposes, including providing a chemistry database underpinning many of our public-facing projects (e.g. ToxCast and ExpoCast). The available data and searches provide a valuable path to structure identification using mass spectrometry as the source data. With an underlying database of ~720,000 chemicals, the dashboard has already been used to assist in identifying chemicals present in house dust. This poster reviews the benefits of the EPA's platform and underlying algorithms used for the purpose of compound identification using high-resolution mass spectrometry data. Standard approaches for both mass and formula lookup are available but the dashboard delivers a novel approach for hit ranking based on functional use of the chemicals. The focus on high-quality data, novel ranking approaches and integration to other resources of value to mass spectrometrists makes the Dashboard a valuable resource for the identification of environmental chemicals.

EPA's CompTox Chemistry Dashboard



- The US EPA's CompTox Chemistry Dashboard is a public chemistry resource developed to support computational toxicology research efforts within the National Center for Computational Toxicology.
- The data underlying the Dashboard was assembled over a period of more than a decade into the DSSTox database.
- Chemical specific pages contain curated structure information related to a unique DSSTox Identifier (DTXSID), calculated intrinsic properties, structural identifiers, and detailed record and citation information.

Dashboard Chemical Page Features

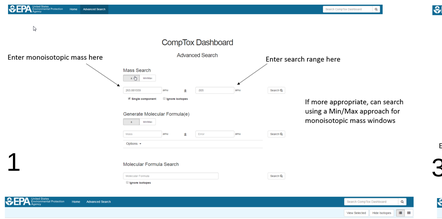


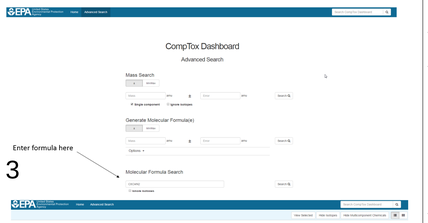
Data under the Chemical Properties tabs include experimental and predicted data (LogP, water solubility, melting points, etc. Predicted properties were modeled using a suite of QSAR models.

Product composition, frequent use and function data are compiled from the US EPA's Consumer Product Category Database (²CPCat, Dionisio *et al* 2015) for over 43,000 chemicals. information.

US EPA's ToxCast program uses high-throughput screening to evaluate, predict, and prioritize toxicity of chemicals. Bioassay data for all chemicals in ToxCast can be viewed.

Advanced Searching for Chemical Identification Using Mass Spectrometry Data





Unknown chemicals can be brought to the top of a list by rank-ordering by the number of associated data sources within a reference database. Monoisotopic masses or chemical formulae associated with spectral features are searched within the Dashboard and the results rank-ordered. Figs. 1 &2 demonstrate searching an unidentified monoisotopic mass observed via NTA and rank-ordering the results to bring the most likely chemicals to the top. Figs. 3 & 4 demonstrates the process when a user has already generated a molecular formula.

Compound class	Number in class	Average Rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical Drug	72	1.3	59	8	3	2	
Manufacturing Chemicals	42	1.2	38	1	1	2	
Personal Care Products	8	2.6	6				2
Steroid Hormones	7	1.0	7				
Perfluorochemicals	6	1.3	5	1			
Pesticides	12	1.3	10	1	1		
Veterinary Drugs	3	1.0	3				
Dyes	2	1.0	2				
Food product/natural compounds	4	1.5	3		1		
Illicit Drugs	2	1.5	1	1			
Misc. Molecules	3	1.0	3				

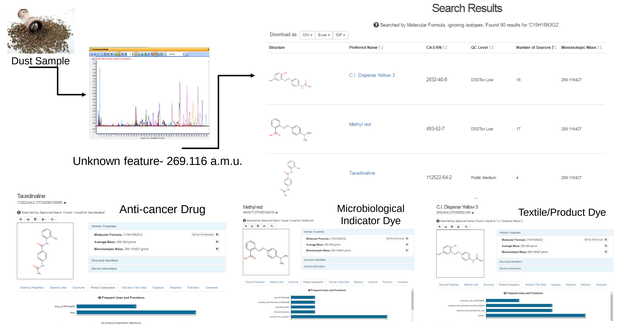
Compound class	Number in class	Average Rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical Drug	72	1.4	55	9	6	2	
Manufacturing Chemicals	42	5.5	28	6	3		5
Personal Care Products	8	6.1	3	1			4
Steroid Hormones	7	1.0	7				
Perfluorochemicals	6	1.2	5	1			
Pesticides	12	2.3	6	2	3		1
Veterinary Drugs	3	1.3	2	1			
Dyes	2	1.0	2				
Food product/natural compounds	4	3.8	2			1	1
Illicit Drugs	2	2.0	1		1		
Misc. Molecules	3	1.3	2	1			

Comparison of results of searching and rank-ordering by data sources between the Dashboard (top) and ChemSpider (bottom) split up by compound classes.

	Mass-based Searching		Formula-based Searching	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Average Rank Position	1.3	2.2	1.2	1.4
Percent in #1 Position	85%	70%	88%	80%

Advanced searching using rank-ordering methods to identify unknowns was conducted in both the Dashboard and in ChemSpider to compare the new application (Dashboard) with the community standard (ChemSpider). 162 unique chemicals were searched by both monoisotopic mass and formula and results rank-ordered by data sources. The Dashboard significantly outperformed ChemSpider in terms of the average rank position within each search method. Additionally, the percentage of chemicals in the #1 position was higher for the Dashboard in both cases.

Functional Use for Structure Identification



- Functional use and product occurrence data can assist in the identification of unknown chemical features when sample information is known
- When data source count alone cannot definitively distinguish an unknown, the next step is to look at functional use information
- In this example, an unknown feature from a house dust sample led to several potential chemicals but was most likely to be a textile/product dye over an anti-cancer drug or microbiological dye

Future Work

- "MS-Ready Structures":
  Mass spectrometry users observe data in spectra as the desalted form of a chemical with no stereochemistry information. These "MS-ready forms" of chemical substances are being incorporated in the Dashboard so that searching for substances is performed on the detectable form of the chemical – the desalted, non-stereo form of the chemical

- Non-Targeted Analysis Collaboration Research Trial:
  The US EPA is conducting a collaborative research trial with more than 20 labs and institutions from around the world in an effort to develop, improve, and standardize analytical and data processing methods in NTA and suspect screening analysis

References

1. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal, A.M. Richard, C.R. Williams, Mutat Res. 2002 Jan 29;499(1):27-52.
2. Dionisio, Kathie L., *et al.* "Exploring consumer exposure pathways and patterns of use for chemicals in the environment." Toxicology Reports 2 (2015): 228-237.
3. McEachran *et al.* Identifying "Known Unknowns" Using the US EPA's iCSS CompTox Dashboard." In review.