Environment International 101 (2017) 19–26

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Full Length Article

Incorporating ToxCast and Tox21 datasets to rank biological activity of chemicals at Superfund sites in North Carolina

Sloane K. Tilley [a], David M. Reif [b], Rebecca C. Fry [a,c,*]

[a] Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina, Chapel Hill, NC 27599, USA
[b] Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA
[c] Curriculum in Toxicology, The University of North Carolina, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 10 May 2016
Received in revised form 9 September 2016
Accepted 6 October 2016
Available online 31 January 2017

Keywords:
North Carolina
Risk assessment
Superfund
Tox21
ToxCast
ToxPi

ABSTRACT

Background: The Superfund program of the Environmental Protection Agency (EPA) was established in 1980 to address public health concerns posed by toxic substances released into the environment in the United States. Forty-two of the 1328 hazardous waste sites that remain on the Superfund National Priority List are located in the state of North Carolina.
Methods: We set out to develop a database that contained information on both the prevalence and biological activity of chemicals present at Superfund sites in North Carolina. A chemical characterization tool, the Toxicological Priority Index (ToxPi), was used to rank the biological activity of these chemicals based on their predicted bioavailability, documented associations with biological pathways, and activity in *in vitro* assays of the ToxCast and Tox21 programs.
Results: The ten most prevalent chemicals found at North Carolina Superfund sites were chromium, trichloroethene, lead, tetrachloroethene, arsenic, benzene, manganese, 1,2-dichloroethane, nickel, and barium. For all chemicals found at North Carolina Superfund sites, ToxPi analysis was used to rank their biological activity. Through this data integration, residual pesticides and organic solvents were identified to be some of the most highly-ranking predicted bioactive chemicals. This study provides a novel methodology for creating state or regional databases of biological activity of contaminants at Superfund sites.
Conclusions: These data represent a novel integrated profile of the most prevalent chemicals at North Carolina Superfund sites. This information, and the associated methodology, is useful to toxicologists, risk assessors, and the communities living in close proximity to these sites.

1. Introduction

To conclude the "decade of the environment" of the 1970s, the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), also known as Superfund, was established to facilitate the remediation of abandoned hazardous waste sites around the United States (EPA, 2015a). Since its induction in 1980, over 1700 sites have been added to the National Priorities List (NPL), which is used by the Environmental Protection Agency (EPA) to prioritize the cleanup of sites that pose the greatest potential risk to public health. In 2006, the EPA celebrated the completion of construction at the one thousandth Superfund site (EPA, 2015a). The cleanup of these hazardous waste sites has had numerous public health and economical benefits on the surrounding communities (Arndt et al., 1999; EPA, 2011; Gamper-rabindran et al., 2011; Henry et al., 2015). However, many Superfund sites remain as hazardous waste sites that pose significant threats to human health through contaminant exposure *via* soil, surface water, or ground water (Carlin et al., 2015; Jung et al., 2011; Messier et al., 2012; EPA, 2013).

Environmental exposures to chemicals contribute to numerous adverse health effects in humans, including carcinogenesis, nephrotoxicity, hepatotoxicity, and immune system dysfunction (Prüss-Ustün et al., 2011). The World Health Organization (WHO) estimated that approximately 23% of deaths around the globe in 2012 were attributable to modifiable environmental risk factors (World Health Organization, 2016). Cleaning up areas known to be contaminated with toxicants harmful to human health, therefore, represents a significant way to reduce disease risk in a community (EPA, 2011). However, there are

Abbreviations: ATSDR, Agency of Toxic Substances and Disease Registry; CASRN, Chemical Abstracts Service Numbers; CERCLA, Comprehensive Environmental Response, Compensation, and Liability Act; CTD, Comparative Toxicogenomics Database; EPA, Environmental Protection Agency; HMDB, Human Metabolome Database; IRIS, Integrated Risk Information Systems; KEGG, Kyoto Encyclopedia of Genes and Genomes; N.C., North Carolina; NPL, National Priorities List; SAA, Superfund Alternative Approach; ToxPi, Toxicological Priority Index; UNC SRP, University of North Carolina Superfund Research Program; WHO, World Health Organization.

* Corresponding author: Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, CB 7431, University of North Carolina, Chapel Hill, NC 27599, USA.
E-mail addresses: sloanekathryn@unc.edu (S.K. Tilley), dmreif@ncsu.edu (D.M. Reif), rfry@unc.edu (R.C. Fry).

http://dx.doi.org/10.1016/j.envint.2016.10.006
0160-4120/

many difficulties associated with remediation of a hazardous waste site, including characterization of likely routes of exposure, the complications of exposures to chemical mixtures, consideration of health effects in susceptible populations, determining the extent of cleanup needed at a particular site, and prioritization of sites for cleanup (Anderson et al., 2015; Harris and Wrenn, 1988). The cleanup process is made more difficult by the paucity of data detailing the mechanism of action by which the >200 chemicals commonly found at Superfund sites could potentially harm human health and the dose at which they pose a threat (Committee on Developmental Toxicology et al., 2000; Cronin, 2003).

In the state of North Carolina, there are currently 42 Superfund sites on the NPL, spread throughout 22 different counties (EPA, 2016a). The University of North Carolina Superfund Research Program (UNC SRP) aspires to further identify, characterize, and communicate the human and environmental health risks associated with chemicals found at these sites (UNC Superfund Research Program, 2016). Although these hazardous waste sites have been added to the NPL, there is an insufficiency of data on which chemicals are present at North Carolina Superfund sites and what potential health risks they pose. While the EPA maintains information about Superfund site status and history, the relative biological activities of the contaminants located at each site are not publically available. Without these data, both the characterization and prioritization of potential risks that remain at Superfund sites remain difficult.

In this study, the ToxPi computational chemical prioritization tool was used to compare the biological activity of chemicals found at North Carolina Superfund sites (Reif et al., 2013, 2010). The advantage of using ToxPi is the ability to integrate different types of data, including experimental and predicted chemical properties of compounds, pathway analyses, and *in vitro* assays into a comprehensive predictive profile of a chemical's biological activity. In addition, our analysis provides a ranking system that could be used for chemical prioritization in remediation efforts at many hazardous waste sites throughout the United States. These results allowed us to aggregate data on each of North Carolina Superfund sites to create a database and methodology useful to both risk assessors and community members.

2. Materials and methods

2.1. Identifying North Carolina superfund sites

Superfund sites on the NPL and Superfund Alternative Approach sites in North Carolina were identified from the EPA Superfund website. Geographic locations were plotted by street address using CartoDB (https://cartodb.com/). Sites that were deleted from the NPL were excluded from further analysis, as remediation has been completed, and the EPA has determined that there is no longer a threat of human exposure. These were Martin-Marietta, Sodyeco, Inc., New Hanover County Airport Burn Pit, and the 210 Mile Roadside PCB Spill. Additionally, sites that are currently undergoing site investigation were excluded from further analysis since no information about contaminants present at the site was available at the time of analysis. These included Kerr-McGee Chemical Corporation and Cristex Drum.

2.2. Data extraction and compilation

2.2.1. Superfund chemical data extraction and compilation

Contaminant lists of concern were obtained for every site, where available. Several sites had incomplete chemical characterization, with only a few chemicals listed under the site description, but no official list of contaminants of concern. For these sites, which were CTS – Asheville, Ecusta Mill, Hemphill Road TCE, Holtra Chem, and Wright Chemical Corporation, the chemicals mentioned in the description of the site were used instead of a list of contaminants of concern. Chemical Abstracts Service Numbers (CASRNs) were obtained by querying chemical names in the Aggregated Computational Toxicology Resource (ACToR) database. Data from chemicals with the same CASRN were combined so that each CASRN only corresponded to one chemical name.

2.2.2. Chemical and physical property data extraction and compilation

From chemicalize.org by ChemAxon, the following data were gathered for each chemical by CASRN: average molecular weight (g/mol), polar surface area (Å^2), 1-ocatanaol-water coefficient (logP), bioavailability score (0/1), Lipinski's Rule of Five (0/1), Ghose Filter (0/1), Veber's Rule (0/1), Lead-like Rule (0/1), and Muegge Filter (0/1). Bioavailability is a measure of whether the compound is orally bioavailable upon ingestion. The remaining five filters are used in pharmacology to estimate the solubility and permeability of orally active compounds based on their physical and chemical properties (Bickerton et al., 2012; Lajiness et al., 2004; Lipinski, 2004; Lipinski and Hopkins, 2004; Muegge, 2002). In addition, the aqueous solubility, logS, of each compound was predicted in ALOGPS 2.1, if it was available in the database (Tetko et al., 2005; "VCCLAB, Virtual Computational Chemistry Laboratory,", 2005). Where available, experimental logP and logS values were recorded instead of the predicted values.

2.2.3. KEGG pathway data extraction and compilation

Using the Comparative Toxicogenomics Database (CTD), each of the toxicants were queried in the "Chemicals" section of the CTD by CASRN, if available. In the cases were the substance did not have a CASRN (*e.g.* pesticides), the name of the compound was queried instead. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that were significantly enriched ($p < 0.01$) among genes that interact with that chemical were downloaded and categorized by the functional annotations provided by the KEGG Pathway Database (http://www.genome.jp/kegg/pathway.html). The significance of enrichment was calculated by the hypergeometric distribution and adjusted for multiple testing using the Bonferroni method (CTD, 2016). Where pathway data were not available for a specific chemical, the parent compound pathway data were downloaded and used instead so all chemicals had KEGG pathway annotations.

2.2.4. ToxCast and Tox21 assay data extraction and compilation

Phase II ToxCast and Tox21 Summary Files were downloaded from the EPA website (https://www3.epa.gov/research/COMPTOX/toxcast_summary.html), with information from 353 *in vitro* assays, corresponding to 1191 assay endpoints. A total of 9076 chemicals had data available. All assays used in the Tox21 program are required to be compatible with the National Center for Advancing Translational Sciences assay criteria, which specify acceptable parameters of assay validation (Sittampalam et al., 2016). Before use in the Tox21 program, all assays are optimized and validated for reproducibility in triplicate against a reference library that includes positive controls (NCATS, 2016). In addition, all standard operating procedures for each assay and the concentration-response curves for all assays for each chemical tested are made publically available (NCATS, 2016). Through this robust screening procedure, all assays used in the Tox21 program adhere to the overall validation framework developed by the Organization of Economic Co-operation and Development in collaboration with both the Interagency Coordinating Committee on the Validation of Alternative Methods and the European Centre for the Validation of Alternative Methods (Judson et al., 2013). Assay data were downloaded from the EPA website in the file 'Assay_Summary_151020.csv'. The hit-call matrix identified chemicals by CASRN number, except for explicit mixtures and pharmaceutical compounds, where arbitrary identifiers were assigned beginning with "NOCAS". If data were available, chemicals found at North Carolina Superfund sites were matched to chemicals assayed in the ToxCast and Tox21 programs. For the remaining chemicals, ToxCast and Tox21 data from isomers or compounds of these chemicals were used instead, where available. Where multiple compounds were available for one chemical, the following preference ranking was used: isomers, acetates, chlorides, sodium anhydrides,