Toxicology Reports 3 (2016) 723–732

Contents lists available at ScienceDirect

Toxicology Reports

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

Characterization and prediction of chemical functions and weight fractions in consumer products

Kristin K. Isaacs[a,*], Michael-Rock Goldsmith[b,1], Peter Egeghy[a], Katherine Phillips[c], Raina Brooks[d], Tao Hong[e], John F. Wambaugh[f]

[a] U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[b] Chemical Computing Group, Suite 910, 1010 Sherbrooke Street West, Montreal, QC H3A 2R7, Canada
[c] Oak Ridge Institute for Science and Education, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[d] Student Services Contractor, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[e] ICF International, 2635 Meridian Pkwy #200, Durham, NC 27713, United States
[f] U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States

ARTICLE INFO

Article history:
Received 17 May 2016
Received in revised form 5 August 2016
Accepted 31 August 2016
Available online 1 September 2016

Keywords:
Chemical function
Exposure modeling
Chemical prioritization
Consumer products
Cosmetics
ExpoCast

ABSTRACT

Assessing exposures from the thousands of chemicals in commerce requires quantitative information on the chemical constituents of consumer products. Unfortunately, gaps in available composition data prevent assessment of exposure to chemicals in many products. Here we propose filling these gaps via consideration of chemical functional role. We obtained function information for thousands of chemicals from public sources and used a clustering algorithm to assign chemicals into 35 harmonized function categories (e.g., plasticizers, antimicrobials, solvents). We combined these functions with weight fraction data for 4115 personal care products (PCPs) to characterize the composition of 66 different product categories (e.g., shampoos). We analyzed the combined weight fraction/function dataset using machine learning techniques to develop quantitative structure property relationship (QSPR) classifier models for 22 functions and for weight fraction, based on chemical-specific descriptors (including chemical properties). We applied these classifier models to a library of 10196 data-poor chemicals. Our predictions of chemical function and composition will inform exposure-based screening of chemicals in PCPs for combination with hazard data in risk-based evaluation frameworks. As new information becomes available, this approach can be applied to other classes of products and the chemicals they contain in order to provide essential consumer product data for use in exposure-based chemical prioritization.

1. Introduction

Assessment of the risks associated with chemicals in consumer products relies not only on characterization of hazard or toxicity, but also on the exposures encountered during use [1,2]. Consumer products contain and can release large numbers of chemicals to which humans are exposed directly during use or indirectly via contact with contaminated household air or dust [3–9]. Consumer product chemicals have been widely found in human blood and urine, and exposures from indoor or proximate "near-field" sources (which include consumer formulations and articles) generally are larger than the doses that result from "far-field" (e.g., industrial) sources of exposure [10–12].

Despite such high potential for exposure, critical gaps exist in both qualitative information describing the variety of chemicals contained in different categories of consumer products and in quantitative data on the weight fractions, both of which are key inputs to numerous exposure assessment frameworks and models [13–18]. However, due to limited public reporting requirements, confidential business considerations, and lack of harmonized chemical and product categorizations, specific data describing the composition of consumer products are often unavailable or incomplete [19,20].

These critical data gaps impede the quantification of exposures due to consumer product sources, and are especially noteworthy when considered in the context of prioritizing thousands of untested commercial chemicals on the basis of risk. The U.S. Environmental Protection Agency (EPA), under its ExpoCast program, is developing high-throughput (HT) computational methods for prediction of chemical exposures for combination with *in vitro* hazard information [1], with a particular goal of developing expo-

* Corresponding author.
E-mail address: isaacs.kristin@epa.gov (K.K. Isaacs).
[1] Work performed while an EPA employee.

http://dx.doi.org/10.1016/j.toxrep.2016.08.011

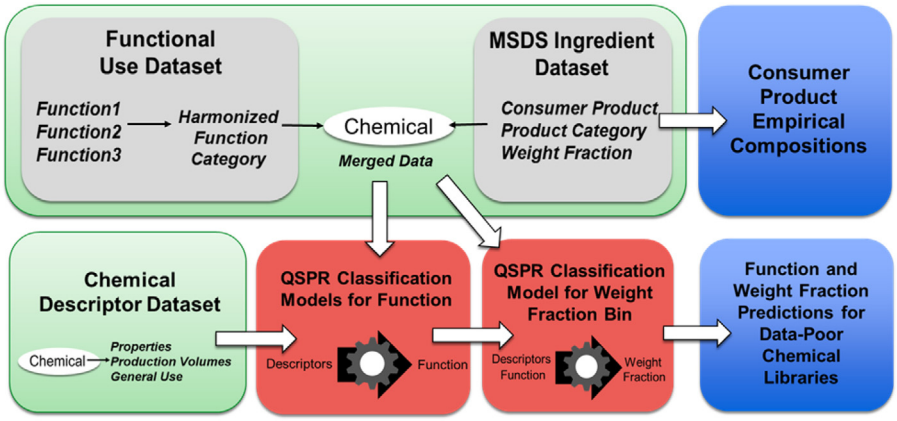

Fig. 1. Workflow for using existing chemical function and weight fraction datasets to build empirical compositions and QSPR classification models for chemical function and weight fraction for use in estimation of chemical exposure.

sure estimates for chemicals being evaluated by the ToxCast [21] initiative and the Tox21 interagency consortium [22]. A recent focus of ExpoCast has been the development of improved near-field (e.g., residential) exposures using both empirical [10,23] and mechanistic [13] approaches. To parameterize these efforts, U.S. EPA has developed new sources of information on how chemicals are used in commerce. EPA's Chemical/Product Categories Database (CPCat) [24,25] is a harmonized index of chemical use in products and sectors based on multiple publicly available data sources. One source within CPCat, the Consumer Product Chemical Profile Database (CPCPdb) [26] contains product ingredients and quantitative weight fractions derived from Material Safety Data Sheets (MSDS) for 1797 chemicals in nearly 9000 consumer products. Unfortunately, these quantitative data are limited to a relatively small fraction of products (and chemicals) currently in commerce. Methods are needed for extrapolating this existing knowledge to additional products and chemicals in a systematic manner.

In this work we present an approach for filling gaps in consumer product chemical use and composition data based on chemical function, and apply it to a case study of chemicals in personal care products (PCPs). Intentionally-added chemicals are present in consumer products because they serve a specific functional role that addresses either product performance or marketability. The functional role of an ingredient is defined by the chemical's properties and aids in determining its weight fraction in products. For example, Chevillotte et al. [27] described an exposure assessment method for cosmetics based on developing a "standard" or "virtual" composition of a product based on the weight fractions associated with chemical "families" across multiple product formulations. These families included functions such as "plasticizer" and "solvent." Here, we build on this approach by collecting and curating publicly-available function categorizations for thousands of chemicals, and combine these function categories with MSDS-based product weight fractions to build empirical compositions (or general formulations) based on real products in commerce for 66 categories of PCPs. These empirical compositions will be useful for parameterizing consumer exposure models for new or existing PCP chemicals when quantitative composition information is not available.

In addition to generating virtual compositions, we also describe a framework for predicting the probability of an arbitrary chemical having a given functional role and associated product weight fraction. This framework combines the function and ingredient weight fraction data to generate a series of machine-learning quantitative structure property relationship (QSPR) classification models for predicting functional role and weight fraction for large numbers of chemicals from chemical properties and other available descriptors (Fig. 1). These supervised learning models make use of known information about the characteristics of chemicals having certain functions to classify chemicals for which function is unknown. We apply these models to predict chemical functions for a library of over 10000 chemicals that are mostly data-poor, and corresponding weight fractions for hundreds of chemicals known to be present in PCPs. These methods are flexible and can be extended to additional chemical functions, products, or use sectors in support of HT prioritization of large numbers of chemicals on the basis of exposure potential or risk.

2. Methods

2.1. Chemical function data

Data describing the functions associated with individual chemicals (identified by Chemical Abstract Service Registry Numbers, CASRNs) were obtained from publicly-available government and industry sources; these data were curated into a harmonized Functional Use (FUse) database. Details (including sources) are provided in the Supplemental Information (SI). The largest source of data was the European Commission's Cosmetic Ingredient Database (CosIng) [28]. CosIng identifies different functional roles for cosmetic ingredients; a cosmetic is defined in CosIng to include a wide range of PCPs including lotions and creams, make-up, hair and body cleansing products, dental care products, fragrances, deodorants and antiperspirants, and suncreens [29]. A total of 10373 unique chemicals in PCPs were identified.

Many of the chemicals in the database were associated with multiple function categories. For the purpose of this study, we harmonized the function categories based on the similarity of the chemical groups associated with each category. For example, the majority of the chemicals classified as surfactants were also identified as cleansers and/or emulsifiers, so these chemicals were combined into a single harmonized category. This harmonization was based on a cluster analysis[30] of the function "fingerprint" of chemicals; a total of 36 harmonized functional categories relevant to PCPs were identified (details of the analysis and the results provided in the SI). A few chemicals had a greater variety of function classifications (e.g., "ethyl alcohol" had six functions: antifoamer, antimicrobial, astringent, solvent, masking agent, and viscosity controller). These chemicals were not assigned a single harmonized function for the purpose of calculating empirical compositions. Instead, these chemicals were categorized by name in the dataset, as they take on a variety of functions across different