Mining of Consumer Product and Purchasing Data to Identify Potential Chemical Co-exposures

Zachary Stanfield[1,2], Cody Addington[1,3], Kathie Dionisio[1], David Lyons[1], Rogelio Tornero-Velez[1], Katherine Phillips[1], Timothy Buckley[2], and Kristin Isaacs[1]

1. Center for Computational Toxicology and Exposure, 2. Center for Public Health and Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 27711, United States, 3. Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee 37830, United States

Zachary Stanfield | Stanfield.zachary@epa.gov | 919-541-5780

Abstract

Background: Chemicals in consumer products are a major contributor to human chemical co-exposures. Consumers purchase and use a wide variety of products containing potentially thousands of chemicals. There is a need to identify potential real-world chemical co-exposures to prioritize *in vitro* toxicity screening. However, due to the vast number of potential chemical combinations, this has been a major challenge.
Objectives: To develop a data-driven procedure for identifying prevalent chemical combinations that humans are exposed through purchase and use of consumer products.
Methods: We applied frequent itemset mining on an integrated dataset linking consumer product chemical ingredient data with product purchasing data from 60,000 households to identify chemical combinations resulting from co-use of consumer products.
Results: We identified co-occurrence patterns of chemicals over all households as well as those specific to demographic groups. We also identified chemicals exhibiting aggregate exposure. Lastly, a case study of endocrine active chemicals revealed priority chemical combinations co-targeting receptors involved in biological signaling pathways.
Discussion: Integration and comprehensive analysis of household purchasing data and product-chemical information provided a means to assess human near-field exposure and inform selection of sets of chemicals for high-throughput screening in *in vitro* assays.

Materials and Methods

Consumer Product Purchase Data: The U.S. EPA established a material transfer agreement in 2013 with The Nielsen Company (US), LLC, to obtain consumer product purchasing (CPP) data for household products purchased during the year of 2012. Individuals used hand-held scanners to catalog products purchased. The data included limited demographic and product information (including Universal Product Code, UPC, and product name). Product categories included personal care products and household care products.

Consumer Product Ingredient Data: Data on chemicals in specific consumer products were obtained from the most recent version of the EPA's Chemical and Products Database (CPDat) (Dionisio et al. 2018). The CPDat ingredient data were obtained via collection and curation of one of three types of data documents: public safety data sheets (SDS), ingredient lists, and manufacturer ingredient disclosures. Chemical identifiers were harmonized to unique substance identifiers using EPA CompTox Chemicals Dashboard (comptox.epa.gov/dashboard). These data were merged with the purchasing data using direct UPC matching and fuzzy matching to product names.

Frequent Itemset Mining (FIM): The merged purchasing and ingredient data were mined to obtain prevalent chemical combinations introduced to individual homes. A presence-absence matrix was created where a row corresponds to all the purchases made by a single household in a single month and has a value of 1 for all chemicals that were in the products and 0 for those not. The Equivalence Class Transformation (Eclat) algorithm (implemented in R with the arules package) was applied to this matrix to identify the most prevalent chemicals and chemical sets.

Chemical Analyzed: The data were limited to the active public chemical inventory of the Toxic Substances Control Act (TSCA), obtained from EPA's CompTox Chemicals Dashboard (US EPA 2020) to eliminate ingredients of limited interest (e.g., water). As a more focused pathway-based case study, a set of potential endocrine active chemicals was analyzed. The chemicals were identified using results from the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (Mansouri et al. 2016) and the Collaborative Modeling Project for Androgen Receptor Activity (COMPARA) (Mansouri et al. 2020).

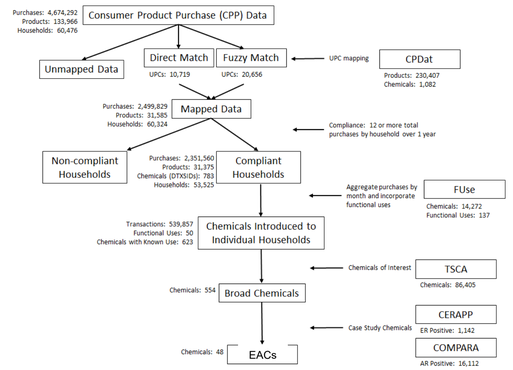
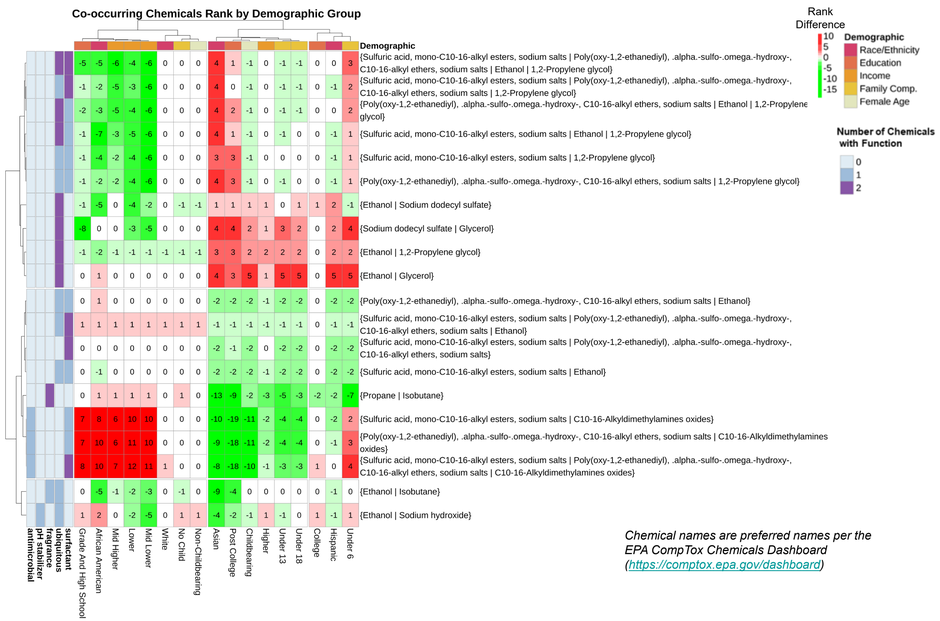


Figure 1. Data processing pipeline for frequent itemset mining of chemicals in consumer purchasing. Consumer purchasing data was obtained from Nielsen, mapped with a database linking chemicals to products (CPDat), integrated with chemical functional use information, and purchases were aggregated by month to focus on chemical co-exposure. For analysis, chemicals were limited to a broad set (TSCA) and a smaller pathway-based case study (endocrine active chemicals).

Abbreviations: CPDat (Consumer Products Database; Dionisio et al. 2018), UPC (Universal Product Code), FUse (functional use data; Phillips et al. 2017), TSCA (Toxic Substances Control Act; US EPA 2020), CERAPP (Collaborative Estrogen Receptor Activity Prediction Project; Mansouri et al. 2016) COMPARA (Collaborative Modelling Project for Androgen Receptor Activity; Mansouri et al. 2020), EACs (Endocrine active chemicals; determined by in silico models)

Results



Chemical names are preferred names per the EPA CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard)

Figure 2. Ranking of co-occurring chemicals. Heat map illustrating the ranked support for the 20 most prevalent chemical combinations. For a demographic, green color denotes a downward shift of rank relative to the global (lower potential exposure) and red denotes an upward shift (higher potential exposure). Cell numbers quantify the unit change in rank relative to the global rank. Column annotations indicate demographic categories and row annotations indicate harmonized functional use of chemicals. Rows and columns were clustered using complete linkage hierarchical clustering based on correlation of rank departures. A prevalence threshold of 2.5% was used.

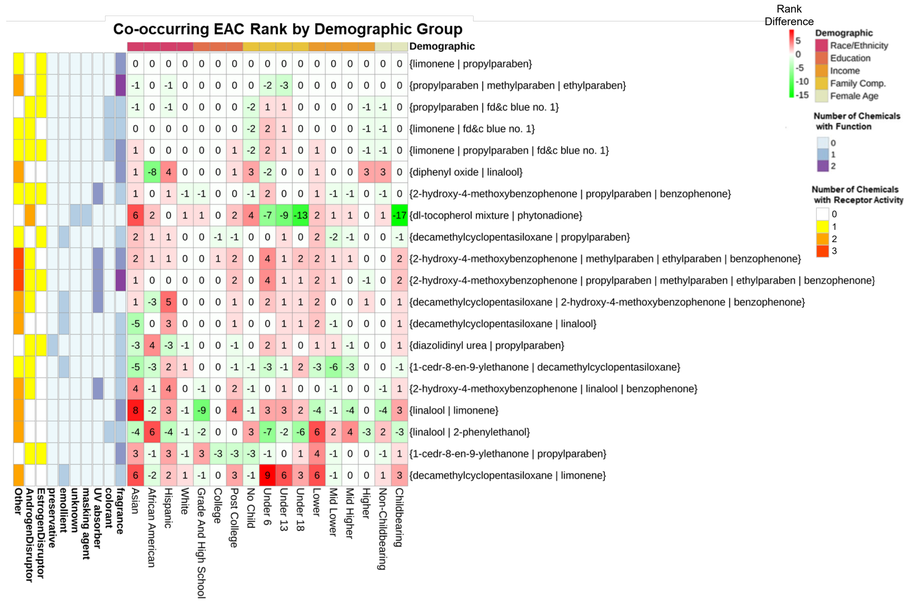


Figure 3. Ranking of co-occurring endocrine active chemicals (EACs). Heat map illustrating the ranked support for the 20 most prevalent EAC combinations. For a demographic, green color denotes a downward shift of rank relative to the global (lower priority) and red denotes an upward shift (higher priority). Cell numbers quantify the unit change in rank relative to the global rank. Column annotations indicate demographic categories and row annotations indicate harmonized functional use of chemicals and their predicted target receptors. A prevalence threshold of 0.1% was used.

Conclusions and Future Work

Conclusions

- Individuals are exposed to thousands of products that together contain hundreds of chemicals.
- It is possible to prioritize chemical combinations, using approaches like FIM, to greatly reduce the potential number of *in vitro* tests that are needed.
- Various demographics are exposed to certain chemicals (and combinations) at different rates. This is also true within certain product categories.
- EACs are present in commonly purchased consumer products, which are used in households with a level of consistency that make them an important factor to consider for hazard or toxicity assessment.
  - There is an observed overlap in prevalent itemsets with previous studies looking at chemical sets in consumer products and human urine.

Future Work

- Great efforts were made to obtain the product purchasing information and chemical ingredient data for consumer products, but the more this work is continued, the more informative the results will be.
  - Improve methods and strategies for determining chemicals present in consumer products.
  - Improve population sampling to have more diversity.
  - Include product purchases from a wider set of retailers (specialty stores, convenience stores, etc.) to expand chemical universe.
- Select itemsets for targeted *in vitro* testing
  - Top broad itemsets occurring in all households
  - Top EAC sets utilizing endocrine-specific assays

References

Dionisio KL, Phillips K, Price PS, Grulke C, Williams A, Biryol D, et al. 2018. The chemical and products database, a resource for exposure-relevant data on chemicals in consumer products. Scientific Data 5:1-9.
Phillips KA, Wambaugh JF, Grulke CM, Dionisio KL, Isaacs KK. 2017. High-throughput screening of chemicals as functional substitutes using structure-based classification models. Green Chem 19:1063-1074.
Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, et al. 2016. Cerapp: Collaborative estrogen receptor activity prediction project. Environmental health perspectives 124:1023-1033.
Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, et al. 2020. Compara: Collaborative modeling project for androgen receptor activity. Environmental health perspectives 128:27002.