Published in final edited form as:

ESCAPE. 2018 ; 43: 421–426. doi:10.1016/B978-0-444-64235-6.50076-0.

Optimal Chemical Grouping and Sorbent Material Design by Data Analysis, Modeling and Dimensionality Reduction Techniques

Melis Onel^{a,b}, Burcu Beykal^{a,b}, Meichen Wang^c, Fabian A. Grimm^c, Lan Zhou^d, Fred A. Wright^e, Timothy D. Phillips^c, Ivan Rusyn^c, and Efstratios N. Pistikopoulos^{a,b}

^aArtie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX, 77843, USA

^bTexas A&M Energy Institute, Texas A&M University, College Station, TX, 77843, USA

^cDepartment of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, 77843, USA

^dDepartment of Statistics, Texas A&M University, College Station, TX, 77843, USA

^eBioinformatics Research Center, North Carolina State University, Raleigh, NC, 27695-7566, USA

Abstract

The ultimate goal of the Texas A&M Superfund program is to develop comprehensive tools and models for addressing exposure to chemical mixtures during environmental emergency-related contamination events. With that goal, we aim to design a framework for optimal grouping of chemical mixtures based on their chemical characteristics and bioactivity properties, and facilitate comparative assessment of their human health impacts through read-across. The optimal clustering of the chemical mixtures guides the selection of sorption material in such a way that the adverse health effects of each group are mitigated. Here, we perform (i) hierarchical clustering of complex substances using chemical and biological data, and (ii) predictive modeling of the sorption activity of broad-acting materials via regression techniques. Dimensionality reduction techniques are also incorporated to further improve the results. We adopt several recent examples of chemical substances of Unknown or Variable composition Complex reaction products and Biological materials (UVCB) as benchmark complex substances, where the grouping of them is optimized by maximizing the Fowlkes-Mallows (FM) index. The effect of clustering method and different visualization techniques are shown to influence the communication of the groupings for read-across.

Keywords

Clustering; dimensionality reduction; predictive modeling; read-across

1. Introduction

Climate change has become one of the major risk factors for chemical contamination events. Therefore, precise and rapid examination of the complexity of the hazardous chemical exposures is essential to identify the potential adverse health impacts, and subsequently to provide immediate solutions and/or prevent further catastrophic events. At Texas A&M Superfund Research Program (TAMU SRP), we aim to develop comprehensive tools and models for addressing exposure to unknown chemical mixtures, and accordingly design solutions for the community during environmental emergency-related contamination events (TAMU Superfund Research Center, 2017). In this paper, we present two applications where data analysis, modeling and dimensionality reduction techniques guide experimental design and decision-making in biomedical and environmental areas.

First, we aim to design a framework for optimal grouping of unknown chemical mixtures based on their multi-dimensional analytical chemistry and bioactivity profiles. Detailed chemical characterization of a chemical mixture is challenging due to the variation in chemical composition during environmental emergencies. To provide rapid solutions, grouping an unknown chemical mixture to a group of well-studied, “known”, chemicals is critical. Here, the hypothesis is that, once an unknown chemical mixture is grouped into a cluster of known chemicals, read-across between cluster members would bridge the gap between data-poor and data-rich chemical substances. To do this, we use an integrated data analysis framework with dimensionality reduction and clustering techniques. Furthermore, we optimize the clustering structures by incorporating the Fowlkes-Mallows (FM) index, a measure for comparative assessment of clustering quality. Secondly, the environmental chemical contaminants can easily get mobilized, subsequently contaminate soil, and threaten the safety of the municipal water and food supply during environmental emergencies. In order to minimize the adverse health effects of chemical exposures, we aim to identify and develop novel broad-acting, high-capacity sorbents, enterosorbents, which can be implemented in diets to reduce the bioavailability of chemical mixtures. Here, the selection of the optimal sorption material for a given chemical mixture is a challenging and iterative task, where the chemical-sorbent property space needs to be explored iteratively to fine-tune and guide the experimental designs. Therefore, we perform predictive modeling of sorption activity of materials via regression techniques.

2. Data Acquisition

2.1. Derivation of Analytical Chemistry and Bioactivity Profiles of Complex Substances

In order to explore a more suitable clustering algorithm that may be used to establish the chemical and biological similarity between complex substances or mixtures, we use several recent examples of substances of Unknown or Variable composition Complex reaction products, and Biological materials (UVCB substances) as a benchmark. These include CON01-CON05 as Straight Run Gas Oils (SRGO); CON12-CON18, CON20 as Vacuum & Hydro-treated Gas Oils (VHGO); A083/13, A087/13, A092/13 as Heavy Fuel Oils (HFO), and CON07, CON09 as Other Gas Oils (OGO). These are complex petroleum substances that are products of oil refining that contain hydrocarbons in the range from C9-C25, C13-C30, C20-C50, and C10-C27 respectively (Grimm et al., 2016). Categorization of petroleum

substances is based on the manufacturing process and physico-chemical properties. Petroleum substances are highly complex and may be variable in composition, hence they are excellent examples of UVCBs that present a major challenge in terms of substance identification, classification and grouping. In addition, gaps in available toxicity data create challenges for regulatory decision-making on these substances. In this study, a measure of chemical composition of the reported UVCB substances was derived using Ion Mobility Mass Spectrometry (IM-MS) analysis (Grimm et al., 2017). IM-MS analysis provides chemical fingerprint of substance complexities by yielding the m/z (mass divided by charge number), drift time (time for each ion to traverse within a homogeneous electric field in the ion mobility spectrometer) and abundance. Here, the IM-MS data involves information on the sample-specific heteroatom class distribution based on relative abundance of individual features yielding 82 unique heteroatom classes as features. In addition, the bioactivity characteristics of the selected UVCB substances are derived from *in vitro* models. Specifically, the dimethyl sulfoxide (DMSO)-soluble extracts of UVCB substances are exposed to induced pluripotent stem cell-derived cardiomyocytes and hepatocytes. The resulting concentration-response curves are used to derive ToxPi scores, which serve as the final bioactivity data with 16 features (Grimm et al., 2016).

2.2. Characterization of Broad-Acting Enterosorbents for Mitigation of Chemicals

Several different sorbent materials (e.g., activated carbon and processed sorbent material) are experimentally tested and characterized for various hazardous chemicals. For each sorbent material, equilibrium isothermal analysis has been performed by fitting the experimental data to multiple isotherm equations. This analysis enables the calculation of various material-toxin properties including, binding affinity (K_d), capacity (Q_{max}) and relative surface binding in water. These material-toxin properties are previously found to be the most important features for identifying optimal binding *in vivo* (Phillips, 2008) and can be used to build predictive models which will guide the selection of toxin binders for future testing. In this study, we use the first set of experimental data on the aforementioned sorbent materials with following chemicals: pentachlorophenol (PCP), benzo(a)pyrene (BaP), lindane, diazinon, zearalenone, aldicarb, 2,4-Dinitrophenol (DNP) and aflatoxin (only with processed sorbent material). The data includes information on chemical binding capacity (Q_{max}) and logP, which characterizes the difference in solubility of a solute in two immiscible phases at equilibrium (i.e. octanol and water). As water and octanol being polar and nonpolar solvents, respectively, the logP value provides a measure for the hydrophobicity or hydrophilicity of the substance.

3. Methodology

3.1. Clustering of Complex Substances

In this study, we use hierarchical clustering with 3 distinct agglomeration methods, i.e. Ward's method, average linkage, and complete linkage along with 3 different correlation metrics, namely Kendall, Pearson, and Spearman correlations. This yields 9 different clustering algorithms for testing. Furthermore, an integrated dimensionality reduction and clustering analysis is performed to assess the effect of principal components analysis (PCA) on the clustering results. For each principal component subset, clustering is performed

iteratively using all listed algorithms. The quality of the clustering structure is optimized by maximizing the FM index. FM index provides a measure of similarity between two clustering trees, between clustering trees generated based on 9 aforementioned algorithms and a reference categorization of the complex substances based on their manufacturing streams (i.e. HFO, SRGO, etc.). The optimal clustering tree and corresponding algorithm are reported when the highest FM index is achieved. In future studies, we aim to utilize Support Vector Machine-based feature selection algorithm as an alternative dimensionality reduction technique, which has been successfully applied in bioinformatics (Kieslich et al., 2016).

3.2. Predictive Modeling of Material Sorption Activity on Chemicals

We implement linear and nonlinear regression techniques (Boukouvala and Floudas, 2017) to postulate a surrogate model that will guide iterative experimental design for maximizing the sorption activity of broad-acting materials. Regression is performed by minimizing the least-squares error between the actual experimental values and model predictions. In this study, linear, general quadratic, and signomial function of order 1 are investigated as candidate surrogate models to predict relative sorption level of the material. The goodness-of-fit for each model is assessed using coefficient of determination (R^2), and root-mean-square error (RMSE).

4. Results

4.1. Optimal Clustering of Complex Substances based on Chemical and Biological Data

We present feature intensity maps of 18 tested complex substances using their relative abundance of heteroatom class distributions in Figure 1A. Intensity map highlights the most informative features that separate the substances into categories by providing qualitative and quantitative compositional comparison. The results show that N1 heteroatom class is the most predominant feature with 33.41% presence among all substances. This is followed by N1 S1, O1, N2 S1, and N1 13C1 heteroatom classes with 13.53%, 8.29%, 6.76, and 3.76% occupancy. Relative occupancy of the 5 most descriptive heteroatom classes within each substance category is presented in Table 1.

Among all the clustering algorithms, highest FM index (0.669) is achieved with the integrated PCA - hierarchical clustering with complete linkage using Kendall correlation. The same clustering algorithm without dimensionality reduction (82 features) reveals the FM index of 0.325 (Figure 1A). Similarly, bioactivity profiles obtained from two cell types, namely cardiomyocytes and hepatocytes, are depicted via feature intensity map in Figure 1B. Figure 1B demonstrates that the most distinctive biological features of HFO category result from 4 hepatocyte phenotypes, e.g. mitochondrial integrity intensity, viability, mitochondrial integrity, and total area of live cells. Moreover, the top informative features for VHGO category belong to 2 cardiomyocyte phenotypes, e.g. peak amplitude and mitochondrial integrity. OGO and SRGO categories reveal similar biological features that are mostly part of cardiomyocyte phenotypes, where OGO diverges with the lack of 2 hepatocyte phenotypes, e.g. mitochondrial integrated intensity and total area of live cells, as well as 3 cardiomyocyte phenotypes e.g. mitochondrial integrity, nuclei count, and viability. In clustering analysis, highest FM index (0.805) is achieved with the integrated PCA -

hierarchical clustering with 4 features through complete and average linkage using Spearman correlation. On the other hand, the same clustering algorithm without dimensionality reduction generates FM index of 0.768 with 16 features (Figure 1B).

4.2. Surrogate Modeling for Sorbent Materials

In order to perform predictive modeling of sorption activity of two sorbent materials based on $Q_{\max}$ and $\log P$, we have used 3 different surrogate model types, namely linear, general quadratic, and signomial of order 1. Table 2 and 3 demonstrate the surrogate model parameters along with the corresponding coefficient of determination (R^2), and root-mean-square error (RMSE). Complex surrogate models are avoided to prevent overfitting in the regression analysis. The results show that quadratic functional forms perform better than the others, where the goodness-of-fit should be improved with additional experimental data. The analysis presented here is performed first on mono-constituent chemicals that are anticipated to be part of environmental mixtures and serves as an initial starting point of an iterative data-driven procedure which aims to maximize sorbent activity by providing feedback for future experiment designs. Here, by locating and reporting the sample space spanned by $Q_{\max}$ and $\log P$, where maximum sorption activity is achieved with the best surrogate models (i.e. quadratic functions), we provide feedback to experimental collaborators to design further set of experiments. In the future studies, more chemical properties such as chemisorption indices, specificity, affinity, capacity, and enthalpy will be included in characterization of chemical-sorbent relationship along with designable (morphological and functional) properties of the sorbent material. Then, predictive modeling of the interconnected relationship between the various sorbent properties and sorbent activity will facilitate the exploration of the large chemical-sorbent property space and guide future experimental designs to maximize sorption. Finally, for a mixture of chemicals, % binding can be estimated by analyzing; (i) fraction of each compound in mixture, and (ii) $Q_{\max}$ of each compound in mixture. These will be used to generate predictive models for mixture of chemicals in solution.

5. Conclusions

Here, we present an application of data analysis, modeling, and dimensionality reduction techniques in biomedical and environmental setting. Our analysis provides comprehensive data-driven tools for (i) optimal grouping of an “unknown” mixture with “known” chemicals that enables read-across, thus rapid decision-making during environmental emergencies, and (ii) guide experimental setups to achieve optimal entero-sorbent material designs that can mitigate the adverse effects of the chemical exposure. This research was funded by U.S. National Institute of Health grant P42 ES027704.

References

- Boukouvala F and Floudas CA. Optimization Letters (2017), 11, 895–913.
- CONCAWE, REACH-Analytical Characterisation of Petroleum UVCB Substances (2012).
- Fowlkes EB, Mallows CL. Journal of the American Statistical Association (1983), 78, 553–569.
- Grimm FA, et al. Green Chemistry (2016), 18, 4407–4419. [PubMed: 28035192]
- Grimm FA, et al. Environmental Science & Technology (2017), 51, 7197–7207. [PubMed: 28502166]

Kieslich CA et al. PLoS One (2016), 11, DOI: 10.1371/journal.pone.0148974.
Phillips TD et al. Food Additives & Contaminants Part A (2008), 25, 134–45.
TAMU Superfund Research Center (2017). <https://superfund.tamu.edu/> (accessed 18 Dec 2017).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

(A) Chemical and (B) bioactivity profiling and cluster analysis of UVCB substances based on IM-MS and human induced pluripotent stem cell analysis data set, respectively. Feature intensity maps (informative features are provided in the side bar; features starting with “C-” and “H-” belong to cardiomyocyte and hepatocyte phenotypes, respectively). FM index from clustering studies are provided in bar plots before and after dimensionality reduction.

Table 1.

Identification of the top 5 descriptive features (heteroatom classes) for each category.

Substance Category	Heteroatom Class	Relative Occup. (%)	Substance Category	Heteroatom Class	Relative Occup. (%)
HFO	N1	32.57	VHGO	N1	33.66
	N1 S1	10.03		N1 S1	13.49
	N2 S1	9.32		O1	8.22
	N2	8.40		N2 S1	7.11
	N1 13C1	6.99		N1 O1	3.49
OGO	O1	34.48	SRGO	N1	45.50
	O1 13C1	7.61		N1 S1	20.29
	N2 O1 S2	7.38		N2 S1	6.56
	N1 O1	5.81		O1 S2	6.00
	N2 O1 S1	4.12		N1 13C1	4.24

Table 2.

Surrogate models for sorption activity of activated carbon using $Q_{\max}$ and $\log P$.

Model Type	R ²	RMSE	Surrogate Model Form, y=%bound
Linear	0.472	0.282	$y = 0.070 + 0.761Q_{\max} + 0.210\log P$
Quadratic	0.872	0.139	$y = -3.480 + 11.714Q_{\max} + 7.904\log P$ $-7.044Q_{\max}^2 - 6.188Q_{\max}\log P - 4.787\log P^2$
Signomial Order 1	0.498	0.275	$y = 1.010Q_{\max}^{0.574}\log P^{0.090}$

Table 3.

Surrogate models for sorption activity of processed sorbent material using $Q_{\max}$ and $\log P$.

Model Type	R ²	RMSE	Surrogate Model Form, y=%bound
Linear	0.546	0.204	$y = 0.248 + 0.576Q_{\max} - 0.132 \log P$
Quadratic	0.627	0.185	$y = -0.253 + 1.279Q_{\max} + 1.576\log P$ $-0.269Q_{\max}^2 - 0.923Q_{\max}\log P - 1.121\log P^2$
Signomial Order 1	0.516	0.211	$y = 0.715Q_{\max}^{0.619}\log P^{-0.049}$