Towards characterizing the galaxies of biosolids chemical classes across the chemical universe

Paul Kruse[1,2], Caroline Ring[1]

1. Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA 2. Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA

Paul Kruse | kruse.paul@epa.gov | ORCID: 0000-0001-5516-9717

Abstract number: 478

Introduction

- **Biosolids (treated sewage sludge)** are applied to land or disposed of in landfills
- **Chemicals in biosolids** may enter food or water through agriculture and landfill leaching, or contact humans through other pathways
- Need for **risk-based screening & prioritization of biosolids chemical contaminants** — but data gaps make it difficult
- Plan: develop a high-throughput **machine learning** consensus model to **predict chemical concentrations in biosolids**
  - **Training data:** National Sewage Sludge Survey (NSSS) monitoring data (744 chemicals)
  - **Prediction set:** TSCA Inventory (68k chemicals) [1]
- **Domain of applicability**: How well does NSSS chemical space represent TSCA chemical space (or chemical space of other prediction sets)?
- Eventually, use domain of applicability findings to guide model design and training.

Methods

- **Chemical space** characterized using **ClassyFire & ChemOnt** [2]
  - Structure-based classification
  - "Tree of life" hierarchical ontology: kingdom, superclass, class, subclass, …
- Visualize chemical space using **tree-based visualizations** [3-10]
- **Quantify & visualize similarity** of TSCA and NSSS chemicals
  - **Similarity of ClassyFire classifications**, rather than similarity of structures (Figure 2 & Figure 4)
  - Leverage ChemOnt taxonomy structure to calculate **information content** (Box 1; Figure 5)
  - Calculate established **similarity measures** for tree-based ontologies (Jaccard, Resnik, Lin, Jiang-Conrath) [11-15] (Figure 4, Figure 5)
  - **Heatmap visualization** of similarity [16] (Figure 3)

Discussion and conclusion

- **NSSS appears to be a fairly-representative subset of TSCA** (Figure 1)
- **NSSS is as similar to TSCA as TSCA is to itself**, and more similar to TSCA than to random subtrees of comparable size simulating the TSCA subtree (Table 1)
- Random trees simulating the NSSS subtree are not as similar to the TSCA subtree as the real NSSS subtree is (Table 1)
- Jaccard similarity explores the structure of the tree based on path lengths while the information content-based similarity measures explore the local structure of the tree as well.
- Heatmap: Identify *which* classes represented by the TSCA associated labels are better/worse represented by NSSS associated labels (Figure 3)
- **Tree-based visualizations, similarity measures, and ClassyFire provide useful tools for analyzing the domain of applicability**

References

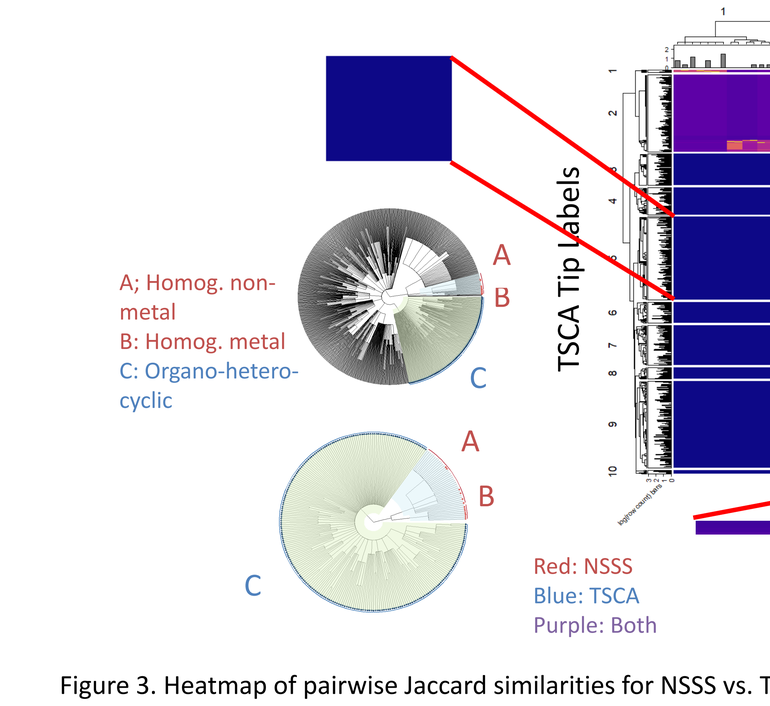


Figure 1. Full ChemOnt tree. Labels color-coded according to representation by classifications associated to TSCA (green), NSSS (orange), both (blue), or neither (gray). Superclasses are labeled (outer arcs).

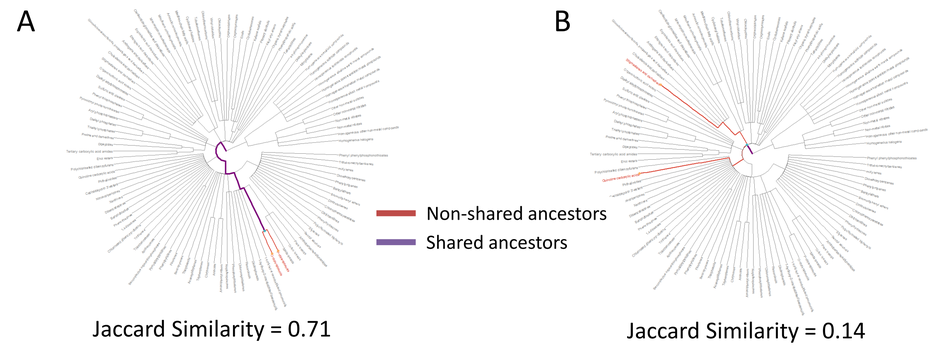


Figure 2. Jaccard similarity of ClassyFire classifications based on "ancestry": sets of labels on each path from root to tip.
A: recent divergence and high similarity.
B: distant divergence and low similarity.

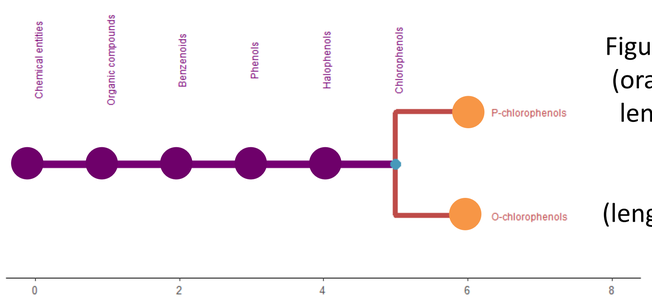


Figure 4. Jaccard similarity of two tips (orange). Shared ancestry (purple) is length 5, while each full ancestry is length 6.
Jaccard similarity =
(length of intersect) / (length of union)
= 5/(6 + 6 – 5) = 5/7 ≈ 0.71

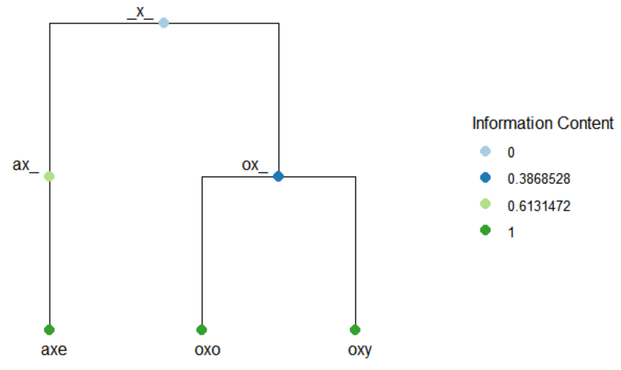
Figure 3. Heatmap of pairwise Jaccard similarities for NSSS vs. TSCA tip labels. Details: Chemical classes represented by two heatmap blocks and by tip dots in tree diagrams. On the left, an example of low similarity; on the right, and example of greater similarity.

Box 1: Information content (IC)

- Quantifies information carried by a label in a tree: **fewer descendants = higher IC** (Figure 5)
- IC(label) = $1 - \frac{\log(1+|desc(label)|)}{\log(N)}$, where $|desc(label)|$ is the number of descendants of the label and N is the total number of nodes and tips in the tree.
- Tip IC = 1; root IC = 0
- Resnik, Lin, and Jiang and Conrath similarity measures all use IC.

Table 1. Average pairwise similarity of labels in Tree1 vs. Tree2: Jaccard, Resnik, Lin, and Jiang and Conrath similarity measures. Random trees: average over n = 100 random trees, each with the same number of tips as NSSS or TSCA.

Tree1	Tree2	Jaccard	Resnik	Lin	Jiang and Conrath
TSCA	TSCA	0.12	0.046	0.050	0.13
NSSS	NSSS	0.17	0.063	0.077	0.24
TSCA	NSSS	0.14	0.046	0.054	0.17
Random "TSCA"	Random "TSCA"	0.12	0.042	0.046	0.11
Random "NSSS"	Random "NSSS"	0.13	0.047	0.055	0.17
TSCA	Random "NSSS"	0.13	0.041	0.046	0.14
NSSS	Random "TSCA"	0.13	0.042	0.048	0.16



←Figure 5. A tree for a three-letter word with 'x' as the second letter. The IC of the node labelled 'ax_' is greater than the information content of the node labelled 'ox_' because it has fewer descendants and thus carries more information about the identity of the word.

Conflict of interest: The authors declare that they have no conflict of interest. ***The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.***
Funding: This poster was supported in part by an appointment to the Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.