



# Linking Chemical And Genomic Databases

Lucas Busta and Edgar B. Cahoon  
 University of Nebraska – Lincoln  
 @PlantsRChemists

Advances in information technology have fueled massive databases for major research fields. However, these databases often do not meet the needs of interdisciplinary research programs that lie at the nexus of major fields.

The rise of genomics has revolutionized natural products research and moved many lines of investigation into the space between chemistry and genomics. Accordingly, an information framework that integrates chemical and genomic data is required.

Here, we find that phylogenetically-resolved metabolic networks integrated with genomic data may identify:

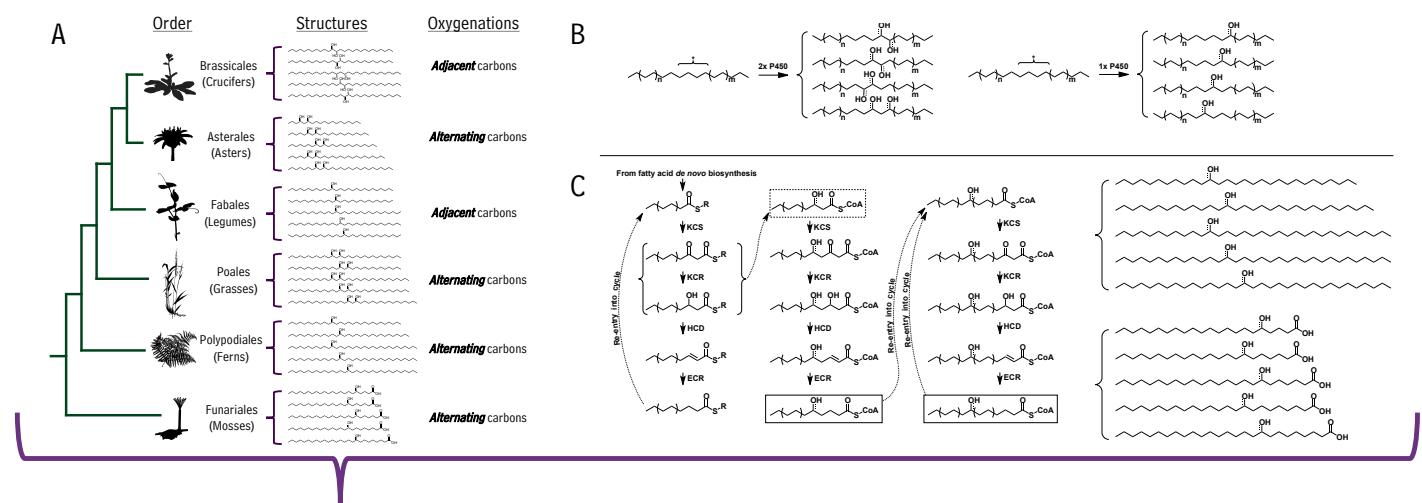
- Genomic hallmarks of medicinal chemistry
- Genetically unexplored lineages of interest
- Novel biochemistries from existing data
- Genetic basis for natural product biosynthesis



Lucas Busta  
 website



This poster  
 as a pdf



1. Biochemical structures encode information, and analyses of diverse structures reveal patterns that enable the extraction of some of this information.
2. In a test analysis of oxygenated very-long-chain fatty acid derivatives, we observed two patterns of oxygenation – alternating and adjacent (A).
3. Based on these patterns, we predicted two distinct modes of secondary group installation: P450-based (B), and discontinuous elongation-based (C). These models were used to predict underlying genes, which were then identified and characterized.
4. This test analysis suggested phylogenetically-resolved chemical structure information as a potential link between chemical and genomic data.
5. To enable phylogenetic analyses of structures on larger scales, we developed a literature and text-mining pipeline using SciFinder.
6. In our pipeline, structure searches are used to create collections of compounds with potentially shared biosynthetic origin. The similarity of these structures is then assessed to construct a putative global metabolic network (D).
7. References for each compound are retrieved and subjected to text-mining to create phylogenetic maps of compound occurrence (E).
8. The metabolic network can then be resolved at any taxonomical level and used and compared with genomic data from the same taxa (F).

