Plant-Microbe Interfaces: Revealing regularities in large datasets with biological annotations using networks and rules of their association

Tatiana Karpinets[1]* (karpinetstv@ornl.gov), Byung Park[2], Guruprasad Kora[2], Mustafa Syed[1], Christopher Schadt[1], Gerald Tuskan[1], Dale Pelletier[1], **Edward Uberbacher**[1], and **Mitchel J. Doktycz**[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [2]Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, TN

http://PMI.ornl.gov

Project Goal: Understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial experimental system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

The expanse of experimental data produced by the Plant-Microbe Interfaces (PMI) project requires the development and application of novel computational tools for their processing. The data, collected from thousands of samples from diverse geographical locations in multiple PMI experimental campaigns, include a complex set of biological annotations, like phenotypic and genotypic characteristics of *Populus*, structure of fungal and bacterial communities in the tree rhizosphere and endosphere, and inherently heterogeneous soil properties. Combined in one table, this mash up of qualitative and quantitative data is challenging to analyze and interpret. To address this challenge we developed a computational framework that expands current approaches to viewing, visualizing, searching, and analyzing information in large databases with biological annotations. The approach is based on two novel concepts, the type-value format and the association network (Anet) that supplement the idea of association rules (Arules) produced by 'Apriori'. Type-value format simplifies computational processing, filtering and grouping of biological annotations by preserving their two-level structural organization in the transaction records and further in Arules and Anets. For annotation types presented by quantitative data, such as genome size or GC content, quantities are replaced with their quality levels based on distributions of the quantities in the dataset records. Association network provides a way to link a diverse set of annotations directly, by the number of transactions where each pair of annotations co-occurred, and indirectly, by considering similarity between profiles of their co-occurrences with other annotations. Monte Carlo simulation is used to assess the significance of similarity by p-value. The resulting Anet is further analyzed and visualized at different levels of resolution, or p-value thresholds, using clustering and visualization techniques. In combination, Anets and Arules provide researchers a powerful tool to create a bird's-eye view of the collected information, to extract hidden biological regularities and to generate hypotheses for further experimental validation.

To test this framework we applied it to the analysis of metadata of sequenced prokaryotic genomes from the GenomeOnLine Database (GOLD). The overlapping structure of the data

provides a good case study for the proposed framework. The generated Anet revealed a hidden structure in metadata of the prokaryotic organisms with three major clusters representing metadata of pathogens, environmental isolates and plant symbionts. The annotations clustered in each group represented a distinct signature profile of metadata for each group and showed a strong link between phenotypic, genomic and environmental features of the organisms.

The Plant Microbe Interfaces Scientific Focus Area is sponsored by the Genomic Science Program, U.S. Department of Energy, Office of Science, Biological and Environmental Research