

# Comparative Genome Content Analysis with Respect to Basic Microbial Phenotypes by Class Association Rule Mining

Makio Tamura<sup>1,2\*</sup>, and Patrik D'haeseleer<sup>1,2</sup>

1. Biosciences and Biotechnology Division, Chemistry, Materials, and Life Sciences Directorate  
Lawrence Livermore National Laboratory, Livermore, CA, USA

2. Biology, Atmosphere, Chemistry, and Earth Division, Computation Directorate  
Lawrence Livermore National Laboratory, Livermore, CA, USA

\*email: tamura2@llnl.gov

## Motivation

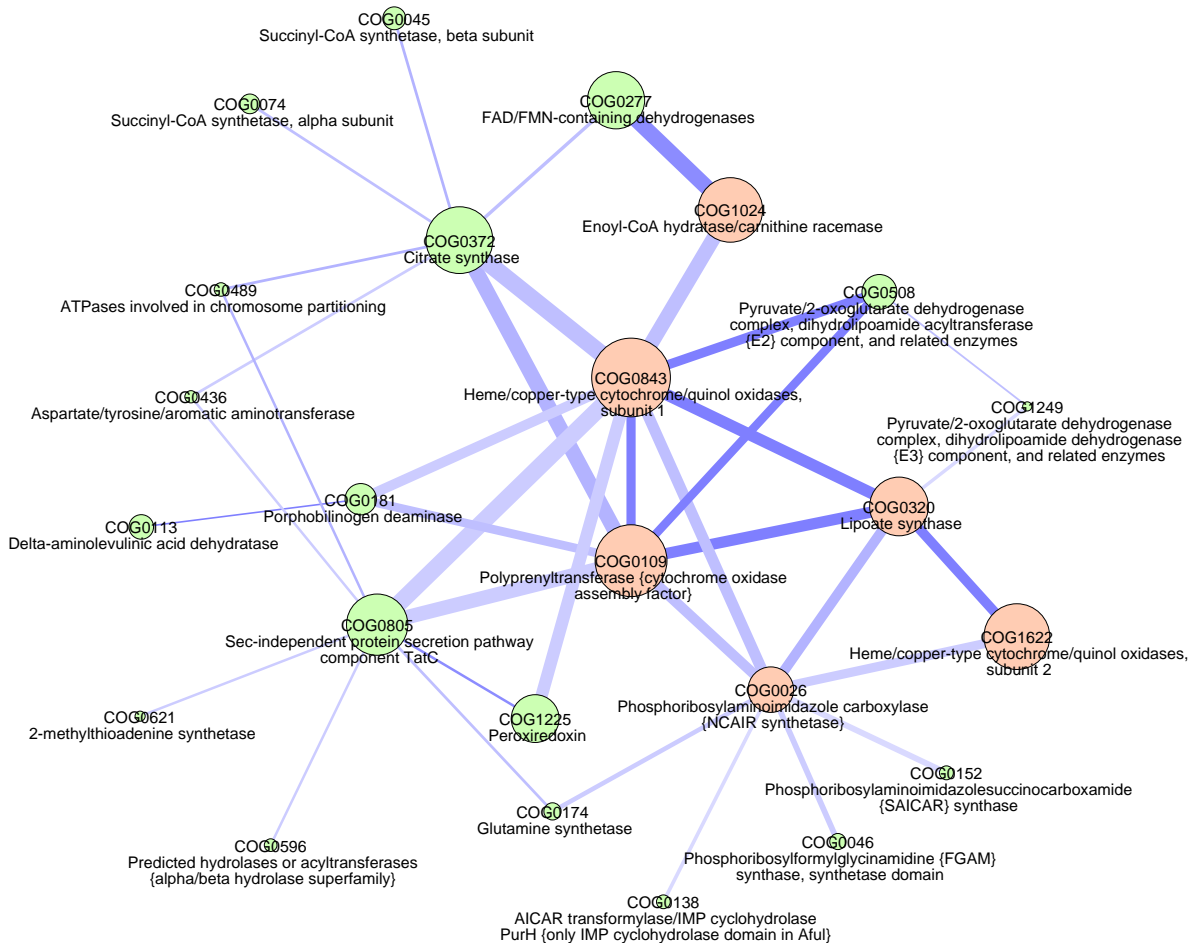
Some microbial phenotypes are closely correlated with a specific gene function, and correlation between the phenotype observation and the corresponding gene presence in the genome may be detected across various organisms. However, microbial phenotypes are typically due to the concerted action of multiple gene functions. Furthermore, the occurrence of corresponding genes may have only weak correlation with the phenotype observation because these genes may be also used in other biological functions depending on interactions with other genes. It may be more appropriate to examine co-occurrence between sets of genes and a phenotype instead of one-to-one relation between genes and the phenotype. Here, we propose an efficient Class Association Rule mining algorithm, NETCAR, to extract sets of genes associated with phenotypes from phylogenetic and phenotype observation profiles. NETCAR takes into account the connectivity graph between genes to restrict hypothesis space, and uses mutual information to evaluate the biconditional relation between sets of genes and phenotypes. In order to evaluate the rules extracted by NETCAR, we compared their predictive capability to those of three alternative algorithms: (1) a standard CAR algorithm, CARapriori, (2) simple sets of genes that have high one-to-one correlation with the phenotype, and (3) Support Vector Machine (SVM), for six microbial phenotypes (aerobic, anaerobic, facultative anaerobic, motility, endospore formation, and Gram staining negativity) across 155 prokaryotic organisms. We also examined the set of COGs used in the extracted rules by contracting a relational network of the COGs for the phenotypes.

## Result

We constructed phylogenetic profiles of Clusters of Orthologous Groups of proteins (COGs), and applied NETCAR to extract sets of COGs associated with the six phenotypes. Based on cross-validation experiments, NETCAR rules show similar or better predictive performance than rules extracted by CARapriori in all phenotypes except for the facultative anaerobic phenotype, and even significantly better than SVM models for endospore formation. This demonstrates that NETCAR can be a powerful tool for extracting sets of genes associated with microbial phenotypes. The rules extracted by the NETCAR contain significant numbers of COGs whose occurrence is only weakly correlated with a phenotype observation, yet phenotype prediction by these rules achieves a higher prediction accuracy than rules combining only highly correlated COGs. And therefore, the NETCAR provides a different perspective for the modularity of genes from previous studies that focused on clustering genes with strong one-to-one relation with phenotypes. A network in Figure 1 shows the relationship among COGs used in the top 30 rules

extracted by NETCAR with respect to the aerobic phenotype. The high frequency of the weakly correlated COGs (green nodes) and strong links to them shown in the network indicates that the gene module can be a combination of genes that span some depth in the network from COGs where we can observe strong one-to-one correlation (orange nodes). The number of fully sequenced microbial genomes has been increasing exponentially over the past decade, and this will increase the power and reliability of computational data-mining methods.

Figure 1: A COG connectivity graph constructed by the extracted rules for the aerobic phenotype. The nodes are COGs involved in the top 30 rules, and edges show that the linked COGs are used in the same rule. The orange nodes are COGs with a strong one-to-one correlation with the phenotype profile while the green nodes represent COGs with a weak one-to-one correlation. The size of each node and the width of each edge are proportional to frequencies of the corresponding COG and link in the extracted rules, respectively. The color intensity of each edge indicates the profile similarity between the linked COGs.



## Acknowledgments

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory (LLNL) under Contract No. W-7405-Eng-48. The project (05-ERD-065) was funded by the Laboratory Directed Research and Development Program at LLNL.