

Comparative Metagenomic Analysis Using Phylogenetic Trees

Paramvir S. Dehal^{1,2*} (PSDehal@lbl.gov), Morgan N. Price^{1,2}, Dylan Chivian^{1,2,3}, Adam P. Arkin^{1,2,3,4}

¹Virtual Institute for Microbial Stress and Survival, <http://vimss.lbl.gov/>; ²Lawrence Berkeley National Laboratory, Berkeley, CA, 94720; ³DOE Joint BioEnergy Institute, Emeryville, CA; and ⁴Department of Bioengineering, University of California, Berkeley, CA, 94720

Acknowledgements

This work was part of the Virtual Institute for Microbial Stress and Survival (<http://VIMSS.lbl.gov>) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

Due to the sheer size of metagenomic datasets, it has been very difficult to compare the gene complements of environments in terms other than high level categorization. Generally, most efforts have focused on comparing metagenomes by determining the relative enrichments of gene family assignments (or functional categories). These methods can give only a very rough idea of how multiple metagenomes differ because gene families are very broad, having multiple subfamilies, and there is not a clear statistical test for significance. Our approach assigns genes to gene families and creates phylogenetic trees of sequences from all metagenomes and sequenced single genomes. Using these trees we can determine if the different metagenomes have significantly different gene families, determine if the differences between metagenomes are within subfamilies of gene family tree, and to determine whether environments or environmental factors associated with the metagenomes are clustering on the gene family trees. Nodes on the phylogenetic trees can be test for significance for both the strength of the association and whether the metagenome was sampled deeply enough to support the node. Here, we describe the results from our analysis of the publicly available metagenomic data with predicted protein sequences. Results of this work will be built into the MicrobesOnline website: www.MicrobesOnline.org.