Characterizing Unknown Viral Genes Through Metabolomics

Tiffany Y. Liang1, Savannah Sanchez2, Jason Rostron2, Jeremy Frank2, Daniel Cuevas3, Daniel Garza2, Anca Segal2, Forest Rohwer2, Robert Edwards2,3

1Bioinformatics and Medical Informatics Research Center, 2Biology Department, 3Computational Science Research Center, San Diego State University, San Diego, CA 92182

Abstract
Viruses are the most diverse biological entities on earth, however, they also have the least characterized genetic, taxonomic, and functional diversity. In metagenomic analyses of viral communities from various environments, most sequences are unrelated to any known sequences; for example, about 90% of the viral sequences found in marine environments are unknown. The goal of this study is to characterize the function of unknown viral genes and identify those that alter host metabolism.

We are building a systematic analysis pipeline that can process metabolomics data for downstream analysis of metabolomics and related data sets.

Generating Metabolomics Data

Metagenome Sequencing
Identify Unknown ORFs
Synthesize ORFs
Express Individual ORF in E.coli

Metabolites
Clones/Genes
28 genes w/ replicates
34% unknowns
417 metabolites
75% unknowns
Together ~12K metabolite-clone pairs

E.coli with an unknown gene
Batch Culture
GC/TOF Mass Spectrometry

Pairwise Correlation Of Abundance Between Known And Unknown Clones

(A) Identify Significant Metabolite-Clone Pairs

(A) Histogram for Z-score of each metabolite-clone pair. Z-score is calculated using log of abundance for all clones in one metabolite. (B) Count of under-abundant metabolites and over-abundant metabolites for each clone.

(B) Correlation based on only the significant metabolites between the pair of clones.

(Pearson correlation of log(abundance) for
(A) Same day, similar clones;
(B) Same day, different clones;
(C) the same clone is ran in GC/TOF MS on two different days;
(D) the same two different sets used in (C), but between different clones.
(E) Overlap of (A) and (B), blue line is the log(abundance) difference from Clone-1 to Clone-2 for the same metabolite.
(F) Histogram of difference in log(abundance) between every 2 clones from different set, using the only control Clone-1 as baseline.

Conclusion

- GC/TOF MS gives huge variation between sets. We need at least one control clone repeated in different sets for normalization between sets.
- Need more unknown clones as references to correlate with unknown clones.

Next Steps

- Identify metabolites significantly altered up or down by expression of a gene.
- Correlate genes of unknown function with genes of known functions by metabolites.
- Automate the analysis pipeline.

Acknowledgements
Thanks for help to Dr. James Nulton, Prof. Barbara Bailey, Prof. Peter Salamon
Funding: NSF, Roche, CRI, Agouron Institute, CIFAR

Contact Information
Tiffany Y. Liang, LiangTiff@gmail.com