Towards a unified framework for analyzing bacteria genomes

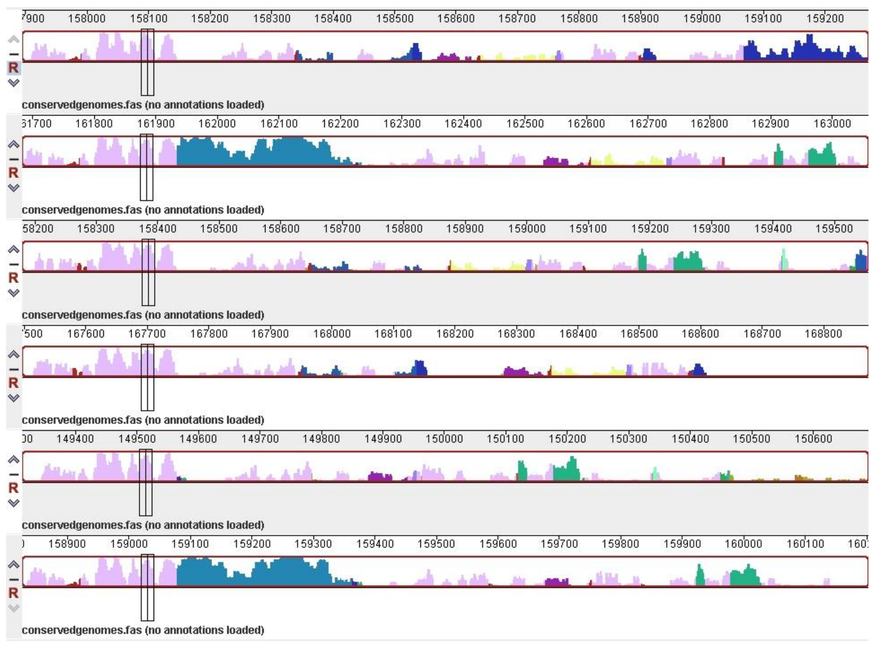
Charles Wigington[1], Alexander Bucksch[1] and Joshua S Weitz[1,2]

[1] School of Biology & [2] School of Physics, Georgia Institute of Technology

http://ecotheory.biology.gatech.edu

Introduction

- Genome evolution in bacteria is a dynamic process of sequence changes through loss and gain of foreign gene material, mutation, deletions, and insertions. Even among closely related organisms such as within a single species of bacteria, we observe large variations in the frequency of the same genes [1].
- Here, the notion of the **pangenome** [2] refers to the common genes between compared bacteria. Within the pan genome we can observe genes of higher occurrence frequency or **core genome**. In contrast, genes of lower occurrence frequency define the **dispensable genome**.
- Typically a frequency pattern results in a U-shape [1,2-5], yet differing in magnitude of the observed U-shape, as a result of grouping similar genes into homologous groups.
- We are investigating the influence of the computational pipeline on the biological interpretation of the result.



The aligned genomes to the right illustrate the concept of the pangenome, the core genome, and the dispensible genome. Mauve areas indicate sequences which are conserved across all genomes analyzed. Teal, blue, green, and other colors indicate sequences which are conserved to a lesser degree across genomes. These non-mauve colors indicate the dispensable genome.

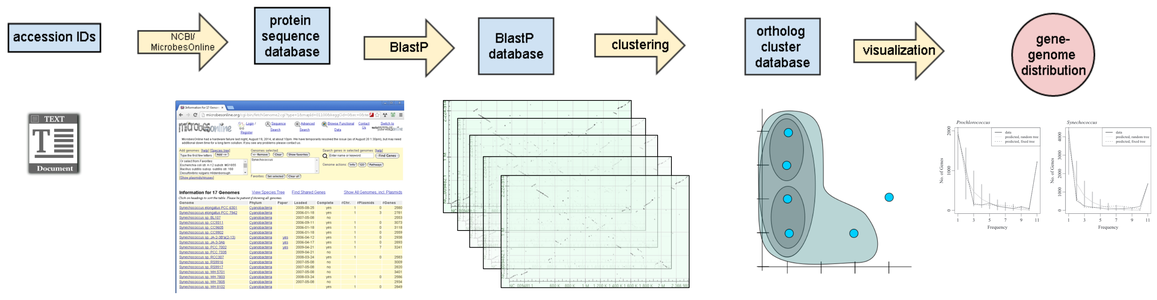
Objectives

- Reproduce existing pipelines for bacteria genome analysis in literature.
- Understand and improve computational elements that led to biological contradictions and differences between different pipelines.
- Develop a pipeline to receive accession IDs of bacteria genomes to robustly constructs an objective U-shape distribution and phylogeny for well studied bacteria species.

Methods

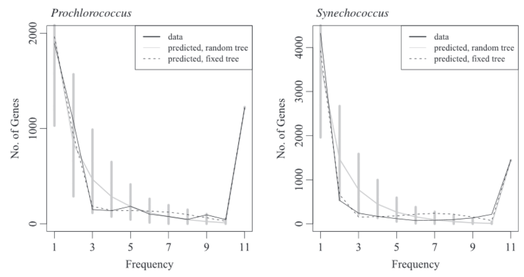
Pipeline

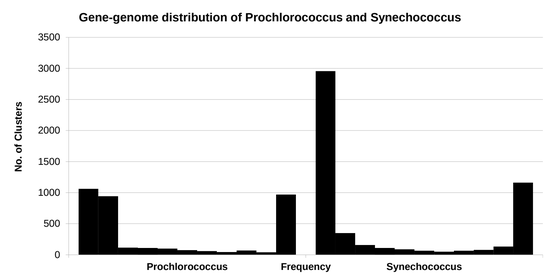
- Protein sequences of 11 *synechococcus* and 11 *prochlorococcus* isolates were accessed from MicrobesOnline.org using accession IDs which mapped to those provided by and Baumdicker [2].
- Preserving the analytic groups from the accession ID sources [6,7], BlastP compared protein sequences to identify those sequences considered to be homologous.
- Homologous genes were clustered using graph theoretic implementations of single-linkage, complete-linkage, and MCL clustering algorithms [8].
- Clusters were interrogated for the genomic source of each member protein sequence.



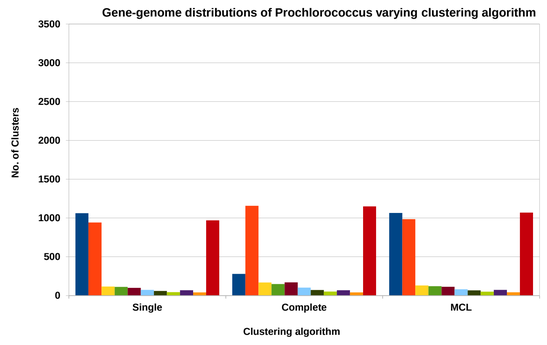
Results

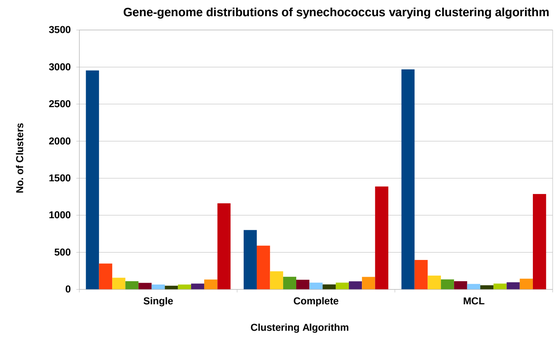
- Despite the use of different clustering algorithms, the recovered gene-genome distributions consistently show the presence of a U-shape, indicating many genes unique to individual genomes, few genes common to multiple genomes, and many core genes conserved across genomes.
- We observed global similarities between the distributions computed in [2,3], but also identified dissimilarities in the magnitudes of individual bars. In the figure below, the second bar on the left differs between Prochlorococcus distributions.
- In reproducing the results of [2] we found obstacles in the choice of the clustering algorithm used to compute U-shapes as well as the choice to exclude paralogs.





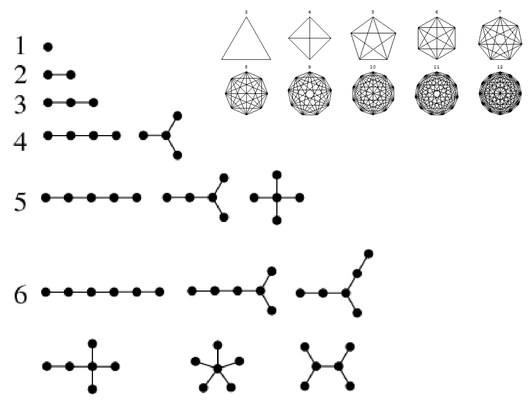
- The observations above indicate that the choice of the clustering algorithm strongly influence the observed U-shape and therefore possible conclusions drawn from it.
- Gene-Genome distributions reflect the core and dispensable genomes of the analyzed prokaryote species.



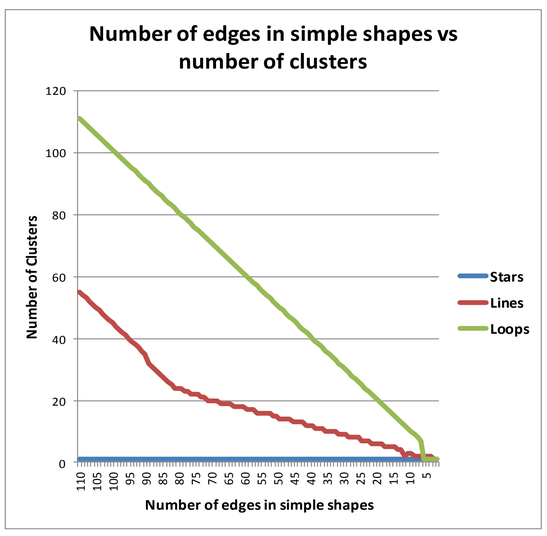


Future Work

- Do common clustering algorithms provide consistent results for elementary motifs at varying scales?



Elementary motifs: Any arbitrary graph can be reconstructed from the motifs linear graph, star graph and loop as a combination at different scales.



- Implement community detection algorithms to test the effect of clustering algorithms which allow for proteins to be members of more than one homolog.
- Investigate alternate algorithms for identifying homologs.

References

[1] A neutral theory of genome evolution and the frequency distribution of genes. Haegeman B. and Weitz J. *BMC Genomics* 2012, 13:196

[2] "Comparative genomics: the bacterial pan-genome". Tettelin H., Riley D., Cattuto C., Medini D. Curr Opin Microbiol. 2008 Oct;11(5):472-7.

[3] "The Infinitely Many Genes Model for the Distributed Genome of Bacteria". Baumdicker F, Hess WR, Pfaffelhuber P. Genome Biol Evol. 2012; 4(4): 443–456.

[4] "Gene Frequency Distributions Reject a Neutral Model of Genome Evolution". Lobkovsky A., Wolf Y., and Koonin E. Genome Biol Evol. 2013; 5(1): 233–242.

[5] "Genomic fluidity: an integrative view of gene diversity within microbial populations". Kislyuk A., Haegeman B., Bergman N., Weitz J. BMC Genomics. 2011; 12: 32.

[6] "Besemer J., Lomsadze A. and Borodovsky M. "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions." *Nucleic Acids Research*, 2001, Vol. 29, No. 12, 2607-2618

[7] "Improved microbial gene identification with GLIMMER" Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Nucleic Acids Res. 1999;27:4636–4641.

[8] Using MCL to extract clusters from networks, in Bacterial Molecular Networks: Methods and Protocols, Methods in Molecular Biology, Vol 804, pages 281—295 (2012).