Final Project Write-Up

Daniella Lato and Jana Taha

25/11/2019

Insert Clever title here

The Data: Our data is biologically based and mostly deals with genome wide trends. We have gene expression data for all genes from four bacterial genomes: *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. This dataset has information about the average expression value of the gene (averaged across multiple datasets) and the genomic location of that gene relative to the origin of replication. Additionally, we have obtained selection information on a few of these genes from each bacterial genome. This selection information tells us about the synonomous substitution rate (dS, mutations that do not cause a change in the amino acid sequence), the non-synonomous substitution rate (dN, mutations that cause a change in the amino acid sequence), and *omega* (dN/dS). The ω ratio allows us to determine if these changes in the sequence cause beneficial or deleterious traits to arise. If ω for a gene is larger than 1, the gene is under positive selection and therefore is beneficial to the organism and will likely be maintained in the genome over time. If ω is less than 1, the gene is under purifying or negative selection, and therefore is deleterious to the organism and will likely not be maintained in the genome over time. If ω is equal to 1, the gene is under neutral selection, and is neither beneficial nor deleterious to the organism. This selection data is again linked to the relative distance from the origin of replication.

Both datasets are looking at how the response variables change with distance from the origin of replication. Near the origin of replication we expect genes to be more conserved and encoding for essential functions than genes located near the terminus of replication. Genes near the origin typically therefore, have higher gene expression and less mutations or substitutions, because they are important to the function of the organism. We expect that most genes (in any genome) are under neutral or purifying selection (removing deleterious traits), regardless of their genomic location (neutral theory or nearly neutral theory). Since genes near the terminus are changing often (mutations) and involved in local environmental adaptation, we could suppose that these genes might be the best candidates for positive selection (increase beneficial traits).

This leaves us with three predictions for our data sets:

1. Gene expression should decrease when moving away from the origin of replication
2. Most genes should be under neutral or purifying selection, any genes that are under positive selection should be located near the terminus