Final Project Proposal

Daniella Lato and Jana Taha

The Data:

We have data from 6 replicons from 4 different species of bacteria: *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. All of the bacteria have their genome contained in one chromosome except *S. meliloti* which is a multirepliconic bacteria. A multirepliconic bacteria means that the genome is made up of multiple replicons or chromosome like structures. For this reason, each replicon of *S. meliloti* (chromosome, pSymA, pSymB) will be analyzed separately. So we effectively have 6 bacterial data sets:

Bacteria	Replicon Name
E. coli	Chromosome
B. subtilis	Chromosome
Streptomyces	Chromosome
S. meliloti	Chromosome
S. meliloti	pSymA
S. meliloti	pSymB

There are three main datasets associated with these bacteria: Substitutions, Gene Expression, and Selection.

Substitutions Data:

The substitutions data set gives information about the number of substitutions (effectively mutations) and the distance from the origin of replication. This data is binary in nature: at each base pair in the genome, there is a substitution present (1) or there is not (0). The data has a phylogenetic component to the analysis and accounts for any substitutions that may also be present in the ancestor of the bacterial strains. Therefore, multiple substitutions may have occurred at a particular base in the genome. The genomic positions in this data set have been scaled to represent base pair distance from the origin of replication, with the furthest distance from the origin of replication being the terminus of replication.

Gene Expression Data:

The gene expression data set has median Counts Per Milltion (CPM) expression values for each gene in the genome. The expression data sets for this analysis were only RNA-seq data sets for control data, where this was defined as the bacteria being grown in environments absent of any stress. Each gene has an associated genomic position (the midpoint between the protein coding start and protein coding end of the gene) which was also scaled to represent base pair distance from the origin of replication.

Selection Data:

The selection data set has information on the non-synonymous (dN) substitution rate, synonymous (dS) substitution rate, and ω (dN / dS) for each gene in the bacterial genomes. This information allows us to make inferences about the selective pressures acting on a gene. This data set contains information about all three of these selective measures for each gene and the associated distance from the origin of replication of that gene. Non-synonymous substitutions cause a change to the amino acid sequence of a gene, which could completely alter the function of the gene and be fatal to the organism. Synonymous substitutions do not alter the amino acid sequence of a gene, and therefore are not expected to significantly impact the function of an organism. 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.

Biological Background

All of the datasets are looking at how the response variables change with distance from the origin of replication. Bacterial replicons begin DNA replication from the origin of replication, and this continues in both directions around the (often) circular chromosome. Some bacteria for example *Streptomyces* have a single linear chromosome that also replicates from a single origin of replication located roughly in the middle of the replicon, and continues in both directions until the ends of the chromosome are reached.

There are certain properties that are believed to be associated with distance from the origin of replication. Replication errors are less frequent near the origin of replication, and as a result, genes that are considered "core" or essential to the survival of the organism are often located in this area. Genes that are more accessory to the function of an organism are therefore located near the terminus of replication (located about half way around the genome in circular bacterial genomes, and at the chromosome ends of linear bacterial genomes). We therefore expect a lower mutation or substitution rate near the origin of replication compared to the terminus. With a lower mutation rate near the origin, we expect genes located in this region to be under purifying or negative selection and removing deleterious traits. Selective pressures at the terminus should be positive, because this region has higher mutation rates and is often recombining, there should be selective pressure to increase beneficial traits. However, we do expect that most genes (in any genome) are under neutral or purifying selection, regardless of their genomic location (neutral theory or nearly neutral theory). These selective pressures and replication processes also impact where genes are located on a bacterial genome. Genes that are essential and highly conserved are generally located near the origin of replication. These genes are additionally often highly expressed because they are so important.

This leaves us with three predictions for our data sets:

1. The number of substitutions should increase when moving away from the origin of

replication

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