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 will be looking at gene expression and selection in four bacterial genomes: *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. The gene expression 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

To begin assessing these predictions we first created some summary graphs of the data.

Gene Expression Data

INSERT FINAL R CODE AND GRAPH HERE FOR GENE EXPRESSION SUMMARY

INSERT EXPLANATION FOR ABOVE GRAPH AND CODE

Selection Data

INSERT FINAL R CODE AND GRAPH HERE FOR SELECTION SUMMARY

When looking at dN and dS substitution rates, we expect that the rate of synonomous substitutions (dS) should be higher than the rate of non-synonomous substitutions. Biologically, mutations that cause a change in an amino acid are more likely to alter the function of the protein than mutations that do not cause a change in an amino acid. As mentioned, a non-functional protein could have catastrophic consequences on the wellbeing of the organism. Across all the bacterial replicons we see that indeed, dS > dN.

We also notice that most of the *omega* values for each bacterial replicon are below 1, this is what we expected. An *omega* value below one means that the genes are likely neutral or under negative selection, meaning that mutations having deleterious impact on the organism will be removed. The notable exception to this is *Streptomyces*, which appears to have a bimodal distribution of *omega* values with a high number of genes with omega values at or above 1. *Streptomyces* creates 80% of the antibiotics that we currently use. This means that the genome of *Streptomyces* would generally benefit from positive selection, where mutations that confer a benefit to the organism are retained.

Since we have some theory about how genes are organized on bacterial genomes, we decided to take a closer look at the selection values for *Streptomyces* and see where these genes fall relative to the origin of replication.

INSERT GRAPH AND R CODE FOR STREP SELECTION GRAPH

This graph shows the mean selection value (dN, dS, or *omega*) calculated over each 10,000 base pairs (bp) region of the *Streptomyces* genome. We observe again that dS is generally higher than dN and most of the *omega* values are less than 1. However, we see that regions of the genome that have an average *omega* value larger than 1 are mostly concentrated near the terminus of the genome. This is reflected in the trend line which is increasing with increasing distance from the origin of replication. Interestingly, the majority of the core and well conserved portion of the *Streptomyces* genome is located in the first ~2 million base pairs (Mbp) near the origin of replication. The rest of the genome is part of the accessory genome which primaraly consists of genes involved in local environmental adaptation and production of antibiotics. It is therefore conceivable that this area of the genome is mostly under positive selection and trying to "hold on" to benificial mutations.

Graphical Decisions

INSERT DECISIONS ABOUT GENE EXPRESSION GRAPHS HERE

For all the graphs we kept a consistant colour scheme so they all look nice together. The selection graphs in particular have the same colours for dN, dS, and *omega* in all graphs so that it is easy for the viewer to follow along. We also wanted to pick colours that were objectively pretty, but also dichromat-friendly. Most of our graphs are based around scatter plots, so the colours needed to be fairly saturated so they would be easy to identify points. Since the values for both the gene expression and the selection data have a wide range of values, we chose to use a log scale to make it easier to read. When considering genomic distance from the origin of replication we scaled the points by 1 million base pairs to make the values on the x-axis more readable. Additionally, all axis labels are clear and have units where applicable. We ensured that greek letters and italic bacteria names were used. The arrangement of bacteria in the facet plots was mostly guided by biological relievance. *E. coli* and *B. subtilis* are the "lab rats", and therefore people often care about them the most, so we put them first. *Streptomyces* is similar to *E. coli* and *B. subtilis* because it has it's genome in one chromosome. *S. meliloti* is a multi-repliconic bacteria (has more than one chromosome-like structure), and therefore we wanted to keep the replicons of this bacteria close together so they could be easily compaired to one another. In the selection summary graphic we decided to add in a redundant legend to aid viewers in determining what colours were linked to which selection measure. We also used direct labeling when we could to avoid the need for a legend. Finally, we utilized trend lines, box-plots, violin plots and reference lines to aid in showing summary statistics and patterns in the data.