Madeline Doak

Anna Ritz

Bio 131

7 May 2018

Project Report

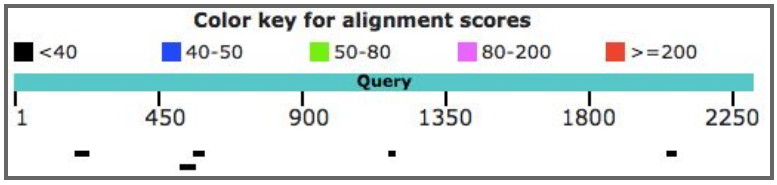
My final project for Bio 131 changed drastically in scope as it progressed. I was initially interested in analyzing the data behind a commonly quoted statistic, that humans and bananas share approximately 60% of our DNA.[1] I was interested in the breakdown of this similarity, the synteny blocks and their location in the genomes of each species. I was planning to perform an analysis in a similar way to the analysis of X chromosomes between humans and mice. However, this proved challenging; I attempted to use a single chromosome against a genome instead, but this still resulted in the BLAST program timing out before completion. I ended up using an analysis of a series of five human housekeeping genes, looking at their local alignments with the banana genome and where these alignments occurred in terms of the gene's coordinates and the banana chromosomes they appeared on. I used NCBI's BLAST program to analyze each of these five genes (NM_000018.3 (acyl-CoA dehydrogenase), NM_000026.3 (adenylosuccinate lyase (ADSL)), NM_000027.3 (aspartylglucosaminidase (AGA)), NM_000033.3 (ATP binding cassette subfamily D member 1 (ABCD1)), and NM_000046.4 (arylsulfatase B (ARSB))) against the banana genome (*Musa acuminata*). I downloaded an Excel file of the BLAST data, pulling the gene coordinates to use in my program. I also manually recorded the banana

[1] http://www.businessinsider.com/comparing-genetic-similarity-between-humans-and-other-things-2016-5

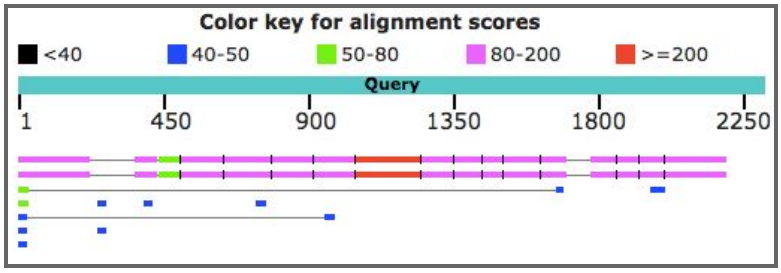
chromosome at which each local alignment occurred. Finally, I recorded the length of the gene to use in the generation of my matrix.

I defined two programs in my repl file, one for generating a graph of a single gene compared to the banana genome, and one for generating a graph of every gene compared to the banana genome, scaled to the longest gene. The first program takes three inputs (an integer for the length of the gene, a list of lists of coordinate pairs for the gene, and a list of chromosome numbers where the local alignments appear). The first step of the program is to add unique chromosomes numbers from the data to the list "chromosomes." This list is then sorted in ascending order. Next, a matrix is generated with a number of rows equal to the number of items in the chromosomes list. The number of columns is set to the length of the gene, from the integer input. Each item in the matrix is set to 0. Next, the program looks at each coordinate pair for the gene, for the entire range of that coordinate pair. For each of these, whichever chromosome N it occurs on, the matrix values at row N, and at the columns equal to the coordinate range, are all set to 1. In the program that combines the genes into a single graph, this last step occurs for each gene, with the matrix set to equal N for the Nth gene it looks at. This combination program also takes different inputs: it takes a list of gene lengths (and uses the longest one to set the number of columns in the matrix), a list of lists of lists of coordinates for each gene, and a list of lists of chromosome numbers. Both of these programs output a matrix, and a list of chromosome numbers to be used as labels for the tick marks on the y-axis. Anna's dot plotter code from Lab 10 was modified slightly to include these labels for the tick marks, and also to take the necessary inputs to label the axes. This code was used to generate the final dot plots.

A problem with this project was the lack of context for these gene comparisons. One could visualize where local alignments occur between human housekeeping genes and similar regions in the banana genome, across different chromosomes. However, this doesn't necessarily answer the question of how similar the human and banana genomes are, especially relative to other species. This was partially answered and contextualized with the use of more BLAST data. I compared the BLAST data for the first gene, NM_000018.3 (acyl-CoA dehydrogenase), between humans and bananas (Fig. 1), and between humans and mice (Fig. 2). The results were



striking; there are significantly more, and significantly higher scored local alignments between humans and mice, compared to humans and bananas. This contextualizes the regions of



similarity between humans and bananas; while the presence of local alignments is interesting, and these alignments are often somewhat well-scoring, these similarities are fairly weak compared to the similarities between humans and more closely related species. Overall, I didn't entirely answer my initial question regarding the data behind the 60% similarity statistic; however, I did provide some visualization of gene similarity between humans and bananas, and I contextualized this using another species. I believe this visualization holds value, in allowing us to understand the degree to which DNA is conserved over time and across species, even when this similarity might not be immediately apparent on the surface.