Homework 4: Finding variable genomic regions

CSCI 5481, Computational Techniques for Genomics
University of Minnesota
Instructor: Dan Knights

Instructions

- Please turn this assignment in on the course site.
- There are multiple files to turn in. All text and code should be placed into a single folder with a name like *lastname_exerciseXX*. The folder should then be compressed and submitted as a single archive (.zip or .tgz)
- You must do this work on your own, although you are encouraged to have general discussions with other students. The work you turn in must be your own. Your code will be checked for overlap and for surprising idiosyncrasies in common with other submissions.
- Please write the names of all students with whom you discussed the assignment at the top of your code.
- Please include copious comments in your code. Full credit will only be given for code that is fully commented, meaning that every line that is not completely obvious needs a comment. Partial credit may be given for broken/non-functioning code if the code is well-commented.
- You may use any programming language you wish.

Background

The purpose of this assignment is to identify variable regions in amplicon sequences, and to compare those results to the conventional wisdom about the locations of variable regions.

Datasets

Download and extract the data: https://canvas.umn.edu/files/16932169/download?download_frd=1.

seqs.fna
File containing a multiple alignment of about 5,000 16S rRNA gene sequences.

Input and Output Format:

This is an analysis project. You do not need to produce a standalone executable program, although you do need to turn in your code (or commandline commands when using commandline tools).

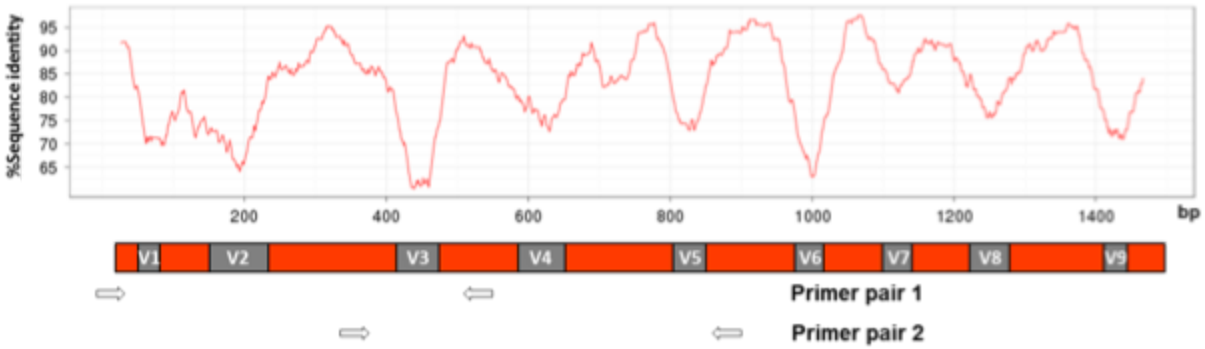
Problems

1. (25 points): Calculate the conservation rate (average identity, or fraction of most common base) at each position in the gapped alignment (1,514 positions). Save the identity values to a text file, one per line. Only a non-gap character (A, C, G, or T) can count as a conserved character, but gaps still count as a non-conserved character. Example: if a position has 10 A's, 5 G's, and 35 gaps, then the most common

non-gap character is “A”, and this base position would be considered 20% conserved (10 / (10 + 5 + 35)). Write this to a file called “solution-problem-1.txt”.

2. (25 points): Plot the variability from step (1) against the position in the gapped alignment. You will need to perform some smoothing on your data before plotting. It is your responsibility to decide on an appropriate approach to smoothing and to describe it in your code comments. You can use a sliding window average or use an external package that does smoothing.

The plot should look somewhat like this plot. Note: this plot is old and based on a small amount of data, and may be using a slightly different approach to calculating conservation (e.g. ignoring gaps in the denominator), and it also includes additional conserved bases on the ends, so do not expect your plot to look exactly like this.

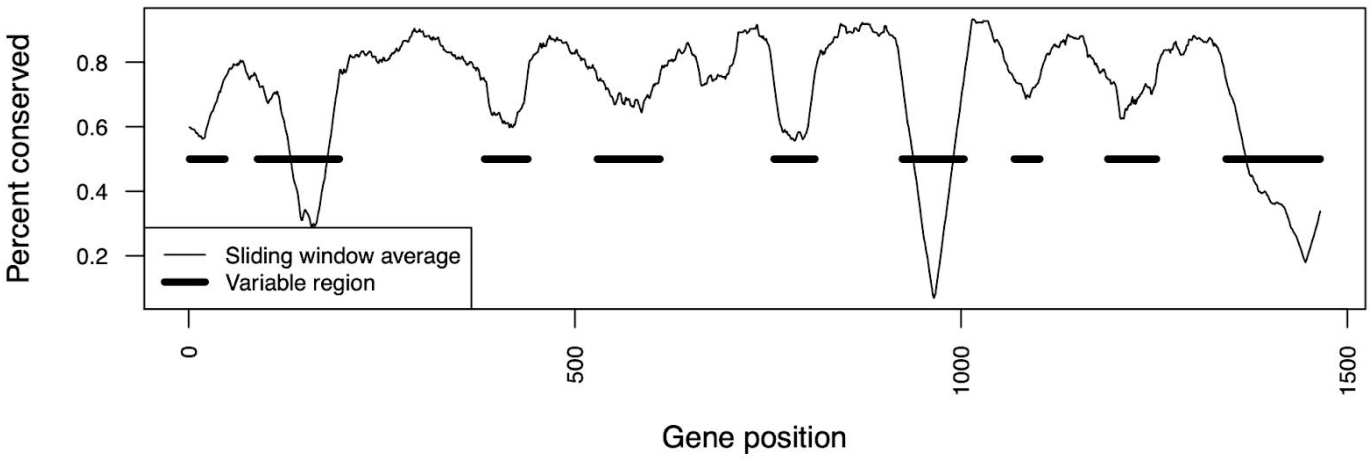


Save this to a file called “solution-problem-2.pdf”.

3. (25 points): Find the variable regions. Most biologists accept that there are 9, but you may find a different number. Use any approach you want, but use code rather than doing it by hand, and justify your answer. Write the start and end coordinates of each v region to a tab-delimited text file with the start in column 1 and the end in column 2. You can use any approach you want as long as it is explained in your code comments.

Save this to a file called “solution-problem-3.txt”.

4. (25 points): Plot the variability again as in step (2) but indicate where your selected variable regions start and end. Here is an example solution plot.



Save this to a file called “solution-problem-4.pdf”.

BONUS. (5 points). Select a random subset of 100 sequences. Extract your longest variable region and your shortest variable region from the DNA file. Use any software you want, including your own, to build and visualize three phylogenies:

1. From your longest variable region
2. From your shortest variable region
3. From the whole 16S.

Show your trees. Can you tell which variable region tree seems closest to the whole-16S tree? Which did you expect? Put all of your visualizations and answers into a single PDF called “solution-BONUS.pdf”.

Deliverables

- Source files (any code that you used for Step 1, 2, 3, 4)
- Readme file explaining how you used your code (text)
- Step 1: File called “solution-problem-1.txt” containing the fraction of conserved bases at each position, one per line.
- Step 2: File called “solution-problem-2.pdf” containing your visualization
- Step 3: File called “solution-problem-3.txt” giving start and end position of each variable region in two tab-delimited columns
- Step 4: File called “solution-problem-4.pdf” containing your visualization
- BONUS: Single PDF with your answers

All files and source code should be added to a folder with your x500 username as the name of the folder. Then zip this folder and upload it on Canvas.