Homework 4: Metagenomics Classifiers

CSCI 5481, Computational Techniques for Genomics
Fall 2018, 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

See “Homework 4 data” folder on the course.

seqs.fna
File containing a multiple alignment of about 200,000 16S rRNA gene sequences.

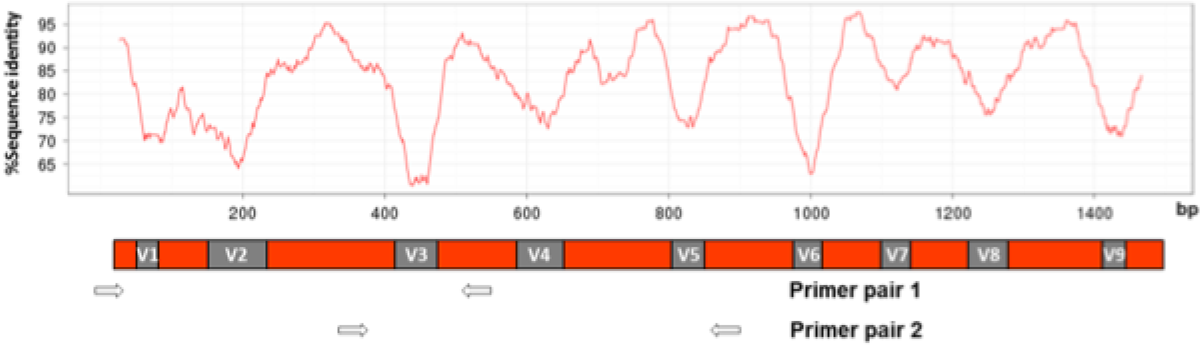
Input and Output Format:

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

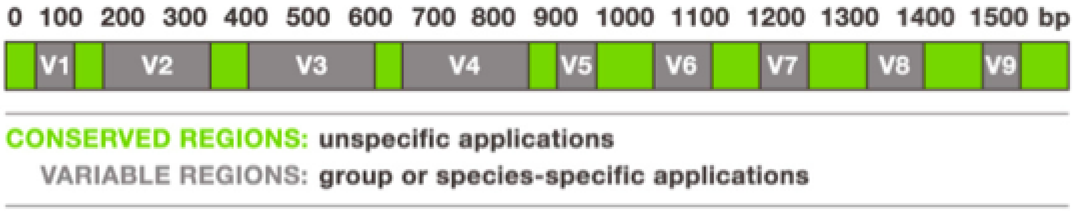
Problems

1. (25 points): Calculate the variability (average identity, or fraction of most common base) at each position in the gapped alignment (1,475 positions). Save the identity values to a text file, one per line. A gap should be treated as non-conserved for the given sequence at the given position. In other words, if a position has 20% A’s, 5% G’s, and 75% gaps, then the position would be considered 20% conserved.

2. (25 points): Plot the variability from step (2) against the position in the gapped alignment. The plot should look somewhat like this plot. 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.



3. (50 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 justify your answer and provide any code used. 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. Here is an approximate guide to the expected lengths and positions of the variable regions, *when excluding gaps* in a representative gene from *E. coli*:


0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 bp
V1
V2
V3
V4
V5
V6
V7
V8
V9
CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

Bonus (10 points). Select a random subset of 100 sequences. Extract variable regions 1 and 4. Use any software you want, including your own, to build and visualize a phylogeny from variable region 1, variable region 4, and the whole 16S. Can you tell which variable region tree seems closest to the whole-16S tree? Which did you expect?

Deliverables

Source files (any code that you used for Step 1, 2, 3, Bonus)
Readme file explaining how you used your code (text)
Step 1: File giving start and end position of each variable region
Step 2: File containing identity values
Step 3: Figure showing plot of identity values.

Bonus: Fasta files containing variable regions 1 and 4. Figures of the V1 tree, V4 tree, and whole-16S tree. Brief text response to questions.

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 Moodle.