**Deliverables:**Source Files: hw4.py, hw4-helper.py, hw4-p4.py, hw3-plot-edges.r, hw3-plot-newick.r, hw3-tip-labels.txt, hw4-100-sequences.fna, Homework4-sequs.fna  
Readme File: This file and the hw4-excel.xlsx file  
Step 1: hw4-variable-regions-indicies.txt  
Step 2: hw4-variability.txt  
Step 3: Shown earlier in this document  
Step 4: hw4-var1.fna, hw4-var2.fna, tree-e-16s.pdf, tree-e-var1.pdf, tree-e-var2.pdf

**HW4: README TO RUN MY SCRIPTS**  
First open a linux based terminal and traverse to the unzipped file that this README is in, it should be something like: /your/root/path/to/my/dir/karlovich\_Homework4/

You will also need python3, and numpy, if they aren’t already installed (I’m pretty sure they should be installed if you’ve run the previous homeworks)

**Problem 1:**Run the command

`python3 hw4.py -i Homework4-seqs.fna`

This will generate the files *hw4-variability.txt* and *hw4-dicitonary.txt*  
- *hw4-dictionary.txt*: A file I used to debug my program  
- *hw4-variability.txt*: The raw variability variables from *Homework4-seqs.fna* before they are smoothed.  
The values in *hw4-variability.txt* are decimal representations of percentages  
 For Example: 0.1 = 10%; 1.0 = 100%; 0.001 = 0.1% etc.  
  
**Problems 2 and 3:**These were done in Excel and are described below  
To import the data into Excel, I used Excel’s Import Data feature on the *hw4-variability.txt*. After importing the data I performed the smoothing and graphing using Excel’s functionality which is explained more later on.

**Problem 4:**Problem 4 requires us to go back the same directory we were in for Problem 1.  
Files:  
- *hw4-100-sequences.fna*: The 100 sequences I randomly chose to use for Problem 4.  
- *hw4-helper.py*: Pulls out the variable regions for V1 and V2 from the *hw4-100-sequences.fna* file based on the indicies indicated in the *hw4-variable-regions-indicies.txt*. This script then puts the variable regions of each sequence into the files *hw4-var1.fna* and *hw4-var2.fna* respectively.  
- *hw4-p4.py*: Generates tree files and edges files from the inputted file.  
- *hw3-plot-edges.r*: The script provided to us in HW3 that I use to build phylogeny trees from edges files.  
- *hw3-plot-newick.r*: The script provided to us in HW3 that I use to build phylogeny trees from tree files.

To run hw4-helper.py, type this into the linux terminal:

`python3 hw4-helper.py -i hw4-100-sequences.fna`

After running this command you should have created two new files: *hw4-var1.fna* and *hw4-var2.fna*

Now we should have all the .fna files we need to create the phylogeny trees:  
First we need to generate the 16S tree so we’ll do that by running the command:

`python3 hw4-p4.py -i hw4-100-sequences.fna`

which will generate the files tree4.tre and edges4.txt

Next to generate the phylogeny tree for Variance Region 1, Type the command in the linux terminal:

`python3 hw4-p4.py -i hw4-var1.fna`

This will generate the files tree4-var1.tre and edges4-var1.txt  
Similarly type the following command to generate the variance region 2:

`python3 hw4-p4.py -i hw4-var2.fna`

and these will generate the files tree4-var2.tre and edges4-var2.txt

*NOTE: the file creation system in hw4-p4.py only works with files that have single digit variance regions identifiers, basically the naming scheme of the output files is based on the last digit before the .fna in the input file, so for example if you ran:*

*`python3 hw4-p4.py -i hw4-var****1****.fna`* and *`python3 hw4-p4.py -i hw4-var****11****.fna`*

*They would both have the same output file since the last digit before the file name extension is a 1. Just something to note if for some reason you wanted to run this with different variance regions, but for this homework it works properly.*  
  
Now all that is left to do is to run either the tree or edge files through the provided R-scripts in order to build the phylogeny trees. The R scripts can be run by running either the tree version:

`Rscript hw3-plot-edges.r [edges\_files.txt] hw3-tip-labels.txt`

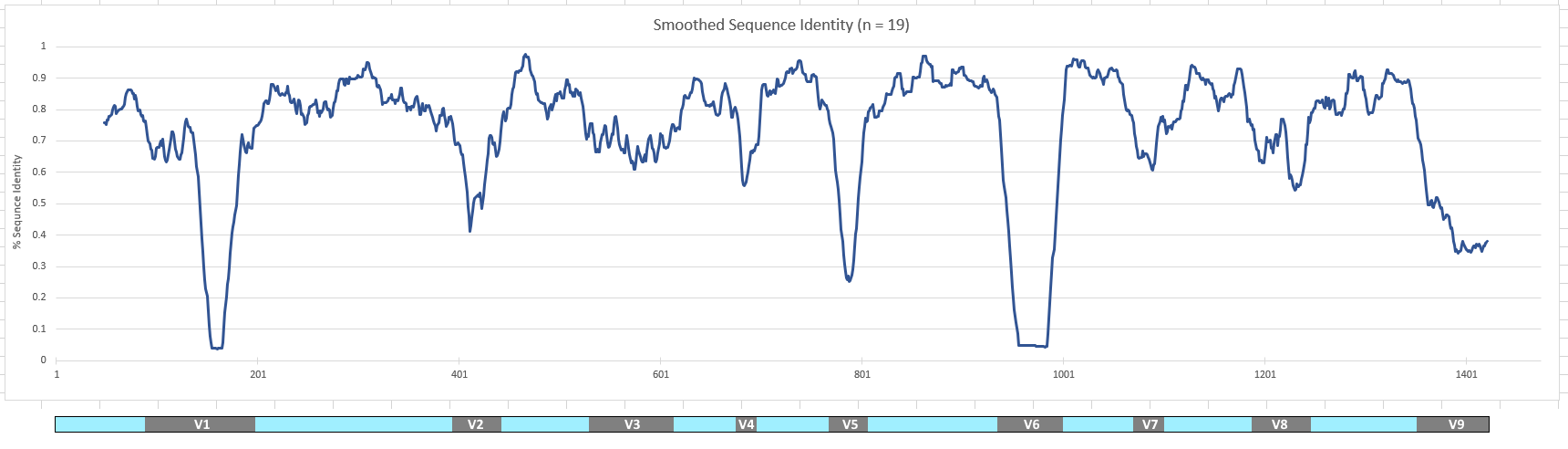
or the version that builds from the .tre file

`Rscript hw3-plot-newick.r [tree\_file.tre] hw3-tip-labels.txt`

where you replace the values in the [] with the .fna file you’re attempting to build the tree for.  
  
*NOTE: the hw3-tip-labels.txt file is only included here so that the R-Script works. I’m not proficient enough in R to remove the parameter requirement. As such, some of the tips will be colored, others will not, but the graphs will still compile properly.*

The output of the Rscript will create a pdf with the filename “tree”. If you want to create multiple phylogny pdf’s you’ll need to manually rename the pdf after running the script for the first time because it will just override the output.

**Data Analysis Write Up:**  
**Problem 1:**  
Didn’t have any Data Analysis, all work was done in the hw4.py file.

**Problem 2:**  
Below you can see a graph of the variability of the sequences. I’ve trimmed the beginning and the end of the sequences as those variabilities were due to high amounts of gap bases rather than actual coded bases and I don’t believe should be counted as “variable regions”. To achieve the smoothing of the graph I took the average values at each index with a balanced window of n = 19. So for example, at base 100, the smoothed value was the average of all the values in the range [91,109] inclusive. The values that were plotted on the graph here are the values in Column C in the excel file *hw4-excel.xlsx*

Conserved Regions are Light Blue Variable Regions are Grey

**Problem 3:**Also shown in the graph above are the conserved and variable regions. From the HW4 handout I knew that there was expected to be about 9 variable regions. I was also advised that I should be looking for a conserved rate between 70 and 80%. To compute what counts as a conserved vs non-conserved region, I computed the average conserved percentage based on a balanced n = 19 window. This is similar to how I smoothed the graph earlier in Problem 2, but this time we are averaging the smoothed sequence values rather than the raw ones. I tested conserved rates between 70 and 85% and found that a conserved rate of 73% gave me the 9 variable regions that I expected and didn’t seem to overestimate or underestimate the variable regions that it did find. The tab-delimited text file: *hw4-variable-regions-indices.txt* has each variable region on a line. The first number is the first base that is in the variable region (inclusive) and the second number is the last base in that variable region (inclusive).  
This file: *hw4-variable-regions-indices.txt* was created based on the formula that is shown in Cell E39 in *hw4-excel.xlsx*. When the average of the balanced n=19 window of the values in Column C respectively were more than the *ConsPercent* then it would return 1 to represent that the region was conserved. When the average of the balanced n=19 window was less than *ConsPercent* then it would return 0.025 to represent that the base was no longer conserved. Then to get the specific indicies where the regions were conserved I took the ranges of regions where there were strings of 0.025’s and built up my variable region from that.

The outputted values by the conserved vs variable regions formula can be found in Column D in *hw4-excel.xlsx*

*Note: ConsPercent is a variable that represents the Cell E3 in hw4-excel.xlsx which in my calculations turned out to be 0.73*

**Problem 4:**I picked variable regions V1 and V2. I ended up picking V2 instead of V4 because V4 seemed like too small of a variable region to compare to the 16S. I used my own software from HW3 in order to compute the phylogeny trees for the 16S, V1, and V2. The V2 region seems to match the 16S tree more closely. Before running this I also expected that the V2 region would closer match the 16S tree than the V1 region. My reasoning behind this was two parts. Part 1, V2 was inherently a shorter variable region so there would be less possible alterations/permutations compared to V1. Part 2, the average region of V2 has a significantly higher sequence identity than V1 (which has a large portion that drops to below 5% (due to a large presence of gaps)) which I hypothesized meant that the variability in V2 was less than that of V1. Based on these two reasonings I believed V2 would more accurately represent the 16S graph. I believe V2 better matches 16S because 16S seems to “spread out” near the center of the graph very quickly which I believe V2 does better than V1 which doesn’t “spread out” as much. V2 and V1 both have one branch that extends very far away from the rest of the tree which matches the 16S phylogeny. Overall it’s still pretty difficult to pick which variable region better represents 16S but I think V2 does a slightly better job.