Ronney Aovida 12/12/16

Final Project - Finding optimal parameters for Trimming and Masking

**Introduction (Biological Background):** In Bioinformatics, Trimming and masking is a process in which alignments are edited per certain parameters to achieve the most optimal output of data from it. The goal of this research is to find the most optimal parameters for this process in order to achieve a more accurate phylogeny.

**Methods:** To conduct the trimming and masking of the data, a few pre-written programs are used. The first important tool used is the TrimAndMaskRawAlignments3, This program takes in 3 variable thresholds that determine the following: X[0,1]= minimum allowable probability of same sites reoccurring, Y[0,20]= minimum amount of good sites (such as GGGG for example), Z[0,96]=missing allowed, which determines how many sites can be missing at a time (the values for Z are more concentrated between 15-40 since most optimal results in previous runs were found in that range). It is these thresholds that this research is oriented towards, by running this program thousands of times over with a different set of parameters each run, the end goal is to then take the data from those runs and analyze them for optimal output. In order to get the data in an organized manner, 3 more programs are used: PlotAlignmentsSummary2- a R script that produces visual plots of different aspects of the alignment, SetupRAxML3- a program that sets up the files needed to run trees and also provides a file that can be uploaded to geneious to be viewed for further inspection of the alignments, and EvalInformativeSitesInPhylipMatrix- a program that provides the results of the trimming and masking. It is the results from the last program that we analyze to determine what thresholds in the first program were optimal.

**Results**: The results provided by the programs are as follows: Number of loci/nGenes, number of taxa, number of sites, variable sites, informative sites, characters, % of missing characters (N’s and –‘s only) and % missing characters all considered).

**In nGenes**: I found that as Z increased, so did the amount of nGenes no matter the Y value. Another observation is as the X value grows alongside the Y value ratio wise, the values become more optimal. However, Y=0 seemed to have no change no matter the given X value.

**In Taxa**: I found that there is no significant change no matter the values given.

**Number of sites**: As the Z value increased so did the number of sites. Most optimal values were found at Y=0 independent of the X value. As X increases, the most optimal values are then found when Y is increased accordingly ratio wise. However its interesting to note that the return values were 0 no matter the Z value at the proceeding Y value after the optimal Y value for every X.

**Number of variable sites**: as Z increases the number of variable sites increases. Y=0 is most optimal independent of the X value. As the X value increases, the Y value that increases at the same rate provides the most optimal results.

**Number of informative sites**: as Z increased, so did the info sites. This provided relatively the same results are the variable sites in regards to the X and Y values.

**Number of total characters**: in this case, values are NOT increasing as Z increases, it is much more chaotic. The most optimal values tend to be high Y values no matter the X values.

**Percent of missing characters (N’s and –‘s only):** as Z increased the amount of missing characters highly increased. Starting at Y=4, the most optimal values are found as Y incremented up with the X value as usual until x=0 or x=2, the optimal value then is found at Y=2.

**Percent of missing characters (all considered):** The results for this were relatively similar to the missing characters only taking N’s and –‘s into account only.

**Conclusion:** After comparing the analysis of all the data independently and finding the most common reoccurrence among them, I found a few distinct patterns. The most prominent being the relationship between X and Y, where if X increases, the most optimal values will be found at the Y value that increased at the same rate as the X. I also noticed for most cases, as the Z value was increased, more data was kept rather than lost. And as a final observation, Y=0 seemed to contain many of the optimal values for each data set regardless of the X value.

(For figures and tables, please refer to the attached excel file)