Aravind Veerappan
BNFO 601
Exam 2
11/07/17

Question 1

a. Assuming that each amino acid has an equal probability of occurring, the maximum amount of information encoded in one column would be $\log_2(20)$, or 4.321928 bits. Since information is the decrease in uncertainty after an observation (information = uncertainty $_{before}$ - uncertainty $_{after}$), and if uncertainty $_{after}$ is equal to 0, then information would equal uncertainty $_{before}$, or $\log_2(M)$.

b. The underlying assumption would be that all the amino acids are equally probable.

c. i) False. A decrease in uncertainty would lead to an increase in information

ii) False. Maximum uncertainty occurs when symbols are equally probable

iii) False. information is gained, or there is less uncertainty, only after an observation is Made

iv) False. Entropy in information theory is the uncertainty before an observation. Uncertainty decreases when information increases

v) True. Less uncertainty after an observation is due to the information gained

Question 2

Output when run with quoll and numbat sequences (word_size=3 and threshold=10):

```
"D:\Program Files (x86)\Python\python.exe" D:/Aravind/Documents/BNFO_601/pycharm_bnfo_601/AravindVeerappan_BLAST_prot.py
Searching for seed DLI at target position 36

Alignment had a score of 18 and is:

Target: 34  AGRDLI
            |  |||
Query:  10  AAPDLI

Searching for seed PIL at target position 36

Alignment had a score of 133 and is:

Target: 12  PDVLVLDIIMPHLDGLAVAAMEAGRPILS
            ||   ||| ||  |||    |  | |||
Query:  12  PDLILLDIMMPGMDGLELGGMDGGKPILT

[(18, 'AAPDLI', 'AGRDLI', 10, 34), (133, 'PDLILLDIMMPGMDGLELGGMDGGKPILT', 'PDVLVLDIIMPHLDGLAVAAMEAGRPILS', 12, 12)]

Process finished with exit code 0
```

Yes it is likely these proteins are orthologs. Even though the alignment isn't perfect, the substitutions that do occur are mainly among amino acids that are found in the same family. For example, changes between valine and leucine are accepted because they share chemical properties, they're hydrophobic. If chemical properties of the sequences are conserved, their

functions would be similar. And since they came from related organisms, it is possible they're orthologous.

The word size and neighborhood threshold affect the sensitivity of the search. Decreasing either of them yields more results.

Question 3

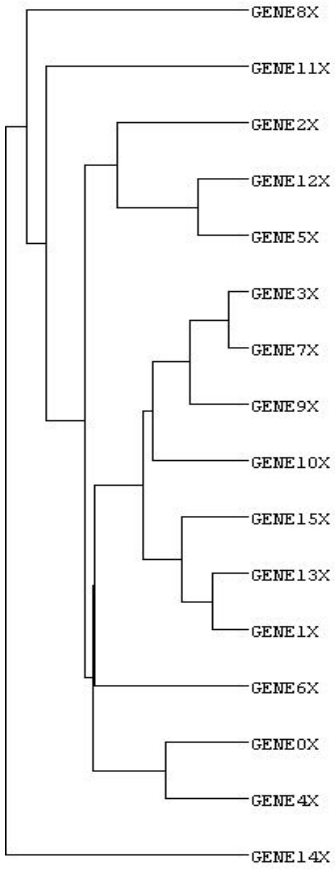
c. Yes the .cdt file works in Java TreeView

d. It is useful to view the same data with different methods because each may provide a unique perspective when analyzing the data. It depends on the data, single linkage is affected more by extreme values while average linkage is not. For example, the shortest distance between two nodes may be due to outliers in the data, which would result in a proxy that doesn't accurately reflect the true distance between the nodes. For this reason average linkage may be a better option.

The complete linkage tree seems to be balanced better. The single linkage tree seems to be more convoluted.

Comparison of the resulting tree views for ratiodata.txt using both single linkage and complete linkage:

Single linkage:



Complete linkage: