EEOB563 – Assignment #2

Dennis Psaroudakis

January 26th 2019

3

GitHub repo is https://github.com/Thyra/EEOB563

4

a) The dinosaur sequence is highly similar to the cloning vector pAgaL as well as some other cloning vectors.

b) The highest rated sequence is erythroid transcription factor from *Gallus gallus*. It is only 96% identical, there are some differences:

```
1   Query   121   MEFVALGGPDAGSPTPFPDeagaflglgggerteaggllaSYPPSGRVSLVPWADTGTLG   300
2                 MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG
3   Sbjct   1     MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG   60

5   Query   301   TPQWVPPATQMEPPHYLEllqpprgspphpssgpllplssgpppCEARECVMARKNCGAT   480
6                 TPQWVPPATQMEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECV     NCGAT
7   Sbjct   61    TPQWVPPATQMEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECV----NCGAT   116

9   Query   481   ATPLWRRDGTGHYLCNWASACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCSHERENCQT   660
10                ATPLWRRDGTGHYLCN    ACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCS    NCQT
11  Sbjct   117   ATPLWRRDGTGHYLCN---ACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCS----NCQT   169

13  Query   661   STTTLWRRSPMGDPVCNNIHACGLYYKLHQVNRPLTMRKDGIQTRNRKVsskgkkrrppg   840
14                STTTLWRRSPMGDPVCN    ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPG
15  Sbjct   170   STTTLWRRSPMGDPVCN---ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPG   226

17  Query   841   ggnpsatagggapmggggdpsmpppppppaaappQSDALYALGPVVLSGHFLPfgnsggf   1020
18                GGNPSATAGGGAPMGGGGDPSMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGF
19  Sbjct   227   GGNPSATAGGGAPMGGGGDPSMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGF   286

21  Query   1021  fgggaggYTAPPGLSPQI   1074
22                FGGGAGGYTAPPGLSPQI
23  Sbjct   287   FGGGAGGYTAPPGLSPQI   304
```

5

These sequences might be homologous (I don't think I can be 100% sure just from the blast result) and if they are, they are more specificially orthologous because they are in different species as opposed to in different genome locations within the same species.

6

I let mafft choose the strategy (auto) and it chose L-INS-i, probably because there were so few sequences and not too long.
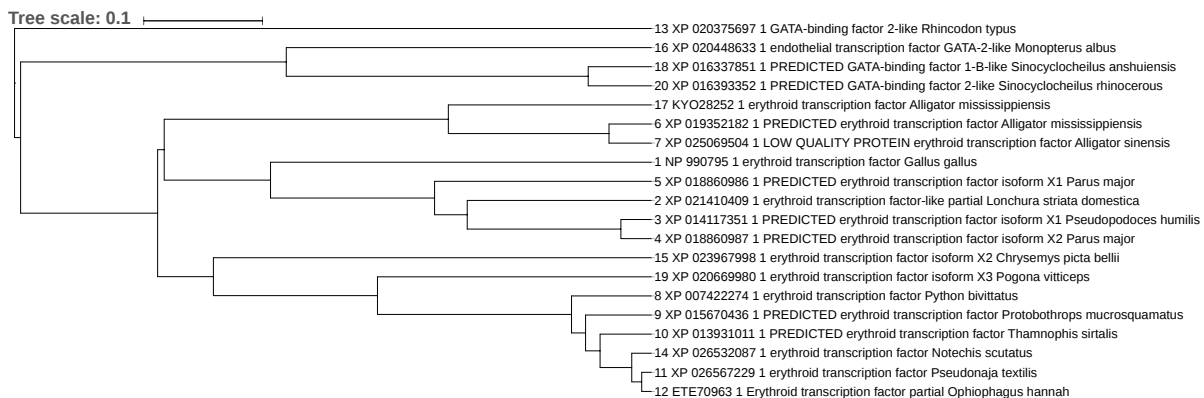

Figure 1: Tree created by mafft

This tree is a guide tree so it is only a rough clustering of the species, not the final phylogenic tree. The actual algorithm to create the phylogenetic tree is run later.