#2: Find genes to discriminate between mammals and fish

Izabela Dusza (70083252)

January 5, 2020

1 Introduction

This short homework project is focused on identification of genes that are most informative to discriminate mammals and fish, using tools for phylogenetic analysis to reconstruct a partial tree of life.

The file attached to the project contained the nucleotide sequences of 20 genes presented in seven species. Six of them belong to mammals, while the seventh belongs to fish. The goal was to determine which of the 20 genes can be used to best distinguish if it is a mammal or a fish.

The project is implemented in Python. The distance matrix for each gene was build and linked with the `linkage` function from `scipy` package. Then, the result was passed to scipy's `dendrogram`.

2 Results

2.1 Description of proposed measure

As a distinguish indicator, I chose a difference between the length of the longest branch (which belongs to fish species in this case), and a maximum length of the rest of the branches (that belongs to others' species, in this case to the mammals) until the last common ancestor.

2.2 Ranking of genes

Ranking of the genes' ability to distinguish between mammals and fish, looks as follow:

1) S100A4	1.0550152046917158
2) INA	0.9651217220362756
3) LGI1	0.8736256814889634
4) SH3KBP1	0.8557548652965721
5) BMP4	0.7596787302896966
6) DLX5	0.7473962846745147
7) GBX2	0.7107237446990087
8) RPS6KA3	0.7005824714927081
9) MRPL21	0.5933626616143963
10) NUP62	0.5808326037925038
11) EXT1	0.5764279289669351
12) DNASE1	0.5709015707136121
13) VAMP2	0.5381595677262943
14) MNS1	0.47944422372434503
15) RPL35	0.2676768633556613
16) RAC2	0.2673804264267681
17) RPL39	0.24226720911177205
18) RPL18A	0.21771628125827397
19) RPL7A	0.19558448550018598
20) UBA52	0.11289706965066121

According to above ranking, S100A4 gene (*S100 calcium-binding protein A4*) is considered the best for distinguish between the species, and the UBA52 gene (*Ubiquitin-52 Amino Acid Fusion Protein*) is considered as the worst one.

Dendrograme of the best gene in terms of species match

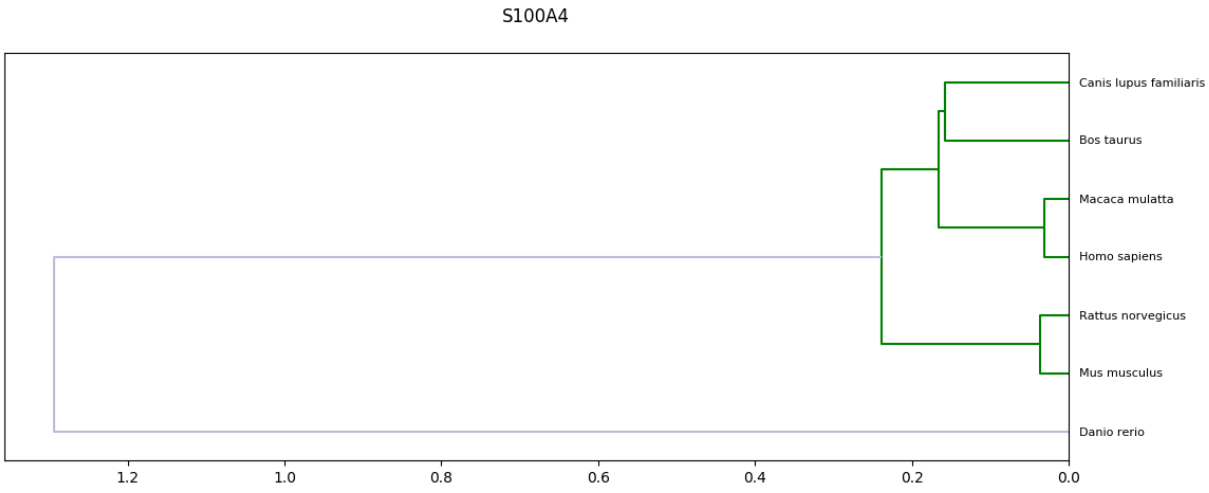

Figure 1: A phylogenetic tree obtained with the best gene

Dendrogram of the worst gene in terms of species match

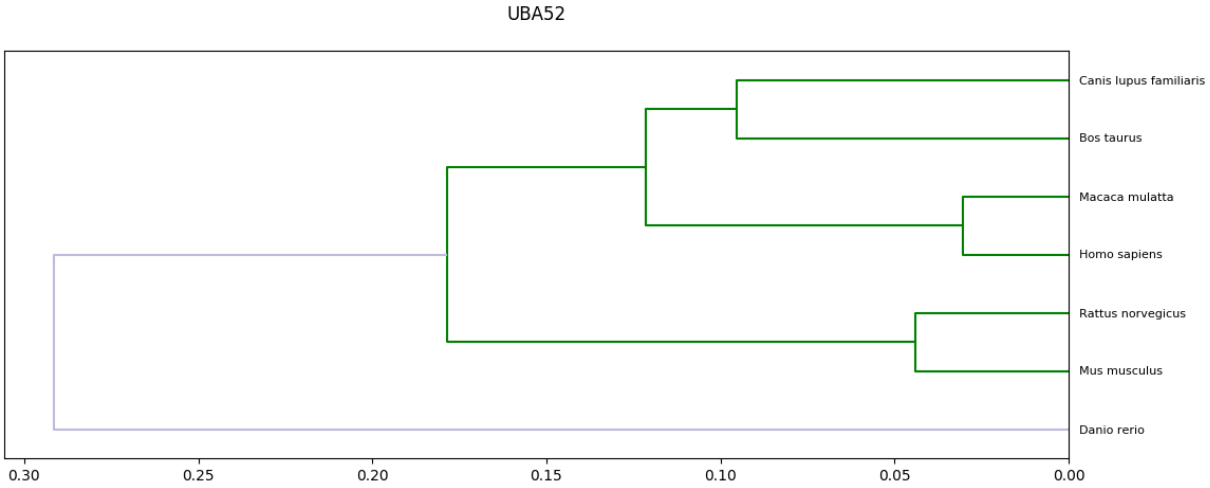

Figure 2: A phylogenetic tree obtained using the least useful gene

2.3 Conclusion

In the project, the best and the worst genes for species recognition were found. Obtained phylogenetic dendrograms, presented in 1 and 2 figures, truly reflects the ranking of genes. In the case of the best gene (fig. 1), the common ancestor of mammal and fish is much further then in the case of the bad distinguishing gene (fig. 2) – the grey branch in the first case is much longer.

To sum up, the best gene for distinguishing one species from the others (e.g. the fish from the group of mammals) is the gene, in which the common ancestor of all species is far away from the outstanding one. The best way to observe it is to create a tree of life, in which the length of branches represents how far the common ancestor is.