Laboratory for Bioinformatics Tools
Assignment 2

Nir Borger

Section b:

FOXP1: A gene encoding for Forkhead box protein P1 protein. The protein is necessary for the proper development of the brain, heart and lung in mammals. It is a transcription factor protein. It has an important role in tissue regulation and cell type-specific gene transcription in both development and adulthood. The gene may act as a tumor suppressor and has an important role in muscle development. Lacking the gene (because of knockout) may cause severe defects in cardiac morphogenesis, which often lead to death. Information courtesy of Wikipedia.

Section C:

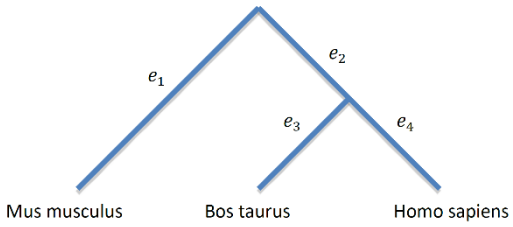
RefSeq Identifiers for the sequences downloaded:
For Homo sapiens: NP_001336267.1
For Bos taurus: NP_001077158.1
For Mus musculus: NP_444432.1

Section D:

Homo sapiens vs. Bos Taurus score is: 1943
Largest insertion/deletion (i.e., indel) is of size 1 amino acid, which is up to 3 bases.

Homo sapiens vs. Mus musculus score is: 1858
Largest insertion/deletion (i.e., indel) is of size 30 amino acids, which are up to 90 bases.

Mus musculus vs. Bos taurus score is: 1808
Largest insertion/deletion (i.e., indel) is of size 30 amino acids, which are up to 90 bases.



Given the tree, it is most likely that the mutation is a deletion mutation and that it had occurred in edge e_2, **or** an insertion mutation in edge e_1. The reason is that in *Mus musculus* there are 30 amino acids that aren't in the other two species, implying either the mouse had an insertion of bases to *FOXP1* gene, or the ancestor of man and cow had a deletion of these very same bases.

Section E:

We will increase the parameter T – The score in which we filter the words of length W. By increasing T, we get less words to compare in the next stages. Since we still must go through all words in the database, once for each word, the preprocessing runtime is not harmed. Increasing T results in less words to expand, because less words will qualify for stage 2, thus improving the algorithm overall runtime.

Section F:

Top 10 results:

Escherichia coli strain YSP8-1 chromosome, complete genome	Escherichia coli	132	132	89%	1e-26	88.60%	4722675	CP037910.1
Escherichia coli strain PJ-T13 chromosome, complete genome	Escherichia coli	128	128	90%	1e-25	87.93%	4935193	CP087110.1
Escherichia coli strain ESY001 chromosome, complete genome	Escherichia coli	128	128	98%	1e-25	86.51%	4833518	CP086342.1
Escherichia coli strain ESY002 chromosome, complete genome	Escherichia coli	128	128	90%	1e-25	87.93%	4832578	CP086556.1
Escherichia coli strain ESY003 chromosome, complete genome	Escherichia coli	128	128	90%	1e-25	87.93%	4830097	CP086340.1
Escherichia coli strain EBJ001 chromosome, complete genome	Escherichia coli	128	128	98%	1e-25	86.51%	4808263	CP086336.1
Escherichia coli strain EJN001 chromosome, complete genome	Escherichia coli	128	128	98%	1e-25	86.51%	4801969	CP086338.1
Escherichia coli strain 1585m1 chromosome, complete genome	Escherichia coli	128	128	90%	1e-25	87.93%	4809054	CP086391.1
Escherichia coli strain elppa4 chromosome, complete genome	Escherichia coli	128	128	98%	1e-25	86.51%	5081390	CP083512.1
Escherichia coli strain elppa8 chromosome, complete genome	Escherichia coli	128	128	90%	1e-25	87.93%	4877488	CP083492.1

The top hit we got is 132, and it belongs to *Escherichia coli*.

Section H:

Total number of STRs: 3509.

Section I:

According to all calculations of this section, the number of 3-length STR's is statistically significant.