**Laboratory for Bioinformatics Tools**

**Assignment 1**

Nir Borger 313580920

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](https://en.wikipedia.org/wiki/FOXP1).

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

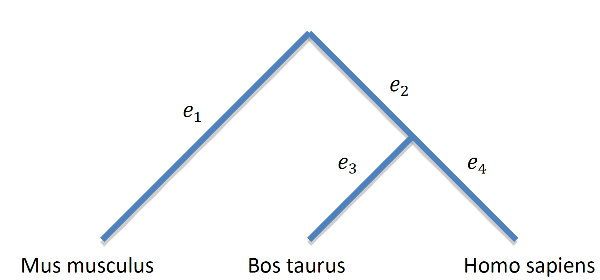
Largest insertion/deletion (i.e., indel) is of size 1 amino acid, which is up to 3 bases.

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

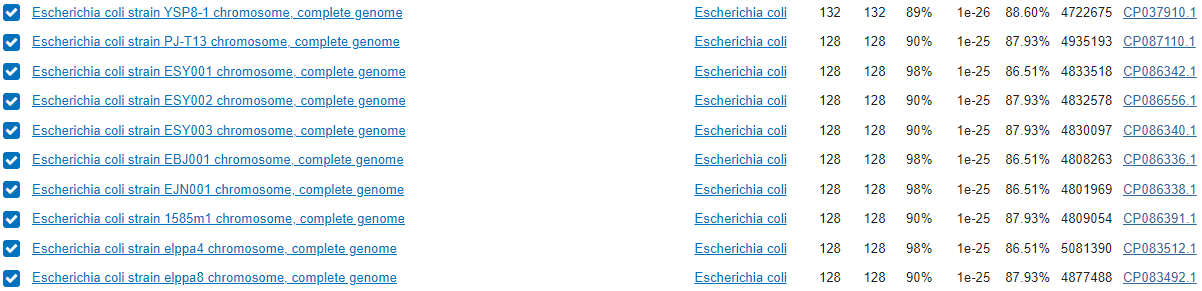
Largest insertion/deletion (i.e., indel) is of size 30 amino acids, which are up to 90 bases.

Given the tree, it is most likely that the mutation is a deletion mutation and that it had occurred in edge , **or** an insertion mutation in edge . 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:  


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.