1. **Introduction** (My introduction is pretty raw, I just used the one from my project idea draft. I still have to work on it, I just included it so you can have an idea what is the disease I am working with which reasons took me to choose this topic).

Osteogenesis Imperfecta is a disease caused by mutations in the genes encoding collagen type I molecule complex. This disorder can have dominant or recessive inheritance; and prevents the collagen from forming an appropriate matrix which leads to a fragile skeletal system. In total, 20 genes are described as causing Osteogenesis Imperfecta. COL1A1 and COL1A2 are responsible for more than 90% of all cases, while genes involved in the background of the recessive forms with relatively high frequency (type VII and VIII) represent less than 10% of the disease (CRTAP).

In this study, the aim is to test whether there is any kind of homology and common ancestor between the twenty genes related with Osteogenesis Imperfecta which could explain why COL1A1 and COL1A2 are critical genes in causing the disease when compared with the others. Also test whether COL1A1 and COL1A2 share similarities with other mammals COL1A1 and COL1A2, and if the mutations that commonly occur in COL1A1 and COL1A2 nucleotide specific sites happen at the same sites in these mammals.

1. **Methods**
   1. **Testing for Homology**

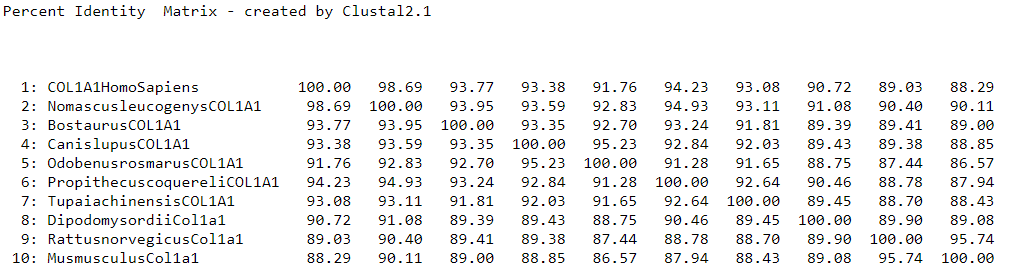
The twenty genes sequences were obtained from the National Center for Biotechnology Information (NCBI) website and <http://www.le.ac.uk/ge/collagen/> which is a databased dedicated to Osteogenesis Imperfecta and Ehlers-Danlos syndrome, where researchers can submit variants found for the genes. The twenty genes are BMP1, IFITMS5, CREB3L1, CRTAP, P4HB, PPIB, FKBP10, SP7, SEC24D, P3H1, SPARC, SERPINF1, COL1A1, COL1A2, SERPINH1, WNT1, MBTPS2, PLS3, TEMEM38B, PLOD2. The sequences were submitted for alignment using MAFFT multiple sequence alignment software. Then, Clustal 2.1 multiple sequence alignment software was also used to provide a percent identity matrix. A rule of 25% of similarity between nucleotide sequences was used to determine whether the sequences were or were not homologous.

* 1. **Building a phylogenetic relationship**

Once the homologous relationship between the twenty genes was proven, the sequences were submitted again to alignment, but changing the outcome file format to phy. The alignment was then used in RAxML software to test for maximum likelihood. A rapid bootstrap followed by the construction of consensus trees were used as means to determine if COL1A1 and COL1A2 have some phylogenetic differentiation from the other genes, since they are responsible for the majority of mutations that result in Osteogenesis Imperfecta.

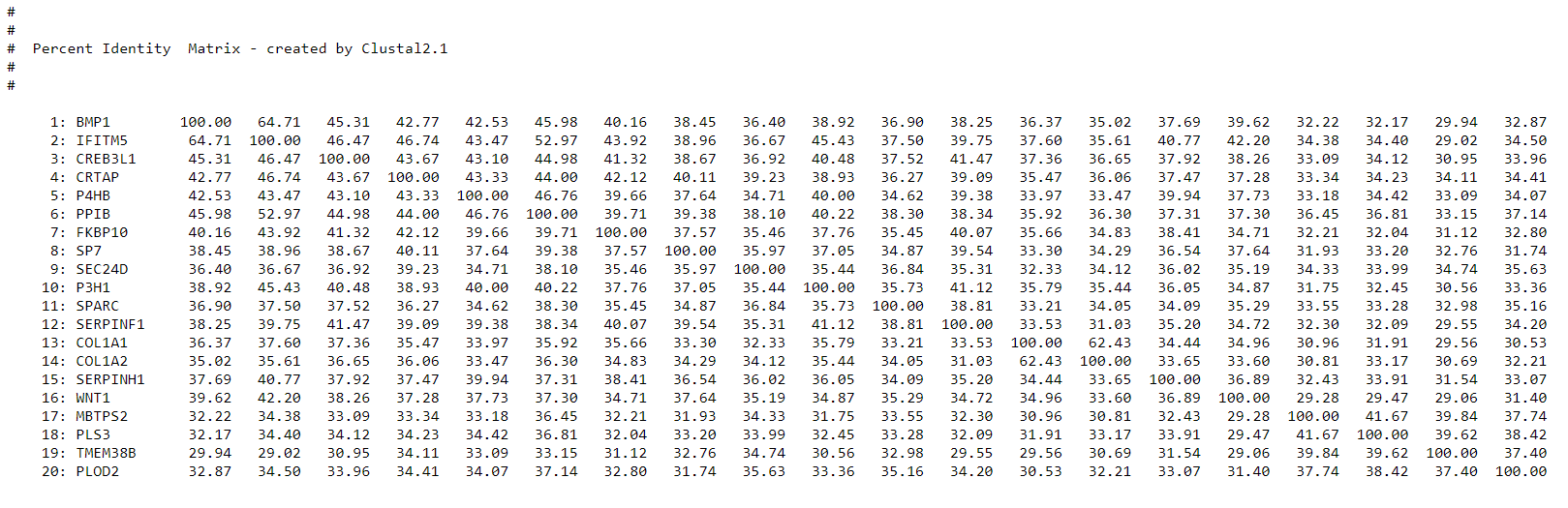
* 1. **Comparing multiple COL1A1 and COL1A2 sequences**

In order to understand how close COL1A1 and COL1A2 genes are, nine different mammal species of COL1A1 and COL1A2 gene sequences were compared with human sequences. The species tested were: *Nomascus leucogenys*, Bos taurus. The sequences were submitted in two trials into MAFFT software followed by CLUSTAL 2.1 to access alignment and percent identity matrix. The aim of this experiment is to test whether other mammals share the same nucleotide as mammals containing the mutations described in COL1A1 and COL1A2, and to see if other mammals could be a potential source of study of mechanisms if these organisms when sharing the same nucleotide in the mutation position, also present mutation cases; or if the same nucleotide in the mutation position is not present, if this change is responsible for the lack of mutation in that specie.



1. **RESULTS AND DISCUSSION**
   1. **Homology connects the genes related with Osteogenesis Imperfecta?**

Based on the percent identity matrix obtained by Clustal 2.1 from the MAFFT sequences alignment, it is possible to infer that the twenty genes related with osteogenesis imperfecta are homologous, since all the alignments attended the criteria of 25% of similarity between nucleotide sequences pre-determined in the methods. As expected COL1A1 and COL1A2 presented one of the higher percentages of similarity (62.43%), after only BMP1 and IFITM5 (64.71). However, it is risky to trust in this homology statement when the similarity rate ranged from 29 to 64%, with most of the sequences being similar in the range of 30-40%. The effort and money investment to approach a specific gene mutation using interpolation from other gene mutations related to Osteogenesis Imperfecta that have been more studied, might not have been worth it.

Fig. 1. Percent Identity Matrix built in Clustal 2.1 after MAFFT sequences alignment of the twenty homologous genes related with Osteogenesis Imperfecta.

* 1. **Phylogenetic relationship**

Fig. 2. Illustrates the majority rule consensus tree built from bootstrap in RAxML of the twenty homologous genes related with Osteogenesis Imperfecta. As expected COL1A1 and COL1A2 are together in the same clade, as well as IFITM5 and BMP1, confirming the Percent Identity Matrix. However, no clear differentiation marks COL1A1 and COL1A2 from the rest of the genes in the tree, indicating that no phylogenetic difference might be the cause of why these two genes cause the majority of the disease cases. COL1A1 and COL1A2 encode for alpha proteins that constitute the helical structure of type I collagen, which could cause a great structural damage when one of them is mutated and might be the cause of why they incur most cases of the disease.

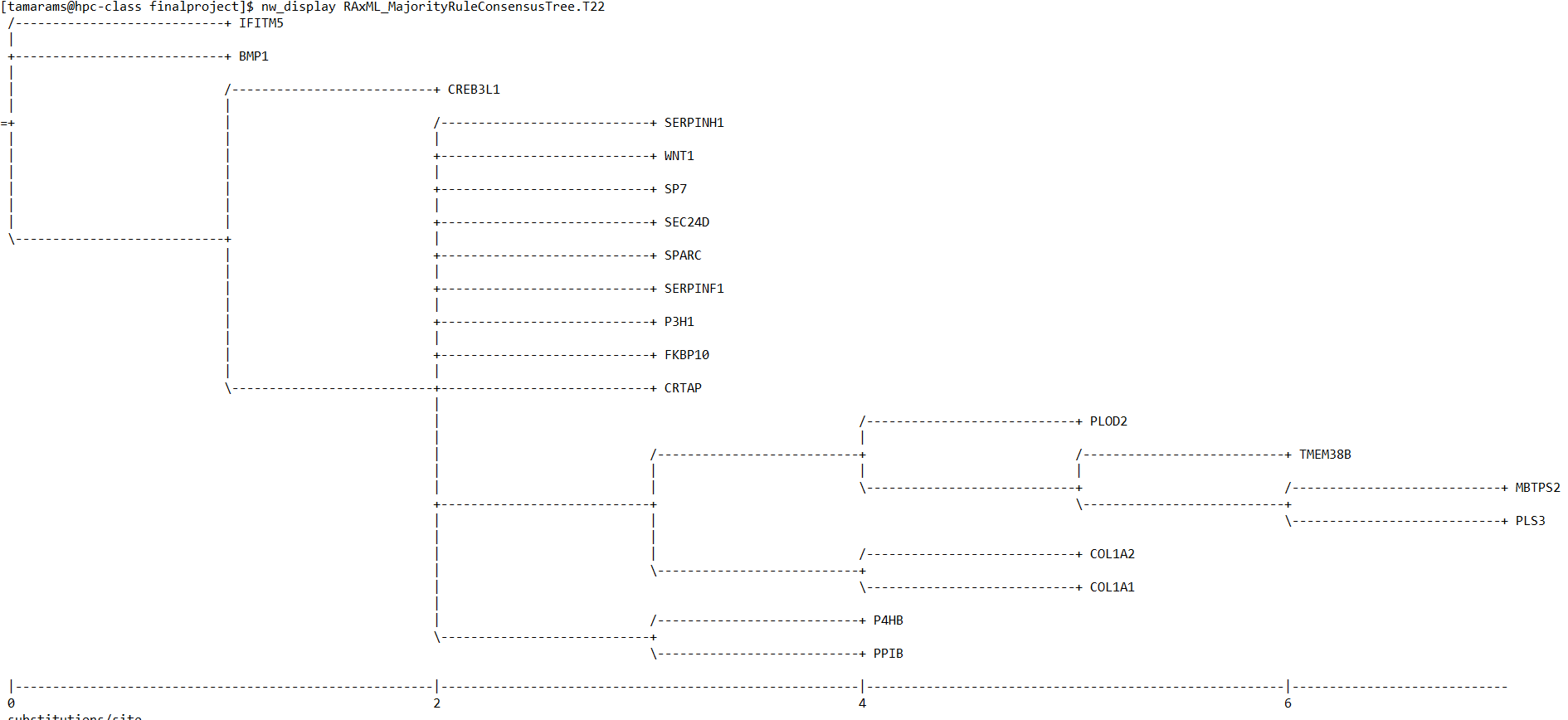


Fig. 2. Majority Rule Consensus Tree built from rapid bootstrap in RAxML of the twenty homologous genes related with Osteogenesis Imperfecta.

**3.3 Mammals COL1A1 and COL1A2**

As we can see in in the percent identity matrix from figure 3, both, COL1A1 and COL1A2 sequences’ alignment from the 9 mammals species were pretty similar with the human genes. COL1A1 similarity ranged from 86.57 to 98.69 which indicates a great homology between the genes. COL1A2 similarly ranged from 84.67 to 98.22, very close to the values found for COL1A1.

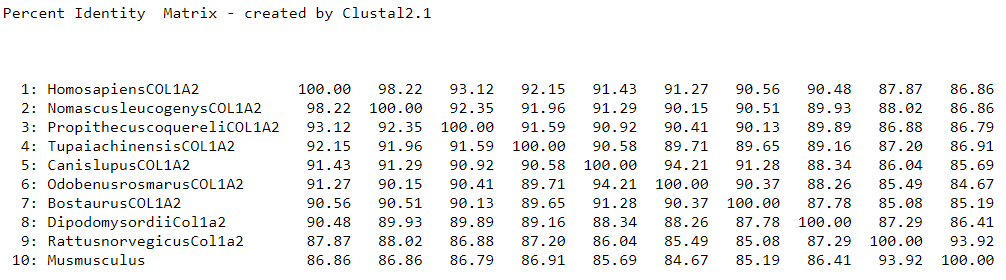
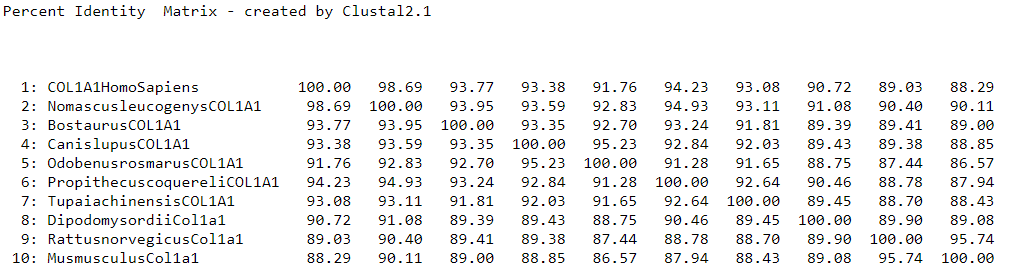
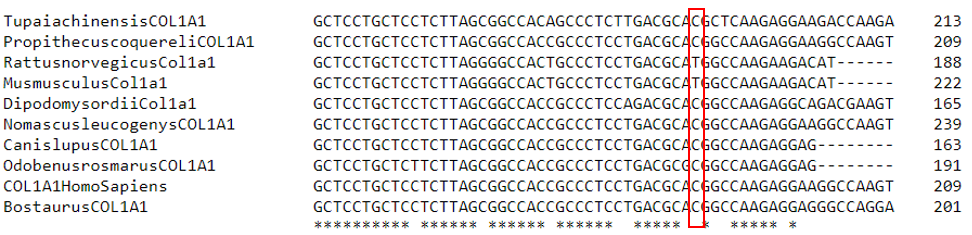
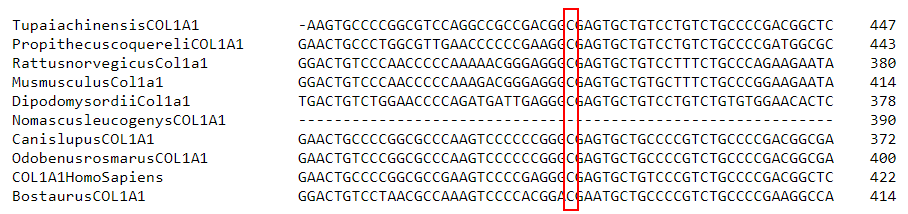
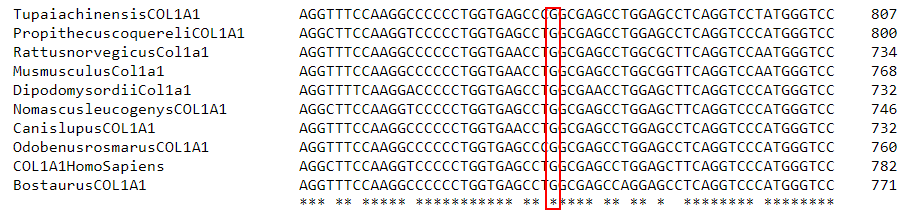


Fig 3. Percent identity matrix built in Clustal 2.1 after MAFFT sequences alignment of ten mammals sequences of COL1A1 and COL1A2.

Arvai et all 2016 list some common nucleotide sites for COL1A1 and COL1A2 mutations. After randomly selecting 3 mutation sites for each, COL1A1 (189, 391, 750) and COL1A2 (246, 811, 2072), the sites were accessed based on the Clustal 2.1 alignment of each of these genes with the other nine mammals (Fig 4).



Ps: I still need to do it for COL1A2, and I am thinking to use the sequences in the snapgene instead to be more precise while looking for the correct mutation site, and to be sure that the sequences match or not in the mutation site, because I think using CLUSTAL is not very precise.