1. **Introduction**

Osteogenesis Imperfecta (OI) is a disease caused by mutations in the genes encoding collagen type I molecule complex1. This disorder can have dominant or recessive inheritance; and prevents the collagen from forming an appropriate matrix which leads to reduced bone mass, increased bone fragility and deformity, and a fragile skeletal system2. In total, 20 genes are described as causing Osteogenesis Imperfecta3. 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)4.

Osteogenesis Imperfecta is a rare disease, occurring in ~1 in 10,000–20,000 births5. To study the mechanisms that cause the mutations to occur and how they influence the collagen encoding/secretion, the use of animal models is a great resource. Unfortunately, most studies are based solely in induced knockout of Mus Musculus genes, with no deep interest in the natural occurrence of the disease in other mammals.

In this study, the aim is to test: (1) 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 other genes associated with OI; (2) whether COL1A1 and COL1A2 are well conserved among other animal species (3) if the mutations that commonly occur in COL1A1 and COL1A2 nucleotide specific sites happen at the same sites in other animals, and if these sites can be used as informative sites for future research.

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

The twenty genes nucleotide sequences (Table 1) were obtained from the National Center for Biotechnology Information (NCBI) website and <http://www.le.ac.uk/ge/collagen/> which is a database dedicated to Osteogenesis Imperfecta and Ehlers-Danlos syndrome, where researchers can submit variants found for the genes. The nucleotide sequences were converted to amino acid using EMBOSS Transeq software7. The amino acid sequences of the genes were submitted for alignment using MAFFT multiple sequence alignment software8. Then, Clustal 2.1 multiple sequence alignment software9 was also used to provide a percent identity matrix. A rule of 25% of similarity between amino acid sequences was used to determine whether the sequences were homologous.

|  |  |  |  |
| --- | --- | --- | --- |
| **GENE** | **ID** | **GENE** | **ID** |
| bone morphogenetic protein 1 | BMP1 | procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2 | PLOD2 |
| collagen, type I, alpha 1 | COL1A1 | plastin 3 | PLS3 |
| collagen, type I, alpha 2 | COL1A2 | peptidylprolyl isomerase B (cyclophilin B) | PPIB |
| cAMP responsive element binding protein 3-like 1 | CREB3L1 | SEC24 family member D | SEC24D |
| cartilage associated protein | CRTAP | serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor) | SERPINF1 |
| FK506 binding protein 10 | FKBP10 | serpin peptidase inhibitor, clade H (heat shock protein 47) | SERPINH1 |
| interferon induced transmembrane protein 5 | IFITM5 | Sp7 transcription factor | SP7 |
| membrane-bound transcription factor peptidase, site 2 | MBTPS2 | secreted protein, acidic, cysteine-rich (osteonectin) | SPARC |
| prolyl 3-hydroxylase 1 | P3H1 | transmembrane protein 38B | TMEM38B |
| prolyl 4-hydroxylase, beta polypeptide | P4HB | wingless-type MMTV integration site family | WNT1 |

Table 1. Description and ID of the twenty genes described as triggers of Osteogenesis Imperfecta. List obtained from the Osteogenesis Imperfecta Variant Database3.

* 1. **Testing for conserved genes**

To test whether COL1A1 and COL1A2 genes are conserved among other species, nine different mammal species of COL1A1 and COL1A2 gene sequences were compared with human sequences. Three outgroup species sequences were also added. The nine mammals species tested were: *Bos taurus, Canis Lupus, Dipodomys ordii, Mus musculus, Nomascus leucogenys, Odobenus rosmarus, Propithecus coquereli, Rattus norvegicus, Tupaia chinensis* (Fig. 1A). The three outgroup species were: *Gekko japonicus, Lepisosteus oculatus, Nipponia nippon* (Fig. 1B) The nucleotide sequences obtained from NCBI6 were submitted into MAFFT software8 followed by CLUSTAL 2.19 to access alignment and percent identity matrix. Mutations in Osteogenesis Imperfecta happen in single base nucleotide, and that is the reason why in this experiment it was preferable to use nucleotide sequences instead of amino acid sequences.



Fig. 1. Species utilized to compare COL1A1 and COL1A2 with the respective human gene. A) Nine mammals species utilized: *Bos taurus, Canis Lupus, Dipodomys ordii, Mus musculus, Nomascus leucogenys, Odobenus rosmarus, Propithecus coquereli, Rattus norvegicus, Tupaia chinensis.* B) Three outgroup species: *Gekko japonicus, Lepisosteus oculatus, Nipponia nippon.*

* 1. **Testing for informative sites**

To access whether other animals can be used as organisms’ models for Osteogenesis Imperfecta, we used the wild type human, same nine mammals plus the three outgroups for test for informative sites to study the disease. We consider here as informative sites, the exactly nucleotide position in the human COL1A1 and COL1A2 gene sequences where previously was described a single nucleotide mutation that triggered Osteogenesis Imperfecta. We obtained the location of three mutations for each one of the genes (COL1A1 and COL1A2) from the ones mentioned in Arvia et al., 2016 work4 (Table 2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **GENE ID** | **Coding variant** | **Protein effect** | **Variant type** | **Position of mutation in this experiment base sequence** |
| COL1A1 | c.189C > A | p.Cys63Ter | Pathogenic | c.325C |
| COL1A1 | c.391C > T | p.Arg131Ter | Pathogenic | c.529C |
| COL1A1 | c.2427C > G | p.(= ).Gly809 | Synonym variant | c.2563C |
| COL1A2 | c.811G > T | p.Gly271Cys | Pathogenic | c.1282G |
| COL1A2 | c.1645C > G | p.Pro549Ala | Missense, non-pathogenic | c.2116C |
| COL1A2 | c.2072G > A | p.Gly691Asp | VUS | c.2543G |

Table 2. List of mutations in COL1A1 and COL1A2. Table based and adapted from the work of Arvia et al., 20164.

The mutations are described in the COL1A1 and COL1A2 coding DNA reference sequence, which is not available in NCBI database, but it is described at Osteogenesis Imperfecta Variant Database3. The version available in NCBI is a slight different version of the gene, and in order to match the correct position of the mutation in the original DNA reference sequence with the one we used in our experiment, we aligned the mutations separately using CLUSTAL 2.19 and mapped the nucleotide position in our sequence for posterior use when comparing with the other animals gene sequence (Fig. 2). From the alignment and using the original coding DNA reference sequence we could stipulate the exact mutation base nucleotide position in our sequence (Table 2, column E).

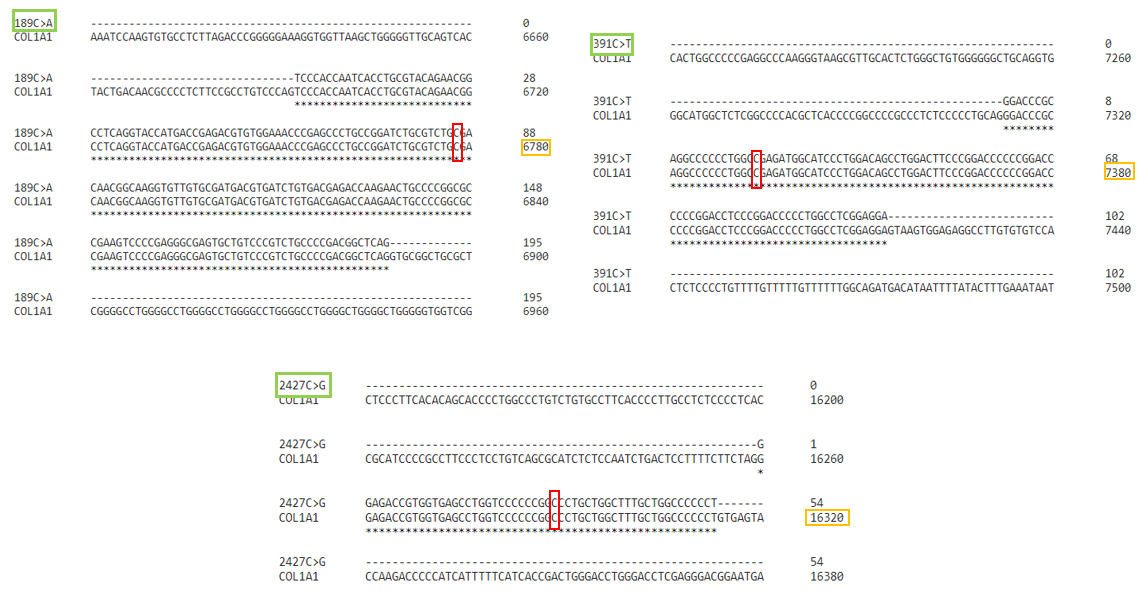


Fig. 2. Example of how in our experiment the NCBI gene sequence was aligned with the DNA reference sequence databank and the exactly base nucleotide position of three different mutation in COL1A1 was found. The green box refers to the mutation nucleotide site position number and for which nucleotide it changes for (ex C>A, C changes to A); red box shows the location of the mutation nucleotide site; orange box shows nucleotide count range for that line. This is just a hypothetical example.

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

Based on the percent identity matrix obtained by Clustal 2.19 from the MAFFT8 sequences alignment, it is possible to infer that the twenty genes related with osteogenesis imperfecta are not homologous, since almost none of the alignments met the criteria of 25% similarity between amino acid sequences pre-determined in the methods (Fig. 3). As expected, only COL1A1 and COL1A2 presented some homologous relationship, with the highest percentage of similarity (30.06%). However, it is risky to trust in this homology statement when the percentage of similarity is very low. 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. This might be the cause of why they incur most cases of the disease, and not because of them being related through any common ancestor. 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 in this case.

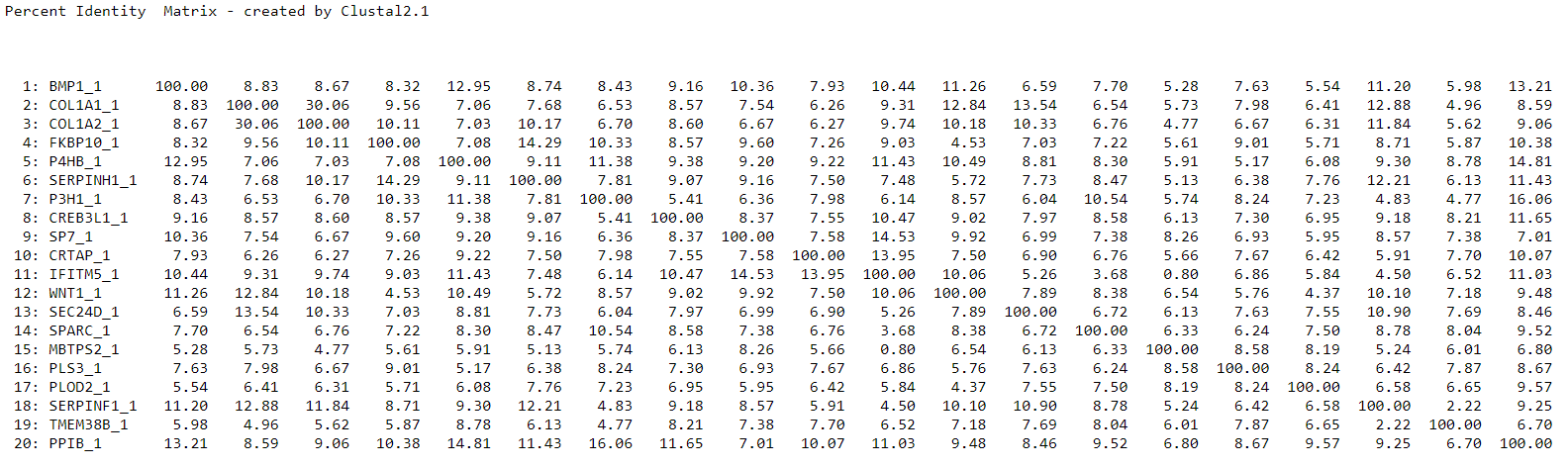
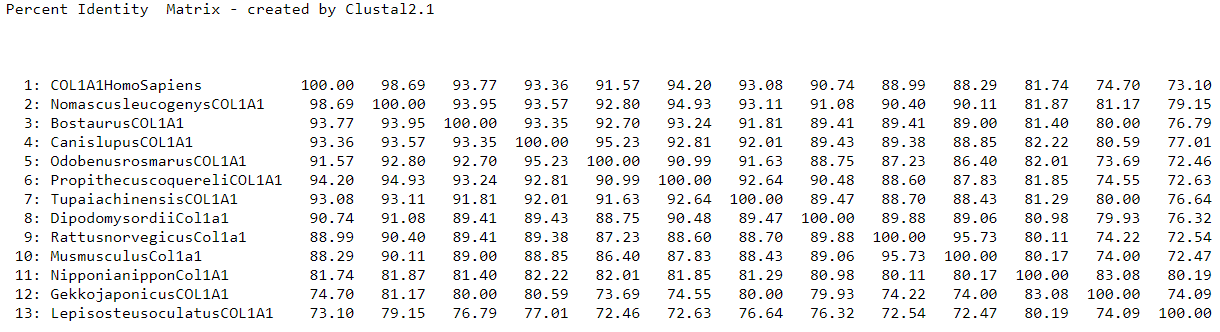


Fig. 3. Percent Identity Matrix built in Clustal 2.19 after MAFFT sequences alignment of the twenty homologous genes related with Osteogenesis Imperfecta.

* 1. **COL1A1 and COL1A2 are conserved among species**

As we can see in in the percent identity matrix from Fig. 4, both, COL1A1 and COL1A2 sequences’ alignment from the 10 mammals species plus 3 outgroup met the homology criteria stipulated (25% similarity). The human COL1A1 and COL1A2 shared more similarities with other mammals species then with other species in general, as was already expected. In general, we can infer from the data that COL1A1 and COL1A2 are well conserved among mammals, once the similarity rates were rarely lower than 85%, and they can be used as a powerful study tool.



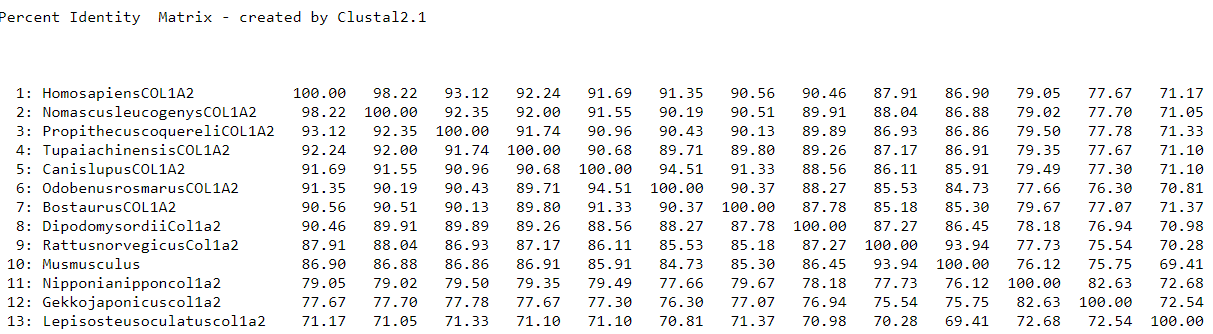


Fig 4. Percent identity matrix built in Clustal 2.1 after MAFFT sequences alignment of ten mammals plus 3 outgroup sequences of COL1A1 and COL1A2.

Even though it was not the focus of the research to show the direct phylogenetic relationship among the different species for COL1A1 and COL1A2, we decided to use RAxML software to find the best tree to describe all 13 organisms tested in this experiment and give more support to our hypothesis that these genes are conserved among species. As we can see in Fig. 5, the tree corroborates the results found by Clustal 2.19 in Fig. 4.

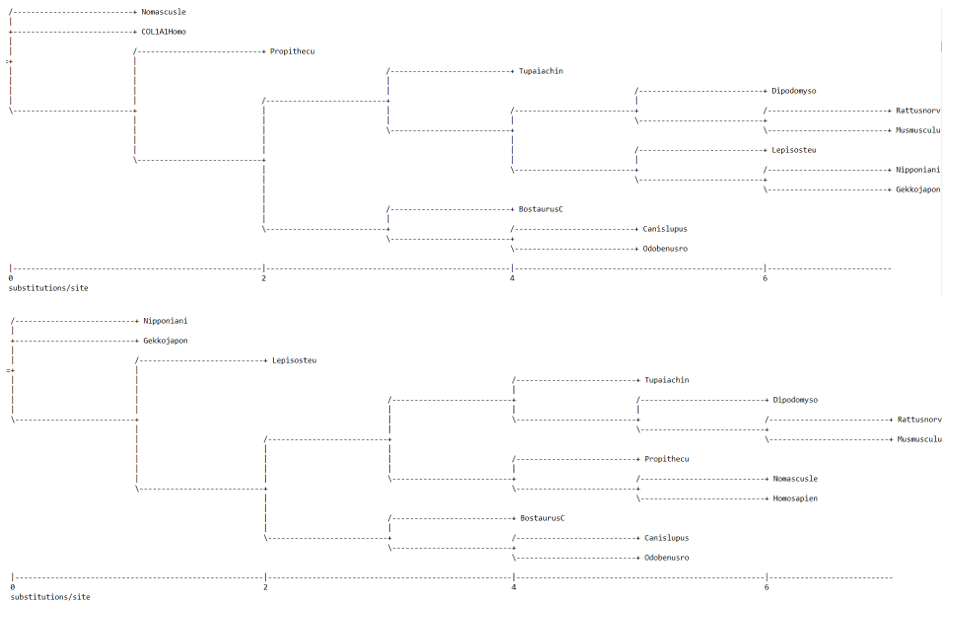
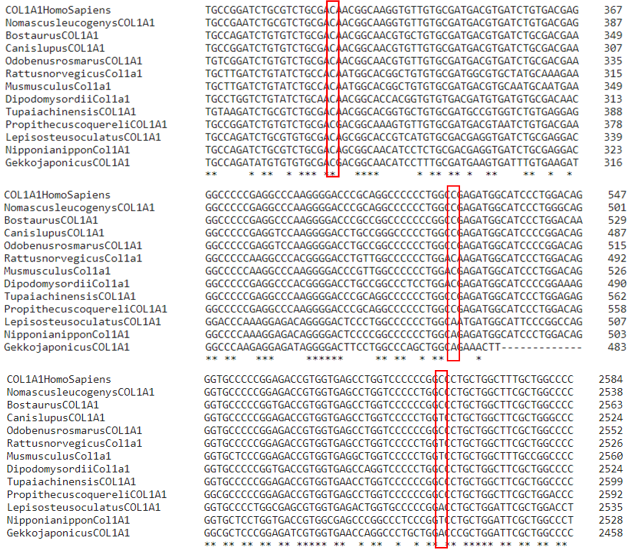


Fig. 5. RAxML tree for COL1A1 (first tree) and COL1A2 (second tree) of 13 different organisms.

* 1. **Informative sites are consistent**

After mapping the informative sites for three distinctive base nucleotide mutations in COL1A1 and COL1A2 genes (Fig. 6) we can notice that they are consistent and bring several options for using other animals, mainly mammals in studying the mechanisms on how single nucleotide mutations can lead to Osteogenesis Imperfecta. Mutation site c.325c in COL1A1 and c.1282G and c.2543G in COL1A2 shows that most of the time, the sites will have the same nucleotides as a result of the genes being conserved among species. These sites are a great opportunity to test whether other mammals or species of animals also suffer from natural Osteogenesis Imperfecta mutations at the same nucleotide position, or if it is an isolated case which happens only in humans. Mutation site c.529C in COL1A1 shows a great example where all mammals present the same nucleotide while non-mammals species present an Adenosine instead of a Cytosine. This example opens up a chance for testing whether the mutation in c.529 position only happens and causes Osteogenesis Imperfecta in individuals that have an Adenosine at that particular spot. Mutation site c.2116C in COL1A2 has only one non-mammal organism with a different nucleotide at the informative site, which allows not only the comparison with mammals such as in c529C, but also with the other non-mammals that present a different nucleotide for the same spot. Mutation site c.2563C in COL1A1 happens to have the occurrence of three different nucleotides and allows for a more complex trial for testing mutation in mammals and non-mammals, with or without the same nucleotide type. In summary, COL1A1 and COL1A2 genes contain mutation sites which confers great information in studying the disease and brings a future opportunity in using other animals besides Mus Musculus as models for understanding Osteogenesis Imperfecta and trying to find a treatment for the disease.



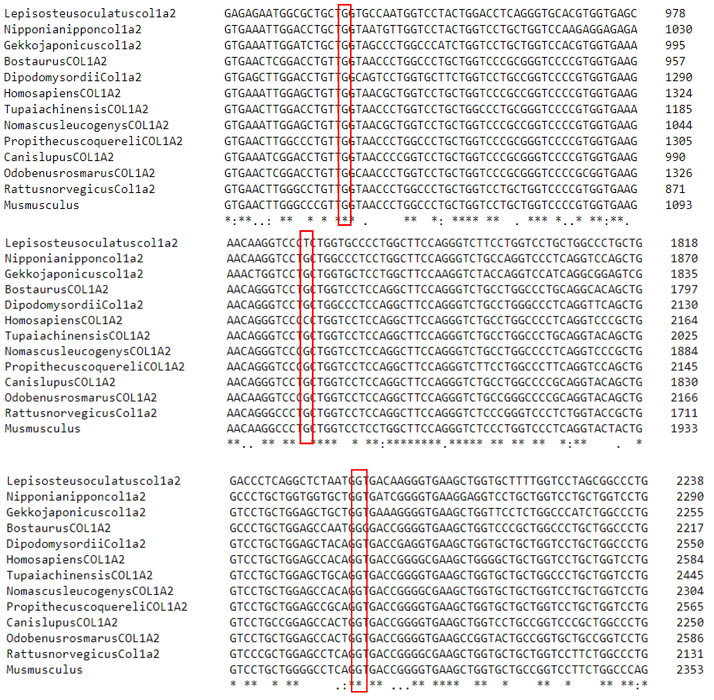


Fig. 6. Informative sites for COL1A1 and COL1A2 gene sequences for three different specific base nucleotide mutation each: COL1A1 (c.325C, c.529C, c.2563C) and COL1A2 (c.1282G, c.2116C, c.2543G).

Here we show that twenty collagen related genes which are known for being related with Osteogenesis Imperfecta do not present any kind of homologous relationship. COL1A1 and COL1A2, which are not paralogs genes, have similar names due to both encode for alpha proteins that are important for the helical structure of collagen. Our results present that these genes are well conserved among several species, mainly mammals. COL1A1 and COL1A2 have specific single nucleotide base mutation sites, which might be used in future experiments while trying to understand how mutations are triggered in the human genome and if the mutations extend to other organisms.

1. **References**
2. Rauch, F. & Glorieux, F. H. Osteogenesis imperfecta. Lancet 363, 1377–85 (2004).
3. Garbes, L., Kim, K., Rieß, A., Hoyer-Kuhn, H., Beleggia, F., Bevot, A., Kim, M.J., Huh, Y.H., Kweon, H.S., Savarirayan, R. and Amor, D., 2015. Mutations in SEC24D, encoding a component of the COPII machinery, cause a syndromic form of osteogenesis imperfecta. The American Journal of Human Genetics, 96(3), pp.432-439.
4. Osteogenesis imperfecta & Ehlers-Danlos syndrome variant databases [https://www.le.ac.uk/ge/collagen/]
5. Árvai, K., Horváth, P., Balla, B., Tobiás, B., Kató, K., Kirschner, G., Klujber, V., Lakatos, P. and Kósa, J.P., 2016. Next-generation sequencing of common osteogenesis imperfecta-related genes in clinical practice. Scientific reports, 6, p.28417.
6. Raunch F, Glorieux FH. Osteogenesis imperfecta. Lancet. 2004;363:1377–1385.
7. National Center for Biotechnology Information (NCBI) [https://www.ncbi.nlm.nih.gov/]
8. EMBOSS Transeq software [https://www.ebi.ac.uk/Tools/st/emboss\_transeq/]
9. MAFFT software [https://www.ebi.ac.uk/Tools/msa/mafft/]
10. Clustal 2.1 software [https://www.ebi.ac.uk/Tools/msa/clustalo/]