The human COL1A1 gene codes for the structural protein, alpha (1) type I collagen chain. The RefSeq curated COL1A1 protein sequence is attached to accession number NP_000079.1. The COL1A1 protein sequence was submitted as a BLASTP query to find homologous proteins. From the results, many possible orthologues were identified. To quantify the likelihood of homology to the COL1A1 protein ten of the possible orthologues were selected to further analyze in the Phylogeny.fr pipeline. The ten possible orthologues, the COL1A1 protein sequence, and an outgroup protein sequence were analyzed through the pipeline multiple times under various parameter modifications to yield the most probable phylogenic tree.

Although local alignments and multiple sequence alignments were previously performed to identify the possible homologs the step was repeated in the Phlogeny.fr pipeline. The possible homologs were identified using the T-COFFEE multiple sequence aligner initially and several times through the pipeline. The best pipeline results were produced with the MUSCLE multiple sequence aligner. Although T-COFFEE is more accurate, MUSCLE is fast. The BLASTP results for the COL1A1 protein query exhibited poor global alignment to the almost 1500 amino acid protein. Many of the hits exhibited conserved domains for a roughly 300 amino acid length. These unique characteristics of the alignments produce aligning difficulty for the more stringent aligners like T-COFFEE. However, possible homologs despite large gaps were already identified and rapid alignment despite these inconsistencies can be performed more easily with MUSCLE.

The FASTA amino acid sequences of the ten possible orthologues, the RefSeq curated COL1A1 protein, and a plant structural protein outgroup were loaded as input to the MUSCLE multiple sequence aligner. Since only proteins produced by vertebrates were identified as possible orthologues in the initial BLASTP, the collagen-like protein from Chlamydomonas

reinhardtii (a type of algae) associated with accession number XP_001697073 was selected to be the outgroup the phylogenic analysis of COL1A1. The algae's collagen-like protein consisted of a similar quantify of conserved amino acids, 387 and while the protein has a similar function to COL1A1 the producing species is most dissimilar to all of the possible orthologous protein producing species. The MUSCLE output consisted of the best alignment given the BLOSUM62 substitution matrix and identified most conserved amino acids with a light blue color. The MUSCLE output only consists of three colors, far fewer than the T-COFFEE output but adequate for the purposes of this analysis. Many of the sequences were of differing length, evident from dashes in place of amino acid letter representatives at the beginning of the alignment. Although the COL1A1 protein is almost 1500 residues, the other proteins only align to roughly 400 residues. Alignment between the proteins was indicated by a similar color at a locus for each sequence. For the 400 residue conserved sequence there is very low variation between the aligned proteins. This high degree of conservation suggests homology or at the very least, similar structure and function.

The next step in the Phylogeny.fr pipeline was curation.