The structural protein product of the human COL1A1 gene, alpha (1) type I collagen chain is sequenced, curated by RefSeq, and attached to accession number NP\_000079.2. This 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 a homologous relationship to the COL1A1 protein, ten of the possible orthologues were selected to be further analyzed 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 with many 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 but the best pipeline results were produced with the MUSCLE multiple sequence aligner. The BLASTP results for the COL1A1 protein query exhibited poor global alignment to the almost 1500 amino acid sequence. Many of the hits exhibited conserved domains for roughly 300 contiguous amino acids. These unique characteristics of the alignments and their biological significance was overly scrutinized by the more stringent aligners like T-COFFEE. These proteins were identified as possible homologs despite these inconsistencies through BLASTP parameter manipulation and were less scrutinized by the MUSCLE multiple sequencer aligner.

The FASTA amino acid sequences of the ten possible orthologues, the RefSeq curated COL1A1 protein sequence, 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 output, the collagen-like protein produced by Chlamydomonas reinhardtii (a type of algae) associated with accession number XP\_001697073 was selected to be the outgroup the phylogenic analysis of the COL1A1 protein. The algae’s collagen-like protein consists of a similar quantity of conserved amino acids, 387 and was not identified by the BLASTP query. The protein has a similar function to the COL1A1 protein but the producing species is biologically dissimilar to the producing species of the ten possible orthologous proteins under observation. MUSCLE produced the best alignment with the BLOSUM62 substitution matrix and identified the most conserved amino acids with a light blue color. The MUSCLE output consisted 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 consists of almost 1500 residues, the other proteins only aligned to roughly 400 residues. Within the conserved sequence, there is very low variation between the proteins. This high degree of conservation suggests homology or at the very least, similar structure and function.

Partial MUSCLE Output:

[gi|algea|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) rvdsYqahearqvad----qladEqRHVsl-------------------FaYGvGrGvDr

[gi|seaturt](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQattAQKNWYISKNPKEKkHVWFGEtMsDGFQ----------FEYG-GEGSnP

[gi|Alligat](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVhPTQatIAQKNWYmSKNPKEKkHiWFGEtMsDGFQ----------FEYG-GEGSnP

[gi|Cricetu](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) qTCVfPTQPtVpQKNWYISpNPKEKeHVWFGESMTDGFQ----------FEYG-sEGSDP

[gi|shrew|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQPSVAkKNWYvSKN-KdKRHVWFGESMThGFQvltrsslfssFpcs-ssdSDP

[gi|Brandtb](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQPtVAQKNWYISKNPKEKkHVWFGESMTgGFQ----------FEYG-GqdSDP

[gi|mouse|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) qTCVfPTQPSVpQKNWYISpNPKEKkHVWFGESMTDGFp----------FEYG-sEGSDP

[gi|rat|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) qTCVfPTQPSVpQKNWYISpNPKEKkHVWFGESMTDGFQ----------FEYG-sEGSDP

[gi|Donkey|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQPqVAQKNWYISKNPKdKRHVWyGESMTDGFQ----------FEYG-GqGSDP

[gi|HumanCO](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQPSVAQKNWYISKNPKdKRHVWFGESMTDGFQ----------FEYG-GqGSDP

[gi|Cow|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQPSVAQKNWYISKNPKEKRHVWyGESMTgGFQ----------FEYG-GqGSDP

[gi|Doggie|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) ETCVYPTQPqVAQKNWYISKNPKEKRHVWyGESMTDGFQ----------FEYG-GqGSDP

[gi|algea|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) nvvflvdgsGsvnaeefeamlgFcvdasnqlAesvpnl---qvAvVqfsnDvrVevgLap

[gi|seaturt](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGFDFSFLPQPPQEKAHtdsRYYRADDANVmRDRDLEVDTTLKS

[gi|Alligat](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGaPSGGFDFSFmPQPPQEKAHDpGRYYRADDANVmRDRDLEVDTTLKS

[gi|Cricetu](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGyDFSFLPQPPQEKsHDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|shrew|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGFDFSFLPQPPQEKAqDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|Brandtb](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGFDFSFmPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|mouse|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGyDFSFLPQPPQEKsqDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|rat|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGyDFSFLPQPPQEKsqDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|Donkey|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSaGFDFSFLPQPPQEKsHDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|HumanCO](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSaGFDFSFLPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|Cow|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGyDlSFLPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS

[gi|Doggie|](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238f01b67a7b2804b182&results_in_list=13&raw=1&file=alignment.html#.) GPPGPPGPPGPPGPPSGGFDFSFLPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS

The next step in the Phylogeny.fr pipeline was curation. The purpose of curation is to clean up inconsistencies in the multiple sequence alignment. The alignment can be made more or less stringent based on parameter selection. Given the presence of large gaps between the proteins, parameters decreasing the stringency were selected. Initially, only gap relaxation from the Gblocks curation algorithm was selected, but the best analysis was produced from the relaxation of all criteria (gaps, flanks and size). The curation step produced a multiple sequence alignment with fewer amino acids . The 453 selected residues not filtered by curation were underlined with blue.

Curation Output:

1510 1520 1530 1540 1550 1560

=========+=========+=========+=========+=========+=========+

[**gi|algea|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) RVDSYQAHEARQVAD----QLADEQRHVSL-------------------FAYGVGRGVDR

[**gi|seaturtle|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQATTAQKNWYISKNPKEKKHVWFGETMSDGFQ----------FEYG-GEGSNP

[**gi|Alligator|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVHPTQATIAQKNWYMSKNPKEKKHIWFGETMSDGFQ----------FEYG-GEGSNP

[**gi|Cricetulus]**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) QTCVFPTQPTVPQKNWYISPNPKEKEHVWFGESMTDGFQ----------FEYG-SEGSDP

[**gi|shrew|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPSVAKKNWYVSKN-KDKRHVWFGESMTHGFQVLTRSSLFSSFPCS-SSDSDP

[**gi|Brandtbat|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPTVAQKNWYISKNPKEKKHVWFGESMTGGFQ----------FEYG-GQDSDP

[**gi|mouse|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) QTCVFPTQPSVPQKNWYISPNPKEKKHVWFGESMTDGFP----------FEYG-SEGSDP

[**gi|rat|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) QTCVFPTQPSVPQKNWYISPNPKEKKHVWFGESMTDGFQ----------FEYG-SEGSDP

[**gi|Donkey|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPQVAQKNWYISKNPKDKRHVWYGESMTDGFQ----------FEYG-GQGSDP

[**gi|HumanCOL1A1|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPSVAQKNWYISKNPKDKRHVWFGESMTDGFQ----------FEYG-GQGSDP

[**gi|Cow|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPSVAQKNWYISKNPKEKRHVWYGESMTGGFQ----------FEYG-GQGSDP

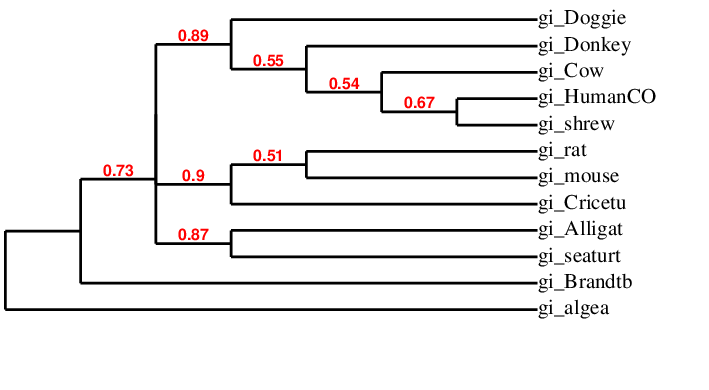
[**gi|Doggie|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPQVAQKNWYISKNPKEKRHVWYGESMTDGFQ----------FEYG-GQGSDP

####################################### ######

Conserved positions: **26%**

The next step in the Phylogeny.fr pipeline was probabilistic method selection. The pipeline offers several choices with regards to the mechanism of tree building (neighbor joining, distance, maximum likelihood, maximum parsimony, and Bayesian). Bayesian was selected because it is the most rigorous statistical tree building option. To produce a possible phylogenic relationship with probability quantifications, the WAG protein substitution model with no pre-supposed variation rate, and a 6 substitution type model were used to simulate 10,000 generations with a tree sampling every 10th generation, and the arbitrary dismissal of the first 250 trees produced. These parameters produced the highest confidence tree for hypothetical phylogenic tree branching between the twelve proteins. This step produced the probabilistic quantification of the many hypothetical phylogenic relationships that were subsequently sampled to represent a population (of hypothetical trees). The samples were then be graphically represented by a tree.

Bayesian Method Tree Rendering Output:



After several analyses, the previously outlined parameters produced a tree with the highest confidence values. Bayesian analysis yielded probability values associated with the hypothetical branching similar to bootstrap values. This analysis properly identified the collagen-like, algae produced, protein outgroup and produced an overall clade branching pattern with a confidence level of 73%. This suggests an adequate outgroup selection and a homologous relationship between the other proteins. The more recent branching patterns and clade separations in terms of evolutionary time have below threshold confidence values and may be an inaccurate representation of recent evolutionary relationships. The overall grouping of the shrew, cow, donkey, dog, and human COL1A1 protein into one clade is supported by the 89% hypothesis probability value. However, the subgrouping amongst the members of the clade may be inaccurate as indicated by the hypothesis probability values of 54%, 55%, and even 67%. The division of the sea turtle and Mississippi alligator COL1A1 proteins into one clade and the rat, mouse, and cricetulus (rodent) COL1A1 proteins into another clade is supported by the respective hypothesis probability values of 87% and 90%. The isolation of the Brandts’ bat COL1A1 protein although supported by the hypothesis probability values seems unlikely due to the bats known close phylogenic relationship to other rodents present in this analysis. An additional benefit of the Bayesian method of tree building in the Phylogeny.fr pipeline is the Newick format which provides a quantification of evolutionary distance based on substitutions. From the Newick output and different types of tree build options, it can be inferred that the algae protein (and the organism itself) is the most ancestral. Of the homologous vertebrate produced COL1A1 proteins, the Brandt’s bat is the most ancestral followed by the sea turtle and alligator based on the pipeline output. From the tree, it can be inferred that humans and shrews are the most recently diverged species followed by the cow but the probability hypothesis values are below the threshold value. The dog and cricetulus (rodent) can be interpreted as similar in evolutionary age to the sea turtle and the alligator but older than the donkey, rat, and mouse who are older than the human, shrew and cow. Although there is some doubt with regards to the most recent (in evolutionary time) branching due to low probability values, the Phylogeny.fr pipeline supported the BLASTP identification of ten COL1A1 protein orthologues. The low confidence values for recent (in evolutionary time) branching are a result of the small sample size used for this analysis. With many more possible homologs, the Phylogeny.fr pipeline could produce more accurate branching with higher confidence values.