**I) Introduction to Structure and Function collagen alpha 1 type 1 protein chain**

a) COL1A1 and collagen type I alpha 1 are the official symbol and official gene name corresponding to RefSeq accession # NG\_007400.1. RefSeq also includes EDSC, OI1, OI2, OI3, and OI4 as aliases of the COL1A1 gene. The COL1A1 gene codes for collagen alpha-1(I) chain protein, also known as alpha-1 type I collagen, alpha1(I) procollagen, collagen alpha 1 chain type I, collagen alpha-1(I) chain preproprotein, collagen of skin, tendon and bone, alpha-1 chain pro-alpha-1 collagen type 1, type I proalpha 1, and type I procollagen alpha 1 chain. The protein coded by COL1A2, alpha 2 type I procollagen chain combines in a 1:2 ratio with the protein coded by the COL1A1 gene to form type I collagen in a triple helix structure (1). Small alpha 1 chains act on either end of the three chain arrangement to twist the chains into the helical structure (1). Type I collagen is a fibril protein and a structural component of bones, ligaments, tendons and the dermal layer of epithelial tissue (1).

**II) In depth structure anlaysis**

a) COL1A1 is a gene contained in the Refseq’s human genome (hg38) assembly of chromosome 17 connected to accession number NC\_000017.11. The COL1A1 gene sequence is contained in the RefSeq gene database connected to accession number NG\_007400.1. There are curated mRNA and protein sequences for theCOL1A1 genes connected to accession numbers [NM\_000088.3](http://www.ncbi.nlm.nih.gov/nuccore/NM_000088.3) and [NP\_000079.2](http://www.ncbi.nlm.nih.gov/protein/NP_000079.2) respectively. The COL1A1 gene codes for a 1464 amino acids collagen alpha-1(I) chains.

b) Domains

i) VWC pfam00093

a) Von Willebrand factor type C domain

ii) several collagen triple helix repeats pfam01391

iii) COLFI smart 00038

a) Fibrillar collagen C-terminal domain Found at C-termini of fibrillar collagens: Ephydatia muelleri procollagen EMF1 alpha, vertebrate collagens alpha(1)III, alpha(1)II, alpha(2)V etc.

**III) Bioinformatics Analysis**

A) BLAST---Section needs editing and the addition of a more thorough paralog search

i) The COL1A1 gene codes for collagen alpha (1) chain protein. The structural protein is associated with RefSeq accession # NP\_000079.2. Using this accession number a BLASTP search was performed with a words size of 6, an expect threshold of 10, the BLOSUM62 substitution matrix, gap extension cost of 1, gap existence cost of 11. Seven in-species protein hits resulted from this query [(AFD28984.1), (AAB94054.3), (AAH36531.1), (BAD92834.1), (CAA67261.1), (P02452.5), and (CAA98968.1)]. These hits are better described as variants with single point mutations than isoforms or homologues. Of the seven in-species hits most had 100% identity, several had one residue substitutions, and one had two residue substitutions. The seven in-species hits couldn’t possibly be pseudogenes because they are derived from a sequenced protein. Pseudogenes do no code for a protein. The BLAST hits from this query were highly conserved, the lowest percent identity hit was 85% with an E-value of 0.

A second BLAST was performed with reduced sensitivity by increasing the expect threshold to 12 and the maximum hits to 500. This query produced 24 in-species hits with several possible homologues. (EAW94630.1) is a COL1A1 isoform with perfect contiguous alignment to half of the amino acid sequence of NP\_000079.2. However, this hit is more likely the result of alternative splicing than a duplication event. Many in-species hits for collagen alpha-1(II) chain proteins are possible candidates of duplication event but this protein is ubiquitous among vertebrates and maximum parsimony suggests they are not paralogs. It is unlikely that many species experienced a similar duplication event.

Based on percent identity and e-value, the BLAST produced several possible orthologues. A possible vertebrate ortholog due to an e-value of 0.0 and 97% identity, is the bovine COL1A1 gene that codes for a protein connected to accession number P02453. Another possible vertebrate orthologue is the canine COL1A1 gene connected to accession number [NM\_001003090.1](http://www.ncbi.nlm.nih.gov/nuccore/50978773) with an e-value of 0.0.

There seem to be well over 100 possible orthologues with e-values of 0 and greater than 70% identity. The two BLASTs did not produce any possible plant, insect, or prokaryote homolog hits.

The HomoloGene Accession number associated with the COL1A1 protein is HomoloGene:73874

The HomoloGene hits for the COL1A1 protein were significantly fewer in quantity than the BLAST hits. This is due to differences in the tools’ algorithms. Homologene is a more stringent search due to its global comparison rather than local alignment. HomoloGene compares sequences at the domain level, rather than the sequence level. This accounts for the fewer returned hits from the HomoloGene query.

B) Amalgamation of the multiple runs of different Multiple Sequence Alignment

i) To produce several global alignments for comparison, T-COFFEE and COBALT were used. Although not included as an option in the Edgar paper, I chose to use COLBALT because of personal familiarity. T-COFFEE was chosen because it is one of the more accurate multiple sequence aligners for under 100 sequences for proteins of under 10,000 residue length. For both tools, regions of low percent identity caused differential alignment. The globally aligning multiple sequence tools produced a different output from BLAST due to model fit. BLAST fits every sequence to one model, the query sequence, while global aligners fit multiple sequences to each-other. Based on the very similar output from the two global aligners, T-COFFEE may be the better option due to its user friendly interface with color coding sequence differentiation. T-COFFEE also produces helpful alignment scores, an option not produced by COBALT. However, it may be useful to use a hybridized approach. COBALT produces useful output regarding to location of a particular sequence’s alignment while T-COFFEE does not. Both T-COFFEE and COBALT found near perfect alignment between the 10 sequences for 250 residues. Within this 250 residue alignment, all but one sequence exhibited an identical deletion of 10 residues. Throughout the rest of the alignment of the sequences there was a high degree of variation. Based on the results of the multiple sequence alignments, the highly conserved areas between the sequences are most likely influenced by selective pressure and any mutation is deleterious. The areas of high variation are most likely under far less selective pressure, indicated by their high degree of variability. Overall, it is very likely that these 10 sequences are orthologues of the human COL1A1 protein.

ii) 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

C) Phylogenic Analysis from Phylogeny.fr

i) 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

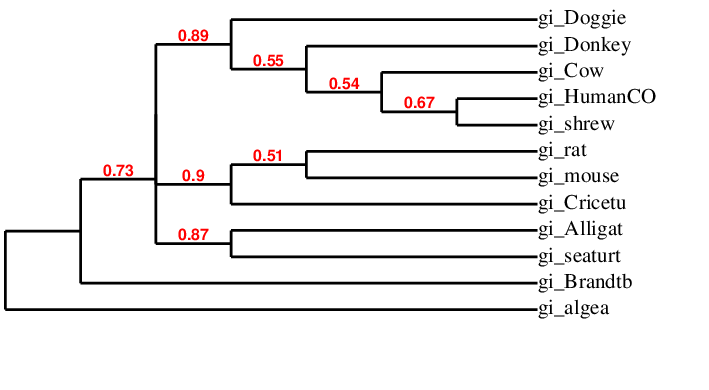
[**gi|Doggie|**](http://www.phylogeny.fr/get_result.cgi?task_id=139846fea965238fc918aa1a2403cf58&results_in_list=13&raw=1&file=align_cured.html#.) ETCVYPTQPQVAQKNWYISKNPKEKRHVWYGESMTDGFQ----------FEYG-GQGSDP

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

Conserved positions: **26%**

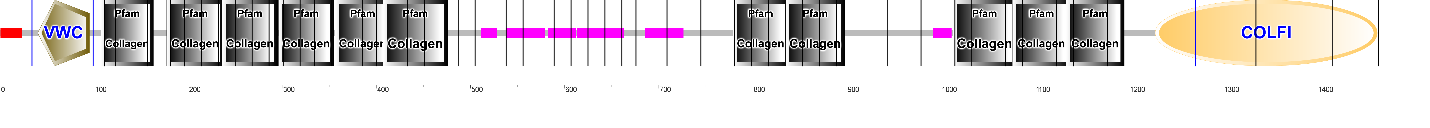
The next step in the Phylogeny.fr pipeline was probabilistic method selection. The pipeline offers several choices with regards to the mechanism of tree building (neighbor joining, distance, maximum likelihood, maximum parsimony, and Bayesian). Bayesian was selected because it is the most rigorous statistical 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.

D) Conserved Domain Analysis



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| http://www.ncbi.nlm.nih.gov/genome/guide/corehtml/transparent.gif | |  |  | | --- | --- | | Smart domain sequence of COL1A1 P02452  Conserved Domains  *Conserved Domains from CDD found in protein sequences by rpsblast searching.* | | |  | | | [COLFI (pfam01410)](http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=pfam01410) http://www.ncbi.nlm.nih.gov/HomoloGene/IMG/domains/color15b.gif  Fibrillar collagen C-terminal domain. |  | | [VWC (pfam00093)](http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=pfam00093) http://www.ncbi.nlm.nih.gov/HomoloGene/IMG/domains/color1b.gif  von Willebrand factor type C domain. |  | | [COLFI (cl02436)](http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cl02436) http://www.ncbi.nlm.nih.gov/HomoloGene/IMG/domains/color10b.gif  Fibrillar collagen C-terminal domain. |  | | [VWC (cl17735)](http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cl17735) http://www.ncbi.nlm.nih.gov/HomoloGene/IMG/domains/color19b.gif  von Willebrand factor type C domain. |  | | [Collagen (cl19732)](http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cl19732) http://www.ncbi.nlm.nih.gov/HomoloGene/IMG/domains/color5b.gif  Collagen triple helix repeat (20 copies). |  | |

i) The COL1A1 seems to be rather simple with regards to domain diversity. As a structural protein there are unique domains at the C and N terminus and then a considerable number of repeats in the interior of the protein.

a) b) Domains

i) VWC pfam00093

a) Von Willebrand factor type C domain

ii) several collagen triple helix repeats pfam01391

!!!!!DIRECT FROM CONSERVED DOMAIN DATABASE AND NEEDS EDITING Members of this family belong to the collagen superfamily. Collagens are generally extracellular structural proteins involved in formation of connective tissue structure. The alignment contains 20 copies of the G-X-Y repeat that forms a triple helix. The first position of the repeat is glycine, the second and third positions can be any residue but are frequently proline and hydroxyproline. Collagens are post translationally modified by proline hydroxylase to form the hydroxyproline residues. Defective hydroxylation is the cause of scurvy. Some members of the collagen superfamily are not involved in connective tissue structure but share the same triple helical structure.!!!!!!!!!!!!!

iii) COLFI smart 00038

a) Fibrillar collagen C-terminal domain Found at C-termini of fibrillar collagens: Ephydatia muelleri procollagen EMF1 alpha, vertebrate collagens alpha(1)III, alpha(1)II, alpha(2)V etc.

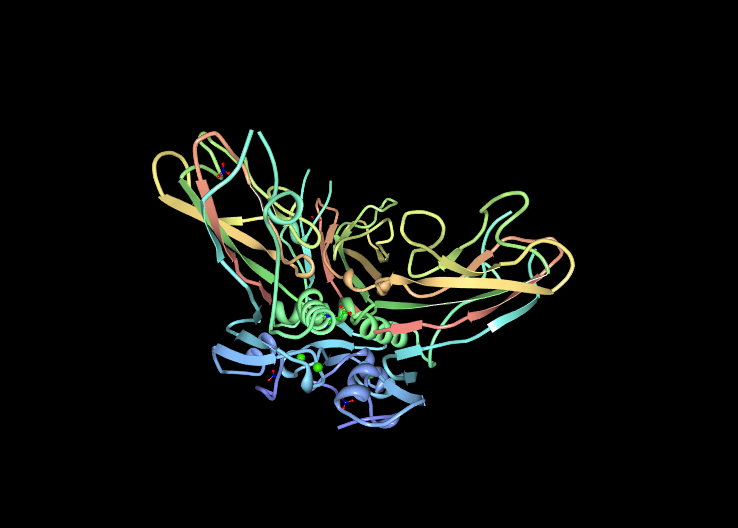
E) Homoglogy Modeling

i) The C-propeptide domain of the collagen type I alpha I chain protein was analyzed to find homologs. The domain is also called the Fibrillar Collagen C-terminal domain or simply COLFI. This domain is coded from the amino acid range 1228-1464 of the sequence associated with RefSeq accession number NP\_000079.2. This domain was chosen because it seems to be the most conserved domain of all of the collagen type I alpha I chain protein domains. Additionally, within the family of collagen proteins sharing the domain there is a correlation between mutation location in the domain and severity of the disease that occurs as a result (1). The COLFI domain has also been identified as the association point for the winding of the collagen chains into the triple helical structure (1). The different human fibrillary procollagens share 46% identity. A portion of this dissimilar identity is attributable to the chain recognition sequence, a 15 amino acid region associated with selectivity during assemblage (1). This variation explains why only homologous domains and not the exact sequence were found in the Protein Data Base. For the COLFI domain sequence, the first possible homomolog entry is the Human fibrillary procollagen type III C-propeptide trimer, associated with accession number 4AE2, elucidated via X-ray diffraction at a resolution of 1.68 angstroms, and an e-value of 1.75729E-91. The second possible homomolog entry is the Human fibrillary procollagen type III C-propeptide trimer, associated with accession number 4AEJ, elucidated via X-ray diffraction at a resolution of 2.21 angstroms, and an e-value of 1.75729E-91. The third possible homomolog entry is the Human fibrillary procollagen type III C-propeptide trimer, associated with accession number 4AK3, elucidated via X-ray diffraction at a resolution of 3.5 angstroms, and an e-Value of 1.75729E-91. The first two entries all represent the A, B, and C chains of the domain while the third entry represents only the A chain. This explains the similar length (of amino acids) of the first two entries and the difference in length of the third entry. No other possible homologs of the COLFI domain currently exist in the PDB. Images of the three dimensional model produced by the PDB are included below.

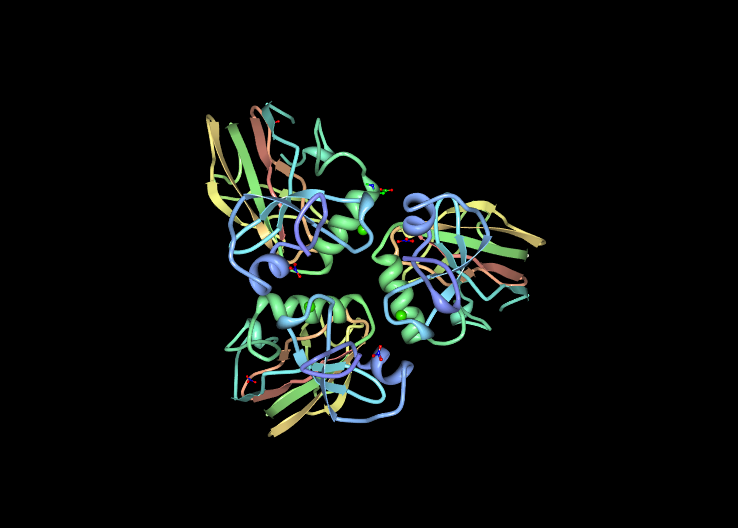
Domain View One:



Domain View Two:

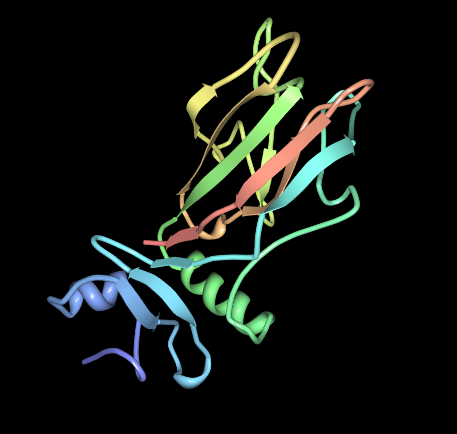


Domain View Three:



The RefSeq curated sequence for the COLFI domain was then loaded into the Phyre2 homology modeling program to find similar domain templates from homologous proteins and produce an accurate 3D structure prediction of the COLFI domain. Phyre2 produced a 3D structure from 213 residues representing 90 % of the COLFI domain sequence. The Phrye2 produced 3D visualization is included below.

Phyre2 Image 1:

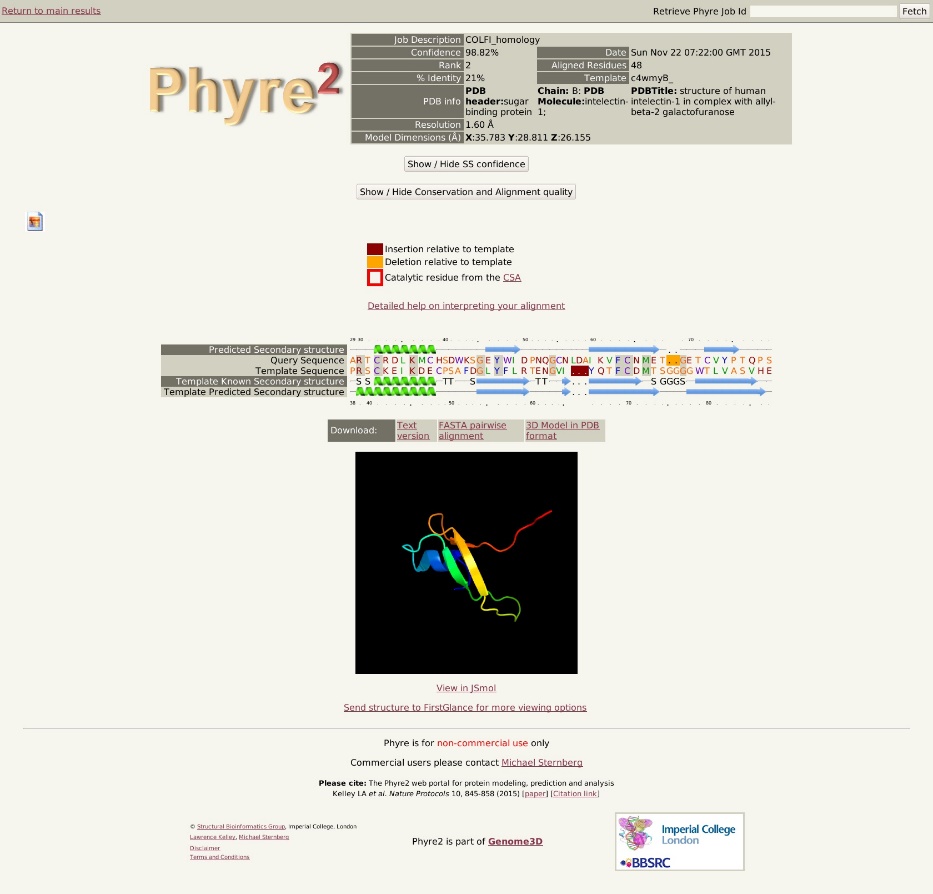


Phyre2 Image 2:



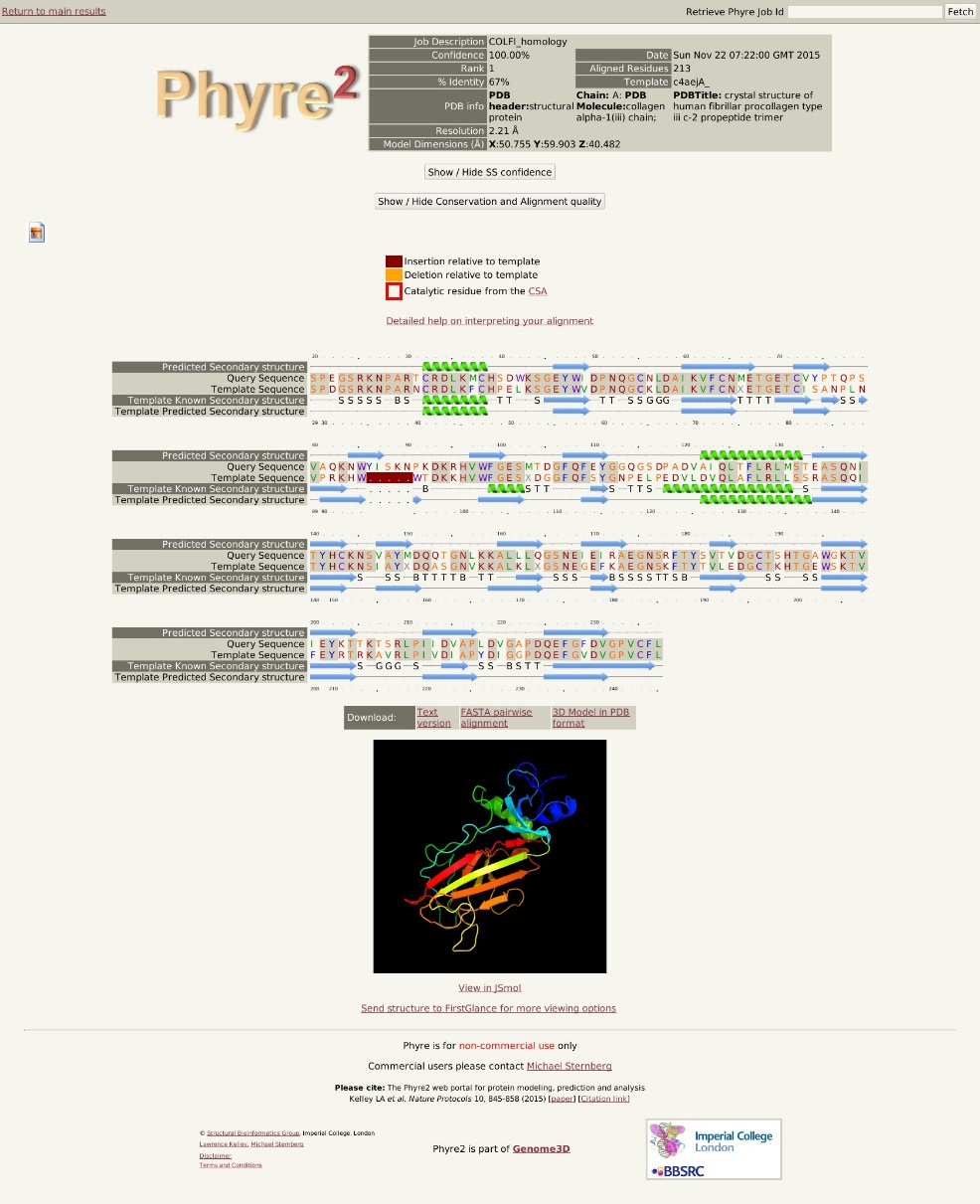
The highest ranked protein model used as a template for the 3D structure produced for Phyre2 for the COLFI domain was from the collagen alpha-1(III) chain, chain A, crystal structure of human fibrillar procollagen type III c-2 propeptide trimer. When compared to the COLFI sequence this first model produced values of 67% identity and 100% confidence. The second template model was from a sugar binding protein, human entelectin-1 complex with glactofurnose. This model produced a 21% identity value and 98.8% confidence value when compared to the COLFI domain sequence. The third template model was from a Fibrinogen C-terminal domain like protein. This template produced a 27% identity value and a 98.5% confidence value when compared to the COLFI domain sequence. An image of the alignment of the second template is included below.

Phyre2 2nd Top Model Alignment:



Overall, the second top model used as a template for COLFI seems to be a poor choice. The local alignment between the sequences represents roughly a quarter of the entire COLFI domain sequence. This region was already aligned to the top model yielding a much higher percent identity. An image of the top model alignment seen below, supports the previously expressed assertion that the second top model seems to contribute little to the overall predicted 3D structure of the COLFI domain.

Phyre2 Top Model Alignment:



It seems that the several points of identity between the second top model template similar to several points of identity between top model and the COLFI domain sequence are used to support the 3D structure and increase the confidence value.

**V) Mutational Analysis**

FH.sapiens MIM:114000

Caffey disease. MIM:120150

Dissection of cervical arteries. MIM:120150

OI/EDS combined syndrome. MIM:130000

Ehlers-Danlos syndrome, type I. MIM:130060

Ehlers-Danlos syndrome, type VII. MIM:166200

Osteogenesis imperfecta, type I. MIM:166210

OI type II. MIM:166220

OI type IV. MIM:166710

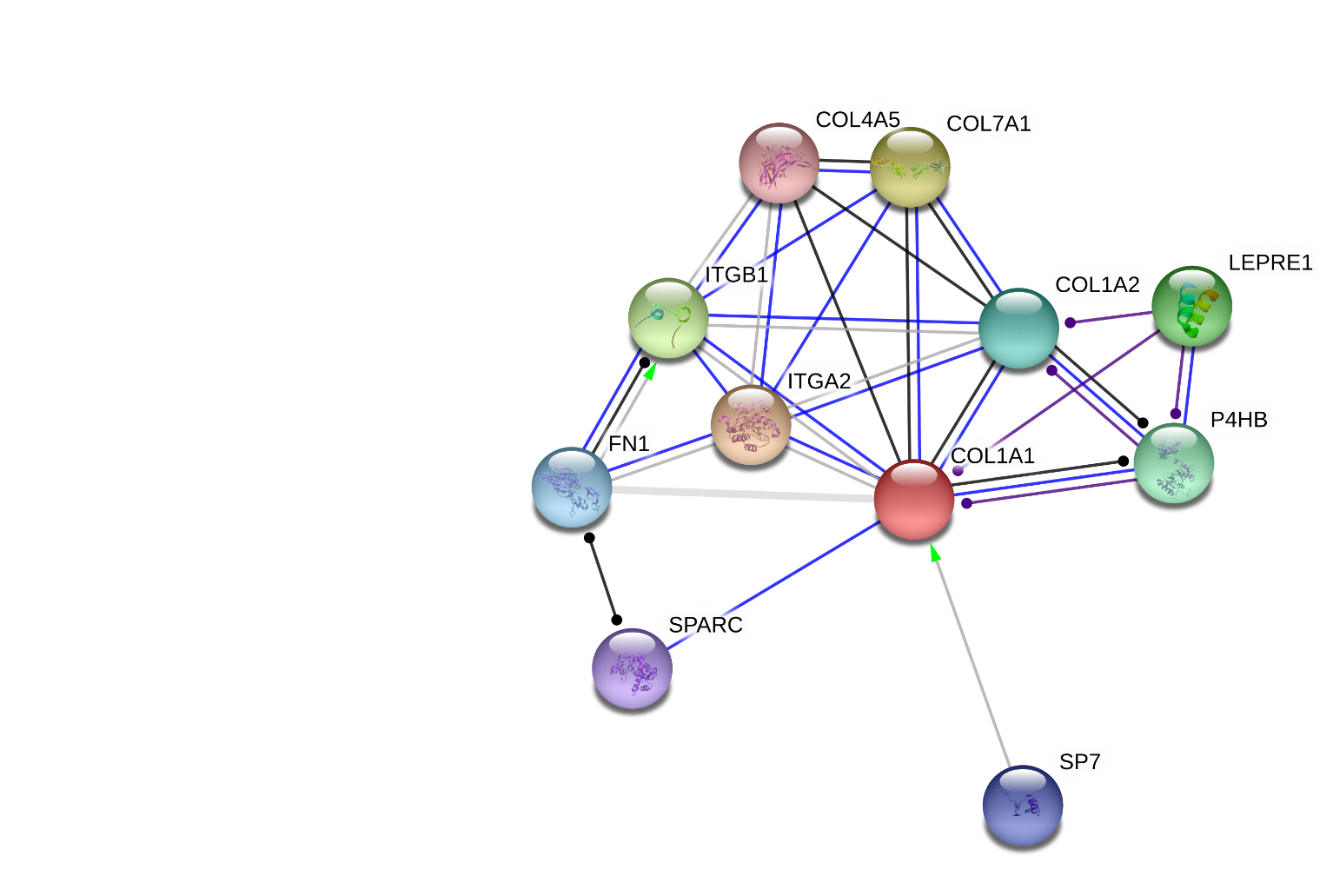
Osteoporosis. MIM:259420 OI type III.

rom HomoloGene accession numbers for COL1A1 isoforms associated with certain disease.

A) Within the COLFI domain there has been discovered a correlation between loci of a point mutation for the various homologs and severity of the disease caused by the mutation.

Mutations in the gene can result in disordered bone formation and bone degeneration manifested in many diseases (osteogenesis imperfecta, types I-IV, Ehlers-Danlos syndrome type VIIA, Ehlers-Danlos, syndrome Classical type, and Caffey Disease) (1). Meiotic events between the COL1A1 gene and the platelet-derived growth factor beta gene on chromosome 22 have been linked to a form of skin cancer (dermatofibrosarcoma protuberans) (1). To determine the extent of the known participation in disease, a query string consisting of “COL1A1” was submitted to string –db for KEGG pathways. The parameters were then modified to exclude all but the interactions yielding the highest confidence value (>.9). This yielded protein to protein interactions between COL1A1 and 9 other proteins yielding a multitude of biological processes and cellular products. Listed below are the protein interactions identified by string-db and the corresponding confidence value. These interactions were elucidate by string –db almost exclusively via text-mining, database parsing, and experimental conclusion. The interactions between COL1A1 and both COL4A5 and COL1A2 were additionally supported via homology or co-expression.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | | | http://string-db.org/images/activation_upr_line.png | http://string-db.org/images/inhibition_upr_line.png | http://string-db.org/images/binding_upr_line.png | http://string-db.org/images/phenotype_upr_line.png | http://string-db.org/images/catalysis_upr_line.png | http://string-db.org/images/ptmod_upr_line.png | http://string-db.org/images/reaction_upr_line.png | http://string-db.org/images/expression_upr_line.png | http://string-db.org/images/score_upr.png | | **Predicted Functional Partners:** | | | | http://string-db.org/images/symbols/item_symbol.7251.P.ffaf65.white.png | [ITGA2](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1847760&targetmode=proteins) | integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor); Integrin alpha-2/beta-1 is a rece [...] (1181 aa) | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=activation&taskId=prMbSJz0PErJ&node2=1847760) |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=binding&taskId=prMbSJz0PErJ&node2=1847760) |  |  |  |  |  | 0.985 | | http://string-db.org/images/symbols/item_symbol.7251.P.b2ab00.white.png | [COL7A1](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1851466&targetmode=proteins) | collagen, type VII, alpha 1; Stratified squamous epithelial basement membrane protein that form [...] (2944 aa) |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=binding&taskId=prMbSJz0PErJ&node2=1851466) |  |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=reaction&taskId=prMbSJz0PErJ&node2=1851466) |  | 0.977 | | http://string-db.org/images/symbols/item_symbol.7251.P.bbff65.white.png | [ITGB1](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1848520&targetmode=proteins) | integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) (798 aa) | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=activation&taskId=prMbSJz0PErJ&node2=1848520) |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=binding&taskId=prMbSJz0PErJ&node2=1848520) |  |  |  |  |  | 0.973 | | http://string-db.org/images/symbols/item_symbol.7251.P.0eb200.white.png | [LEPRE1](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1847720&targetmode=proteins) | leucine proline-enriched proteoglycan (leprecan) 1 (736 aa) |  |  |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=catalysis&taskId=prMbSJz0PErJ&node2=1847720) |  |  |  | 0.966 | | http://string-db.org/images/symbols/item_symbol.7251.P.65ffa3.white.png | [P4HB](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1850962&targetmode=proteins) | prolyl 4-hydroxylase, beta polypeptide; This multifunctional protein catalyzes the formation, b [...] (508 aa) |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=binding&taskId=prMbSJz0PErJ&node2=1850962) |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=catalysis&taskId=prMbSJz0PErJ&node2=1850962) |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=reaction&taskId=prMbSJz0PErJ&node2=1850962) |  | 0.964 | | http://string-db.org/images/symbols/item_symbol.7251.P.00b29d.white.png | [COL1A2](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1847845&targetmode=proteins) | collagen, type I, alpha 2 (1366 aa) |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=binding&taskId=prMbSJz0PErJ&node2=1847845) |  |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=reaction&taskId=prMbSJz0PErJ&node2=1847845) |  | 0.963 | | http://string-db.org/images/symbols/item_symbol.7251.P.65c7ff.white.png | [FN1](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1853098&targetmode=proteins) | fibronectin 1 (2477 aa) |  |  |  |  |  |  |  |  | 0.957 | | http://string-db.org/images/symbols/item_symbol.7251.P.001cb2.white.png | [SP7](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1848454&targetmode=proteins) | Sp7 transcription factor; Transcriptional activator essential for osteoblast differentiation. B [...] (431 aa) | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=activation&taskId=prMbSJz0PErJ&node2=1848454) |  |  |  |  |  |  |  | 0.956 | | http://string-db.org/images/symbols/item_symbol.7251.P.9665ff.white.png | [SPARC](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1843274&targetmode=proteins) | secreted protein, acidic, cysteine-rich (osteonectin); Appears to regulate cell growth through [...] (303 aa) |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=binding&taskId=prMbSJz0PErJ&node2=1843274) |  |  |  |  |  | 0.951 | | http://string-db.org/images/symbols/item_symbol.7251.P.ef7575.white.png | [COL4A5](http://string-db.org/newstring_cgi/display_single_node.pl?taskId=prMbSJz0PErJ&node=1851408&targetmode=proteins) | collagen, type IV, alpha 5; Type IV collagen is the major structural component of glomerular ba [...] (1691 aa) |  |  |  |  |  |  | [**•**](http://string-db.org/newstring_cgi/show_set_evidence.pl?data_channel=reaction&taskId=prMbSJz0PErJ&node2=1851408) |  | 0.949 | | | |
| |  | | --- | |  |   The three dimensional network elucidated by string-db is involved with a multitude of proliferative KEGG pathways. . |



An abridged list of identified KEGG pathways excluding all pathways with a p-value of >.05.

04512 ECM-receptor interaction 1.95E-12

04510 Focal adhesion 3.82E-10

04151 PI3K-Akt signaling pathway 8.1E-9

04974 Protein digestion and absorption 8.54E-8

05146 Amoebiasis 2.02E-7

04611 Platelet activation 4.38E-7

The KO identifiers of each of these pathways were then used to search the KEGG database for visual representations of the biological pathways, included at the conclusion of this paper.

B) Mutations of the COL1A1 gene are linked to numerous degenerative bone disease such as osteogenesis imperfecta, Ehlers-Danlos syndrome, infantile cortical hyperostosis, and osteoporosis. 90% of Osteogenesis imperfecta type I (the least severe), II, III, and IV (the most severe) are coupled with COL1A1 or COL1A2 mutations (1). The disease can be associated with a single point mutation or many, with the most severe cases of the disease arising from mutations in the most highly conserved region of homologous proteins (3). The mechanism of degeneration is impaired inter-chain disulfide bonds and subsequent abnormal chain integration (3)

C) COL1A1 may also have suppressor function for particular cancers. In a study regarding Hepatocellular carcinoma (HCC), COL1A1 was found to be significantly downregulated at tumor sites (5). No chromosomal mutation could be found and the study identified promoter methylation as the mechanism of expression interference (5) Upregulation of COL1A1 may also be associated with particular forms of cancer. Reciprocal translocation of the COL1A1 gene on chromosome 17 with the platelet-derived growth factor beta gene on chromosome 22 is associated with the skin tumor dermatofibrosarcoma protuberans (4). This is attributable to the growth factor beta gene’s subsequent unregulated expression (4).

**VII) Conclusion**

A) Restate biological significance

B) Based on the existing information and disease association it seems a worthy cause to further analyze the COL1A1 protein beyond this simple cursory examination in an effort to curb certain diesaese.

**WORKS CITED needs to be organized and include programs used**

1) [Xiran Wang](http://www.ncbi.nlm.nih.gov/pubmed/?term=Wang%20X%5Bauth%5D),[Yu Pei](http://www.ncbi.nlm.nih.gov/pubmed/?term=Pei%20Y%5Bauth%5D),[Jingtao Dou](http://www.ncbi.nlm.nih.gov/pubmed/?term=Dou%20J%5Bauth%5D),[Juming Lu](http://www.ncbi.nlm.nih.gov/pubmed/?term=Lu%20J%5Bauth%5D),[Jian Li](http://www.ncbi.nlm.nih.gov/pubmed/?term=Li%20J%5Bauth%5D),and [Zhaohui Lv](http://www.ncbi.nlm.nih.gov/pubmed/?term=Lv%20Z%5Bauth%5D)Genet Mol Biol. 2015 Mar; 38(1): 1–7 “Identification of a novel *COL1A1* frameshift mutation, c.700delG, in a Chinese osteogenesis imperfecta family”

2) Chessler SD and Byers PH, J. Biol. Chem. 267 (11), pgs7751-7757 (1992) “Defective folding and stable association with protein disulfide isomerase/prolyl hydroxylase of type I procollagen with a deletion in the pro alpha 2(I) chain that preserves the Gly-X-Y repeat Pattern,”

3) [Chessler SD](http://www.ncbi.nlm.nih.gov/pubmed/?term=Chessler%20SD%5BAuthor%5D&cauthor=true&cauthor_uid=8349697)1, [Wallis GA](http://www.ncbi.nlm.nih.gov/pubmed/?term=Wallis%20GA%5BAuthor%5D&cauthor=true&cauthor_uid=8349697), [Byers PH](http://www.ncbi.nlm.nih.gov/pubmed/?term=Byers%20PH%5BAuthor%5D&cauthor=true&cauthor_uid=8349697). [J Biol Chem.](http://www.ncbi.nlm.nih.gov/pubmed/8349697) 1993 Aug 25; 268(24):18218-25. “Mutations in the carboxyl-terminal propeptide of the pro alpha 1(I) chain of type I collagen result in defective chain association and produce lethal osteogenesis imperfecta.”

4) [Nakamura I](http://www.ncbi.nlm.nih.gov/pubmed/?term=Nakamura%20I%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Kariya Y](http://www.ncbi.nlm.nih.gov/pubmed/?term=Kariya%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Okada E](http://www.ncbi.nlm.nih.gov/pubmed/?term=Okada%20E%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Yasuda M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Yasuda%20M%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Matori S](http://www.ncbi.nlm.nih.gov/pubmed/?term=Matori%20S%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Ishikawa O](http://www.ncbi.nlm.nih.gov/pubmed/?term=Ishikawa%20O%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Uezato H](http://www.ncbi.nlm.nih.gov/pubmed/?term=Uezato%20H%5BAuthor%5D&cauthor=true&cauthor_uid=26332510), [Takahashi K](http://www.ncbi.nlm.nih.gov/pubmed/?term=Takahashi%20K%5BAuthor%5D&cauthor=true&cauthor_uid=26332510).[JAMA Dermatol.](http://www.ncbi.nlm.nih.gov/pubmed/26332510) 2015 Sep 2 “A Novel Chromosomal Translocation Associated With COL1A2- PDGFB Gene Fusion in Dermatofibrosarcoma Protuberans: PDGF Expression as a New Diagnostic Tool.”

5) [Hayashi M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Hayashi%20M%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Nomoto S](http://www.ncbi.nlm.nih.gov/pubmed/?term=Nomoto%20S%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Hishida M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Hishida%20M%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Inokawa Y](http://www.ncbi.nlm.nih.gov/pubmed/?term=Inokawa%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Kanda M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Kanda%20M%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Okamura Y](http://www.ncbi.nlm.nih.gov/pubmed/?term=Okamura%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Nishikawa Y](http://www.ncbi.nlm.nih.gov/pubmed/?term=Nishikawa%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Tanaka C](http://www.ncbi.nlm.nih.gov/pubmed/?term=Tanaka%20C%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Kobayashi D](http://www.ncbi.nlm.nih.gov/pubmed/?term=Kobayashi%20D%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Yamada S](http://www.ncbi.nlm.nih.gov/pubmed/?term=Yamada%20S%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Nakayama G](http://www.ncbi.nlm.nih.gov/pubmed/?term=Nakayama%20G%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Fujii T](http://www.ncbi.nlm.nih.gov/pubmed/?term=Fujii%20T%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Sugimoto H](http://www.ncbi.nlm.nih.gov/pubmed/?term=Sugimoto%20H%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Koike M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Koike%20M%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Fujiwara M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Fujiwara%20M%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Takeda S](http://www.ncbi.nlm.nih.gov/pubmed/?term=Takeda%20S%5BAuthor%5D&cauthor=true&cauthor_uid=24552139), [Kodera Y](http://www.ncbi.nlm.nih.gov/pubmed/?term=Kodera%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=24552139).[BMC Cancer.](http://www.ncbi.nlm.nih.gov/pubmed/24552139) 2014 Feb 19 “Identification of the collagen type 1 α 1 gene (COL1A1) as a candidate survival-related factor associated with hepatocellular carcinoma.”

Works Cited

1) Chessler SD and Byers PH, J. Biol. Chem. 267 (11), pgs7751-7757 (1992) “Defective folding and stable association with protein disulfide isomerase/prolyl hydroxylase of type I procollagen with a deletion in the pro alpha 2(I) chain that preserves the Gly-X-Y repeat

Pattern,”

2) [Young SM](http://www.ncbi.nlm.nih.gov/pubmed/?term=Young%20SM%5BAuthor%5D&cauthor=true&cauthor_uid=25756146), [Bansal P](http://www.ncbi.nlm.nih.gov/pubmed/?term=Bansal%20P%5BAuthor%5D&cauthor=true&cauthor_uid=25756146), [Vella ET](http://www.ncbi.nlm.nih.gov/pubmed/?term=Vella%20ET%5BAuthor%5D&cauthor=true&cauthor_uid=25756146), [Finelli A](http://www.ncbi.nlm.nih.gov/pubmed/?term=Finelli%20A%5BAuthor%5D&cauthor=true&cauthor_uid=25756146), [Levitt C](http://www.ncbi.nlm.nih.gov/pubmed/?term=Levitt%20C%5BAuthor%5D&cauthor=true&cauthor_uid=25756146), [Loblaw A](http://www.ncbi.nlm.nih.gov/pubmed/?term=Loblaw%20A%5BAuthor%5D&cauthor=true&cauthor_uid=25756146), [Can Fam Physician.](http://www.ncbi.nlm.nih.gov/pubmed/25756146) Jan 61 (1) pgs26-35 (2015), “Systematic review of clinical features of suspected prostate cancer in primary care.”

Works Cited:

1) Nat Struct Mol Biol. 2012 Oct; 19(10): 1031–1036, Published online 2012 Sep 23. “Structural Basis of Fibrillar Collagen Trimerization and Related Genetic Disorders”

Jean-Marie Bourhis,1,2 Natacha Mariano,1 Yuguang Zhao,3 Karl Harlos,3 Jean-Yves Exposito,1 E. Yvonne Jones,3 Catherine Moali,1 Nushin Aghajari,4 and David J.S. Hulmes1