A Bioinformatics Exploration of COL1α1

Bioinformatics- BIOL-360, Professor Todd Riley

Andrew Judell-Halfpenny

Introduction to COL1α1

COL1α1 is a gene that encodes for a collagen type I alpha 1which is a structural protein. COL1A1 is a member of the collagen type I family, the most abundant form of collagen in humans (1, ***HHS***). an alpha-1 subtype of collagen type I. The cytogenic location of The COL1A1 gene is on chromosome 17q21.33 of Refseq's human genome (hg37) assembly which is connected to accession number NC_000017.11 [Appendix Image 1] (2, ***RefSeq***). The COL1α1 gene encodes a length of 24,544 nucleotide base-pair (bp) and is associated with the RefSeq accession identifier NG_007400. The COL1α1 gene is transcribed as a monocistronic mRNA transcript (a single gene mRNA transcript unlike polycistronic/multiple gene transcripts which are frequent in prokaryotes). The COL1A1 transcript is 5927 nucleotides long and is associated with the RefSeq accession identifier NM_000088 (2, ***RefSeq***). The COL1α1 mRNA transcript is the reverse complement of the genomic sequence and serves as the translation template sequence for the transition of the sequence to a 1464 residue (amino acid) functional collagen alpha 1 type 1 protein chain associated with the RefSeq accession identifier NP_000079 (2, ***RefSeq***). The RefSeq genomic sequence, the mRNA transcript, and the collage type I alpha 1 protein product are the canonical forms of the COL1A1 gene/nucleotide-sequence/amino acid-sequence and each feature represents one stage of the transcription of DNA to RNA, and the translation of RNA to protein as per the central dogma of molecular biology.

COL1α1 is the gene symbol, a pseudo-vernacular name for the gene that encodes the collagen type I-αpha1 protein chain. RefSeq database curators suggest that the preferred name for the COL1α1 protein is the "collagen alpha-

1(I) chain." Unfortunately the lack of a universal-standardized gene-naming convention has lead to the development of many different nicknames. Many databases that curate and host bioinfomtics data use unique identifiers for COL1A1 (i.e. HGNC:2197; OMIM:120150; Ensembl: ENSG00000108821). ThermoFisher Scientific's alias list for the COL1A1 gene includes "Col1a-1"; "COL1A1"; "Cola-1"; "Cola1"; "COLIA1"; "EDSC; "Mov-13"; "Mov13"; "OI1"; "OI2"; "OI3"; "OI4" (*4*, ***ThermoFisher***). ThermoFisher Scientific's alias list for collagen type I alpha 1 protein chain includes "alpha-1 type 1 collagen"; "Alpha-1 type I collagen"; "alpha1(I) procollagen"; "COL1A"; "Collagen 1"; "collagen alpha 1 chain type I"; "Collagen alpha-1(I) chain"; "collagen alpha-1(I) chain preproprotein";"collagen of skin, tendon and bone, alpha-1 chain"; "collagen, type 1, alpha 1"; "collagen, type I, alpha 1"; "pro-alpha-1 collagen type 1"; "procollagen type I, alpha 1"; "procollagen, type 1, alpha 1"; "procollagen, type I, alpha 1"; "Type I Collagen"; "type I proalpha 1"; "type I procollagen alpha 1 chain"; and "type I-alpha 1 collagen (*4*, ***ThermoFisher***)."

Introduction to Structure and Function of COL1α1

Collagen type I alpha-1 is a fibril protein and a major component of structural and connective tissue. The COL1α1 protein is a structural component of bones, ligaments, tendons and the dermal layer of epithelial tissue (15, ***Jean-Marie Bourhis***). COL1Aa1 forms a triple helix with COL1A2, a homologous collagen protein. The helical structure is produced from a one to two (1:2) ratio of the protein coded by COL1A2 (alpha 2 type I procollagen chain) and the COL1A1 protein (15, ***Jean-Marie Bourhis***). Small alpha 1 chains act on either end of the three chain arrangement to twist the chains into the helical structure

(15, ***Jean-Marie Bourhis***). Although the COL1A1 transcript is monocistronic it can be inferred from the different lengths of its canonical forms that post translational modification is included in its processing which is parallel to the central dogma of molecular biology. The finished product is a collagen type I alpha 1 protein chain unit that is 1464 residues long (equal to 4392 nucleotides given a triplet code) and ~82% shorter than the original genomic COL1A1 transcript sequence. The difference in length between the canonical genomic sequence, the mRNA transcript, and the protein product suggests that one or more intronic regions were spliced out during COL1A1 processing.

The convenience of exploring bioinformatics data through the National Center for Biotechnology Information RefSeq API is the accessibility of information related to a sequence with a RefSeq accession identifier. To explore the splicing landscape of COL1A1 there is an Ensembl hyperlink to the right of the RefSeq interface that leads to information about COL1A1. It is immediately obvious that the massive number of splicing events and features are beyond the scope of this analysis. Ensembl curates a repository similar to RefSeq's and also maintain unique identifiers for COL1A1's gene [ENSG00000108821], transcript [ENST00000225964.9], and protein [ENSP00000225964.5] (*3,* ***T. Hubbard***). Ensembl lists 51 exons and 1838 different single nucleotide polymorphisms or multi nucleotide polymorphisms (SNPs, MNPs, and variants) for the COL1A1 transcript. At the gene level, Ensembl lists 13 different transcripts (splice variants), 65 orthologues, 16 paralogues, and 66 associated phenotypes (*3,* ***T. Hubbard***). PMAP-CutDB, a data bank of proteolytic events listed 13 known proteolytic events associated with alternative splicing of COL1A1. Unfortunately,

the RefSeq and Ensembl databases do not always agree and Ensembl lists the slightly different lengths and genomic locations for COL1A1. Ensembl provides an ideogram of chromosome 17 that is included in the Appendix as image #3. The ideogram is a graphical depiction of Ensembl's list of known variants for COL1A1 in pileup format. Many other curated databases offer similar ideograms. The Protein Data Bank (PDB) is a curated database that catalogs the sequences and 3-dimensional structures of protein domains, the independently folding units of a full length protein sequence.

Protein Domains

The PDB is one of several protein-centric databases. The Conserved Domain Database (CDD), UniProt, and many more esoteric databases like the Plasma Proteome Database curate sequence and structural information about proteins. The COL1A1 entry from each of the previously mentioned databases was explored and included in the Appendix of this analysis. There is general agreement between the databases regarding the domains of the COL1A1 protein chain. The Collage alpha-(I) chain full length protein is associated with accession # P02452 (9, **Pundir**). The PDB entry is associated with a protein feature view, ideogram-like graphical representation of the COL1A1 consensus. The Protein feature view is similar to Ensembl's ideogram and is included in the appendix.

The CDD entry for the COL1A1 protein is accessible via link from the PDB and describes COL1A1 as a member of the Collagen gene family which includes 45 other different types of collagen (4, ***Marchler-Bauer***). The CDD includes PFAM and SMART accessions for the domains of the COL1A1 protein chain. As

an incredibly abundant structural protein, COL1A1 protein is fairly simple. It is comprised of three main domains with a repetitious sequence in the middle which serves to elongate the structural protein. The VWC domain is the von Willebrand factor type C domain which is associated with the PFAM accession # pfam00093 and the Conserved Domain accession # cl17735 (4, ***Marchler-Bauer***). The other C-terminus-like domain is the COLFI domain which is associated with the SMART accession number smart00038, the Conserved Domain accession # c102436 and the PFAM accession # pfam01410 (4, ***Marchler-Bauer***). The third conserved domain of the COL1A1 protein is the Collagen triple-helix repeat region which occurs roughly ~20 times throughout the length of the COL1A1 protein and is associated with the Conserved Domain accession # cl19732 (4, ***Marchler-Bauer****)*. As a structural protein, COL1A1 lacks an N-terminus to allow for protein assembly of multiple COL1A1 repeats. The interior of the COL1A1 is composed variable lengths of triple helix G-X-Y repeat regions before being terminated by a fibrillar collagen C-terminal domain associated with PFAM accession pfam01410. The G-X-Y triplet repeats are highly conserved at the first (Glycine) position but often exhibit variable residue conservation at the second and third amino-acid positions of the repeats. The second and third amino acid positions are often held by proline and hydroxyproline (*10,* ***Chessler***).

To explore the sequence conservation of the ubiquitous COL1A1 protein between species string parsing was performed with BLAST. COL1A1's RefSeq protein accession identifier (NP_000079.2) was used to perform an initial query with a word size of 6 characters, an expect threshold of 10, the BLOSUM62

substitution matrix, gap extension cost of 1, gap existence cost of 11 and the maximum hits to 500 (7, **Altschul**). This query produced 24 in-species hits with several potential COL1A1 homologs. (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 potential COL1A1 paralogs were produced by the blast query. Since there are roughly 50 members of the collagen protein family with marginal sequence variability, it is probable that many of these collagen family proteins are paralogs rather than alternative slice-o-forms. Seven in-species protein hits from this query [(AFD28984.1), (AAB94054.3), (AAH36531.1), (BAD92834.1), (CAA67261.1), (P02452.5), and (CAA98968.1)] can be presumed to be paralogs. These hits are most likely not variants with single point mutations than because of the relatively high proportion of differences when compared to COL1a1. Of the 24 in-species hits, seven had ~95% identity, many of the other hits had one or two residue substitutions. (EAW94630.1) is a COL1A1 isoform with perfect contiguous alignment to half of the amino acid sequence of NP_000079.2. It is difficult to determine whether these potential paralogs are the result of alternative splicing or a duplication event.

The BLAST query produced several possible orthologs identifiable by arelatively high percent identities and low e-values. The bovine COL1A1 protein associated with accession number P02453 was identified as a potential ortholog due to its low e-value of 0.0 and high percent identity ~ 97% . The BLAST query produced an e-value of 0.0 for the canine COL1A1 gene associated with the

accession number NM_001003090.1 which is indicative of a potential vertebrate ortholog (7, **Altschul**).

There seem to be well over 100 possible orthologues with e-values of 0 and greater than 70% identity. The BLAST query did not produce any possible plant, insect, or prokaryote homolog hits. The Homologene database is linked to the the BLAST-output API and provided a direct link to the Homologene entry for the Collagen type I alpha 1 gene associated with HomoloGene Accession number:73874 (8, **NCBI-NLM**).

The HomoloGene hits for the COL1A1 gene include D. rerio (zebra fish), X. tropicalis (tropical clawed frog), R. norvegicus (rat), M. musculus (mouse), B.taurus (cow), C. lupus (dog), and P. troglodytes (chimpanzee) (8, **NCBI-NLM**). Each of these species expressed a protein product of the COL1A1 between 1447 and 1463 nucleotides long. Homologene also contained links to several COL1A1 related diseases throigh the Online Mendelian Inheritance in Man database. The list of diseases associated with mutations in the COL1A1 gene include Caffey disease (MIM:114000), Dissection of cervical arteries (MIM:120150), OI/EDS combined syndrome (MIM:120150), Ehlers-Danlos syndrome type I (MIM:130000), Ehlers-Danlos syndrome type VII (MIM:130060), Osteogenesis imperfecta type I (MIM:166200), Osteogenesis imperfecta type II (MIM:166210), Osteogenesis imperfecta type III (MIM:259420), Osteogenesis imperfecta type IV (MIM:166220), and Osteoperosis (MIM:259420) (6, **Ada Hamosh**). The large number of diseases that arise from mutations in the COL1A1 gene may be related to the body's widespread production of COL1A1 structural protein as a component of the extra-cellular matrix. There are at least 28 KEGG and

REACTOME biochemical pathways in which COL1A1 plays a role. Some of these COL1A1 involved pathways included, IL-4 mediated signaling, focal adhesion, extra cellular receptor interactions, extra cellular matrix organization, assembly of collagen fibrils, collagen chain trimerization, protein digestion, platelet activation, P13K-Aakt signaling, osteoblast signaling, scavenging by class A receptors and the inflammatory response pathway among many others. To explore the sequence similarity and relative importance of COL1A1 in other species (9, **Pundir**).

GEOquery and StringDB: Network, Coexpression, Protein-Interactions

To more thoroughly explore the different networks and biochemical pathways which COL1A1 plays a role, the Gene Expression Omnibus Repositories were queried using “COL1A1” as the query. The output from the GEOquery search included several historical datasets which compared the gene expression of a patient group suffering from some disease or ailment to the gene expression of a control group. Many of the datasets were comparative analyses that explored the differential expression of patients suffering from a particular form of cancer. The first dataset retreived by the GEOquery service was an expression profiling array of 22 patients with advanced primary gastric cancer using an Affymetric Human Gene Expression chip (8, **Szklarczyk**). The 22 tissue samples were compared to a control group of 8 samples. In the control group, expression of COL1A1 was generally uniform and expressed at significantly lower levels than the expression of COL1A1 in the gastric cancer patient group. The patient group did not exhibit uniform expression of COL1A1. After a cursory examination of the data, the variance of COL1A1 expression within the patient

group and the the mean expression of COL1A1 within the patient group may be roughly equivalent.

The second data set explored the gene expression profiles of of patients with Papillary Thyroid Carncinoma in comparison with a control group. Tissue samples from patients with cancer were analyzed via micro array. The gene expression profiles of 7 patients with advanced Papillary Thyroid Carcinoma (PTC) exhibited similar expression profiles to the patients of the gastric cancer comparative analysis from the first GEO Profile dataset (8, **Szklarczyk**). When compared to a 7 sample control group, the patient group expressed significantly higher levels of COL1A1. In this analysis the differences in expression between the two groups was not as large as in the gastric cancer analysis. Additionally the variance of the expression of COL1A1 for the patient group was nowhere near as high as that of the gastric cancer analysis patient group.

The last GEO Profile data-set came from an experiment that was performed on an immortalized pancreatic carcinoma cell line. The experimenters were interested in certain biological processes and phenomena associated with epithelial mesenchymal transition (EMT) (8, **Szklarczyk**). The experimenters repeatedly treated the cell line with TGFB prior to analyzing gene expression with an Affymetrix chip. The results were consistent through each repetition of the procedure. The control group had significantly lower expression of COL1A1 than the group treated with TGFB which was very similar to the first two comparative analyses accessed from the Gene Expression Omnibus Repository.

Given the consistent differential expression exhibited by the patient groups in the historical datasets of the GEO repository it seems likely that the functional

partners that cause differential expression of COL1A1 as a result of cancer may be documented. To determine whether there are known cancer networks the StringDB data repository was queried for high confidence networks in Homo sapiens that include the expression of COL1A1. This query only produced two networks with high confidence scores, an extra-cellular matrix related pathway and a pathway centered around COL1A1 and NFK-beta (8, **Szklarczyk**).

Exploring Conservation with Multiple Sequence Alignments

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.

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.

Phylogenic Analysis from Phylogeny.fr

The first step of the Phylogeny.fr is a multiple sequence alignment of the 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 (*9*). 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.

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 10^{th} 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 (9).

Phylogeny.fr 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 out-group and produced an overall clade branching pattern with a confidence level of 73%. This suggests an adequate out-group 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 (9).

Protein Domain Homology Modeling

The structure of the C-propeptide domain of the collagen type I alpha I chain protein was explored using some of the previous described technologies. The domain is also called the Fibrillar Collagen C-terminal domain or simply COLFI. This domain is encoded by a 236 amino acid sequence (~14.28% of the entire sequence's length) of the 3' end of RefSeq's NP_000079.2 entry. 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 (14). 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 (9, **Pundir**). 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 (9, **Pundir**). 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 (9, **Pundir**). 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.

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.

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.

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 is included in the Appendix, supports the previously expressed assertion that the second top model seems to contribute little to the overall predicted 3D structure of the COLFI domain. 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.

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) (12, **Chessler**). 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) (12, **Chessler**).

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

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 COL1A1 gene can result in disordered bone formation and bone degeneration. These COL1A1-mutation related degenerative diseases are include osteogenesis imperfecta, types I-IV, Ehlers-Danlos syndrome type VIIA, Ehlers-Danlos, syndrome Classical type, and Caffey Disease) (8). The COL1A1 region of chromosome #17 and a similar sequence on chromosome #22 exhibit a propensity for trans-locution. These 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) (10). The KEGG database of biochemical pathways was queried to determine the extent of COL1A1 mutations responsible for dysfunction or disease. A query string consisting of “COL1A1” was submitted to stringDB through the R statistical programming environment to parse the 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.

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. Mutations of the COL1A1 gene are linked to numerous regenerative 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 (12, **Chessler**). 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 (12, **Chessler**). The mechanism of degeneration is impaired inter-chain disulfide bonds and subsequent abnormal chain integration (14, **Hayashi**)

COL1A1 may also have suppressor function for particular cancers. In a study regarding Hepatocellular carcinoma (HCC), COL1A1 was found to be significantly down regulated at tumor sites (13). No chromosomal mutation could be found and the study identified promoter methylation as the mechanism of expression interference (14, **Hayashi**). Up-regulation 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 (14, **Hayashi**). This is attributable to the growth factor beta gene's subsequent unregulated expression .

A cursory review of GEO profiles produce many instances of COL1Aa1 differential expression. For example, an experiment that performed expression profiling of 22 primary gastric cancer tissues identified significant COL1Aa1 over-expression by the cancer positive group. Although this bioinformatics exploration of COL1A1 is not exhaustive, the presence of a highly conserved COL1A1 gene in many species and the high correlation between diseased function or dysfunction and mutations in the COL1A1 gene suggest that a much more thorough exploration and meta-analysis of COL1A1 is warranted. Further analysis may elucidate the mechanisms of disease and may serve as drug design targets.

Works Cited

1)U.S. Department of Health & Human Services 200 Independence Avenue, S.W.Washington, D.C. 2020. “**HHS Headquarters.**“

2) O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** Nucleic Acids Res. 2016 Jan 4;44(D1):D733-45 PubMed

3) T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik and M. Clamp.**The Ensembl genome database project.** *Nucleic Acids Research* 2002 30(1):38-41.

4) a) Marchler-Bauer A et al. (2017), "**CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.**", Nucleic Acids Res.45(D)200-3.
*4) b)Marchler-Bauer A et al. (2015), "**CDD: NCBI's conserved domain database.**", Nucleic Acids Res.43(D)222-6.
*4) c) Marchler-Bauer A et al. (2011), "**CDD: a Conserved Domain Database for the functional annotation of proteins.**", Nucleic Acids Res.39(D)225-9.
*4) d) Marchler-Bauer A, Bryant SH (2004), "**CD-Search: protein domain annotations on the fly.**", Nucleic Acids Res.32(W)327-331.

5) EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. Copyright © EMBL-EBI 2017 | **EMBL-EBI is part of the European Molecular Biology Laboratory** +44 (0)1223 49 44 44 Intranet

6) Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick Nucleic Acids Res. 2005 Jan 1; 33(Database Issue): D514–D517. Published online 2004 Dec 17. doi: 10.1093/nar/gki033 PMCID: PMC539987 “**Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**”

7) Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database**

search programs. Nucleic Acids Res. 1997, 25 (17): 3389-3402. 10.1093/nar/25.17.3389.

8) NCBI-NLM, National Center for Biotechnology Information, U.S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, 20894 USA **Homologene**

9) Pundir S, Magrane M, Martin MJ, O'Donovan C, UniProt Consortium. **Searching and Navigating UniProt Databases** Curr. Protoc. Bioinformatics 50:1.27.1-1.27.10 (2015)

8) Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C." **The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.** " Nucleic Acids Res. 2017 Jan; 45:D362-68.PubMed

9) a) Dereeper A.*, Guignon V.*, Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. **Phylogeny.fr: robust phylogenetic analysis for the non-specialist**. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19.

*9) b) Dereeper A., Audic S., Claverie J.M., Blanc G. **BLAST-EXPLORER helps you building datasets for phylogenetic analysis**. BMC Evol Biol. 2010 Jan 12;10:8. (PubMed)

10) 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**,"

11) Young SM, Bansal P, Vella ET, Finelli A, Levitt C, Loblaw A, Can Fam Physician. Jan 61 (1) pgs26-35 (2015), "**Systematic review of clinical features of suspected prostate cancer in primary care**."

12) Chessler SD[1], Wallis GA, Byers PH. J Biol Chem. 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**."

13) Nakamura I, Kariya Y, Okada E, Yasuda M, Matori S, Ishikawa O, Uezato H, Takahashi K.JAMA Dermatol. 2015 Sep 2 "**A Novel Chromosomal Translocation Associated With COL1A2- PDGFB Gene Fusion in Dermatofibrosarcoma Protuberans: PDGF Expression as a New Diagnostic Tool.**"

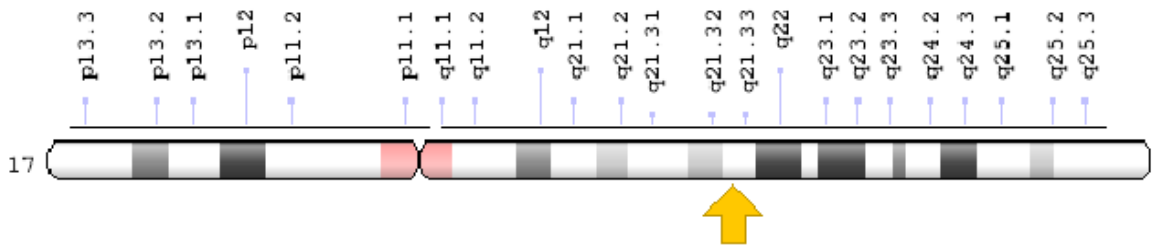
14) Hayashi M, Nomoto S, Hishida M, Inokawa Y, Kanda M, Okamura Y, Nishikawa Y, Tanaka C, Kobayashi D, Yamada S, Nakayama G, Fujii T, Sugimoto H, Koike M, Fujiwara M, Takeda S, Kodera Y.BMC Cancer. 2014 Feb 19 "**Identification of the collagen type 1 α 1 gene (COL1A1) as a candidate survival-related factor associated with hepatocellular carcinoma.**"

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

16) Xiran Wang, Yu Pei, Jingtao Dou, Juming Lu, Jian Li, and Zhaohui Lv Genet Mol Biol. 2015 Mar; 38(1): 1–7 "**Identification of a novel *COL1A1* frameshift mutation, c.700delG, in a Chinese osteogenesis imperfecta family**"

17) Bairoch A., Boeckmann B., Ferro S., Gasteiger E. **Swiss-Prot: juggling between evolution and stability** Brief. Bioinform. 5:39-55 (2004)

Image 1: Cytogenic Chromosomal Location of COL1A1:



C
OL1
A1

Image 2: RefSeq Graphical View of Collagen Alpha-1(I) Chain Protein [Homo sapiens]

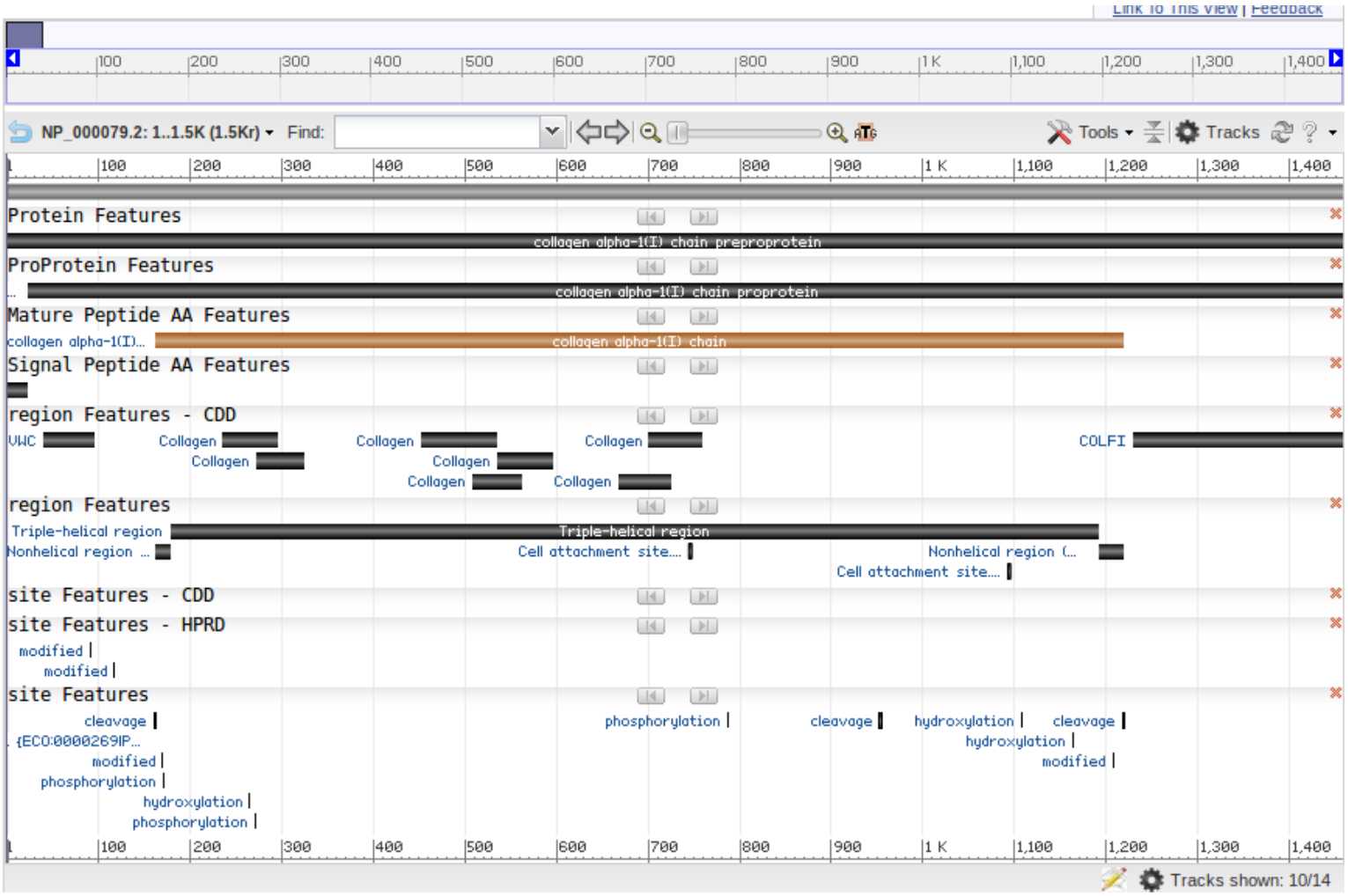


Image 3: Ensembl's Graphical View of Collagen Type I Alpha-1(I) Splice Variants and Domains:

(Chromosome #17 from Ensembl's Genome release #88)

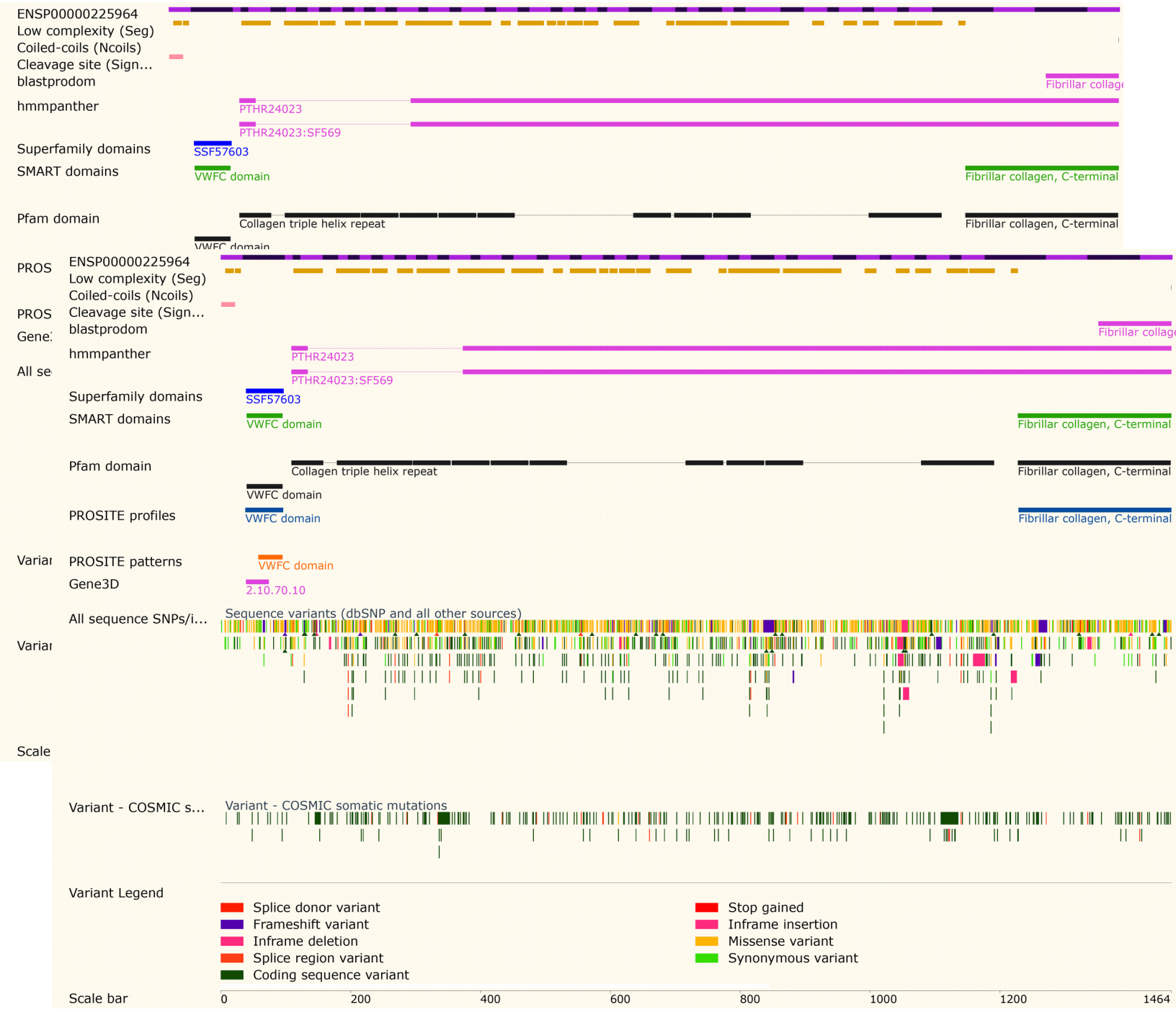


Image 4: Ensembl's Graphical View of Collagen Type I Alpha-1(I) Protein:

(Chromosome #17 from Ensembl's Genome release #88)

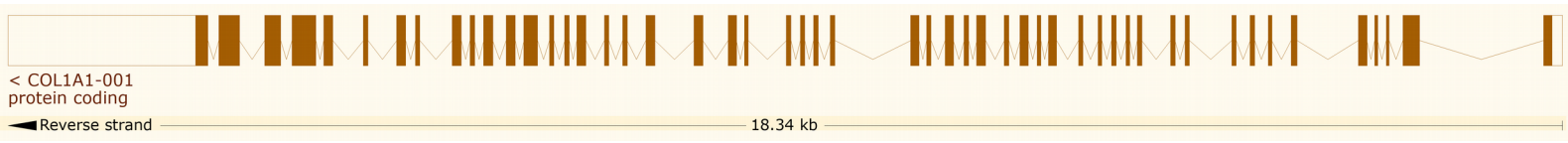


Image 5: COL1A1 Entry from the Plasma Proteome Database

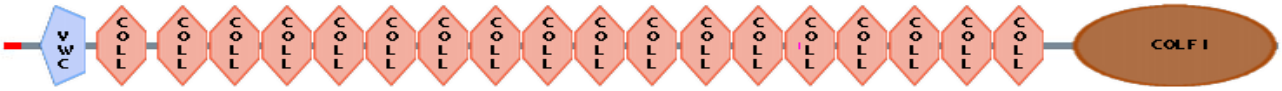


Image 6: COL1A1 Conserved Domain Analysis with Non-Specific Hits from CDD

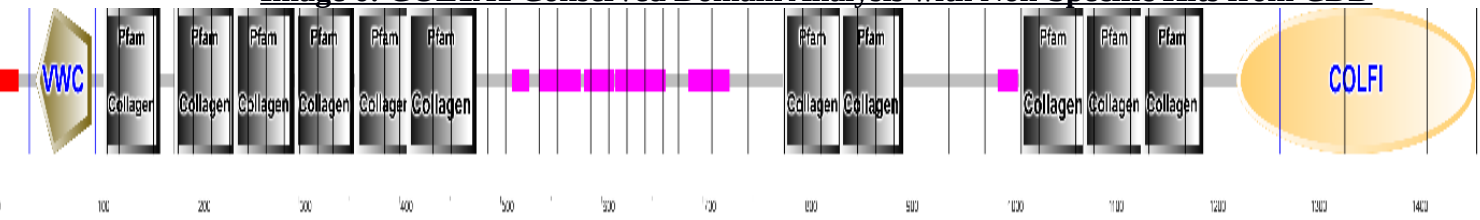


Image 7: CDD's COL1A1's Conserved Domains Specific Hits

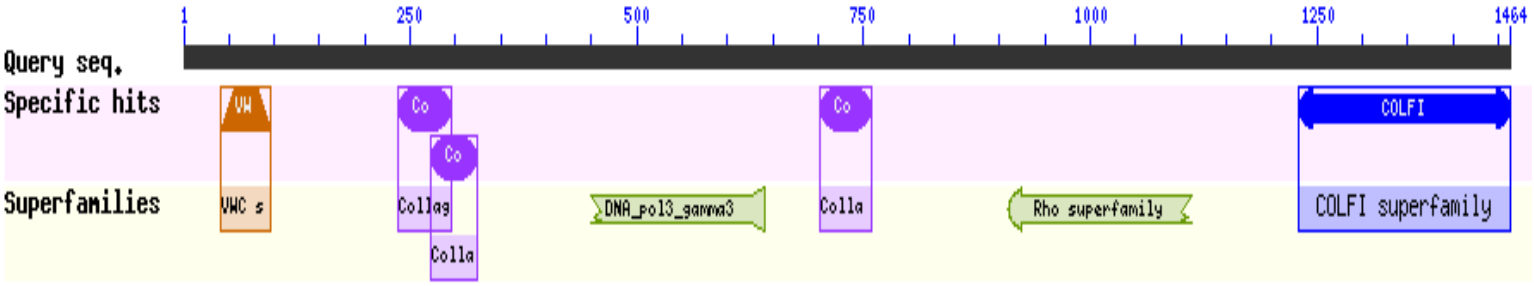


Image 8: CDD's COL1A1's Conserved Domains at Residue Level and w/ Non Specific Hits

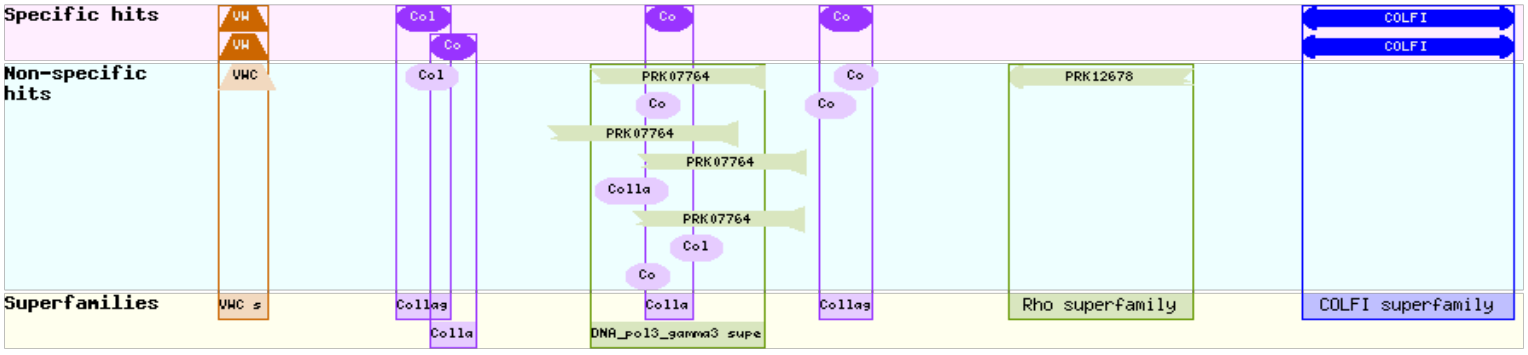


Image 9: COL1A1's Conserved Domains from Genenames.com

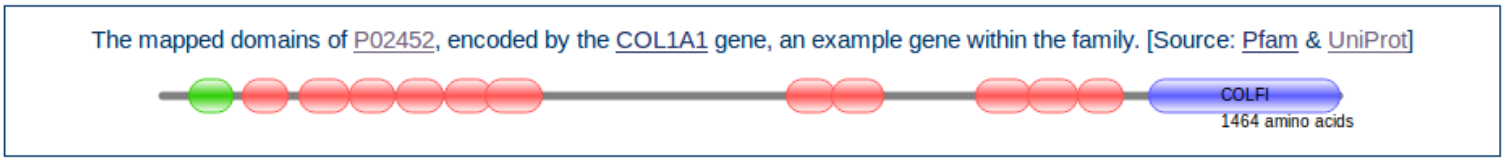


Table 1: CDD List of Domain Hits for COL1A1:

Name	Accession	Description	Interval	E-value
COLFI	pfam01410	Fibrillar collagen C-terminal domain; Found at C-termini of fibrillar collagens: Ephydatia ...	1230-1463	2.65e-154
VWC	pfam00093	von Willebrand factor type C domain; The high cutoff was used to prevent overlap with ...	40-95	3.05e-21
Collagen	pfam01391	Collagen triple helix repeat (20 copies); Members of this family belong to the collagen ...	236-295	6.95e-09
DNA_pol3_gamma3 super family	cl26386	DNA polymerase III subunits gamma and tau domain III; This domain family is found in bacteria, ...	449-640	8.55e-07
Collagen	pfam01391	Collagen triple helix repeat (20 copies); Members of this family belong to the collagen ...	701-759	2.66e-06
Collagen	pfam01391	Collagen triple helix repeat (20 copies); Members of this family belong to the collagen ...	272-324	6.70e-05
Rho super family	cl28310	Transcription termination factor Rho [Transcription];	908-1111	8.23e-03

Table 2: Multiple Sequence Alignment MUSCLE Output (Outgroup is algea sequence):

```
gi|algea|   rvdsYqahearqvad----qladEqRHVsl------------------FaYGvGrGvDr
gi|seaturt  ETCVYPTQattAQKNWYISKNPKEKkHVWFGEtMsDGFQ----------FEYG-GEGSnP
gi|Alligat  ETCVhPTQatIAQKNWYmSKNPKEKkHiWFGEtMsDGFQ----------FEYG-GEGSnP
gi|Cricetu  qTCVfPTQPtVpQKNWYISpNPKEKeHVWFGESMTDGFQ----------FEYG-sEGSDP
gi|shrew|   ETCVYPTQPSVAkKNWYvSKN-KdKRHVWFGESMThGFQvltrsslfssFpcs-ssdSDP
gi|Brandtb  ETCVYPTQPtVAQKNWYISKNPKEKkHVWFGESMTgGFQ----------FEYG-GqdSDP
gi|mouse|   qTCVfPTQPSVpQKNWYISpNPKEKkHVWFGESMTDGFp----------FEYG-sEGSDP
gi|rat|     qTCVfPTQPSVpQKNWYISpNPKEKkHVWFGESMTDGFQ----------FEYG-sEGSDP
gi|Donkey|  ETCVYPTQPqVAQKNWYISKNPKdKRHVWyGESMTDGFQ----------FEYG-GqGSDP
gi|HumanCO  ETCVYPTQPSVAQKNWYISKNPKdKRHVWFGESMTDGFQ----------FEYG-GqGSDP
gi|Cow|     ETCVYPTQPSVAQKNWYISKNPKEKRHVWyGESMTgGFQ----------FEYG-GqGSDP
gi|Doggie|  ETCVYPTQPqVAQKNWYISKNPKEKRHVWyGESMTDGFQ----------FEYG-GqGSDP
```

Table 2 (continued): MSA- MUSCLE Output (Outgroup is algea sequence):

```
gi|algea|          nvvflvdgsGsvnaeefeamlgFcvdasnqlAesvpnl---qvAvVqfsnDvrVevgLap
gi|seaturt         GPPGPPGPPGPPGPPSGGFDFSFLPQPPQEKAHtdsRYYRADDANVmRDRDLEVDTTLKS
gi|Alligat         GPPGPPGPPGPPGaPSGGFDFSFmPQPPQEKAHDpGRYYRADDANVmRDRDLEVDTTLKS
gi|Cricetu         GPPGPPGPPGPPGPPSGGyDFSFLPQPPQEKsHDGGRYYRADDANVVRDRDLEVDTTLKS
gi|shrew|          GPPGPPGPPGPPGPPSGGFDFSFLPQPPQEKAqDGGRYYRADDANVVRDRDLEVDTTLKS
gi|Brandtb         GPPGPPGPPGPPGPPSGGFDFSFmPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS
gi|mouse|          GPPGPPGPPGPPGPPSGGyDFSFLPQPPQEKsqDGGRYYRADDANVVRDRDLEVDTTLKS
gi|rat|            GPPGPPGPPGPPGPPSGGyDFSFLPQPPQEKsqDGGRYYRADDANVVRDRDLEVDTTLKS
gi|Donkey|         GPPGPPGPPGPPGPPSaGFDFSFLPQPPQEKsHDGGRYYRADDANVVRDRDLEVDTTLKS
gi|HumanCO         GPPGPPGPPGPPGPPSaGFDFSFLPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS
gi|Cow|            GPPGPPGPPGPPGPPSGGyDlSFLPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS
gi|Doggie|         GPPGPPGPPGPPGPPSGGFDFSFLPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKS
```

Table 3: Multiple Sequence Alignment preparation for Phylogeny.fr Analysis

```
                        1510      1520      1530      1540      1550      1560
                        =========+=========+=========+=========+=========+=========+
gi|algea|               RVDSYQAHEARQVAD----QLADEQRHVSL-------------------FAYGVGRGVDR
gi|seaturtle|           ETCVYPTQATTAQKNWYISKNPKEKKHVWFGETMSDGFQ----------FEYG-GEGSNP
gi|Alligator|           ETCVHPTQATIAQKNWYMSKNPKEKKHIWFGETMSDGFQ----------FEYG-GEGSNP
gi|Cricetulus]          QTCVFPTQPTVPQKNWYISPNPKEKEHVWFGESMTDGFQ----------FEYG-SEGSDP
gi|shrew|               ETCVYPTQPSVAKKNWYVSKN-KDKRHVWFGESMTHGFQVLTRSSLFSSFPCS-SSDSDP
gi|Brandtbat|           ETCVYPTQPTVAQKNWYISKNPKEKKHVWFGESMTGGFQ----------FEYG-GQDSDP
gi|mouse|               QTCVFPTQPSVPQKNWYISPNPKEKKHVWFGESMTDGFP----------FEYG-SEGSDP
gi|rat|                 QTCVFPTQPSVPQKNWYISPNPKEKKHVWFGESMTDGFQ----------FEYG-SEGSDP
gi|Donkey|              ETCVYPTQPQVAQKNWYISKNPKDKRHVWYGESMTDGFQ----------FEYG-GQGSDP
gi|HumanCOL1A1|         ETCVYPTQPSVAQKNWYISKNPKDKRHVWFGESMTDGFQ----------FEYG-GQGSDP
gi|Cow|                 ETCVYPTQPSVAQKNWYISKNPKEKRHVWYGESMTGGFQ----------FEYG-GQGSDP
gi|Doggie|              ETCVYPTQPQVAQKNWYISKNPKEKRHVWYGESMTDGFQ----------FEYG-GQGSDP
                        ##########################################          #######
```

Conserved positions: **26%**

Image 10: Bootstrapped Phylogenic Tree of COL1A1 Non-Human Homologous Proteins

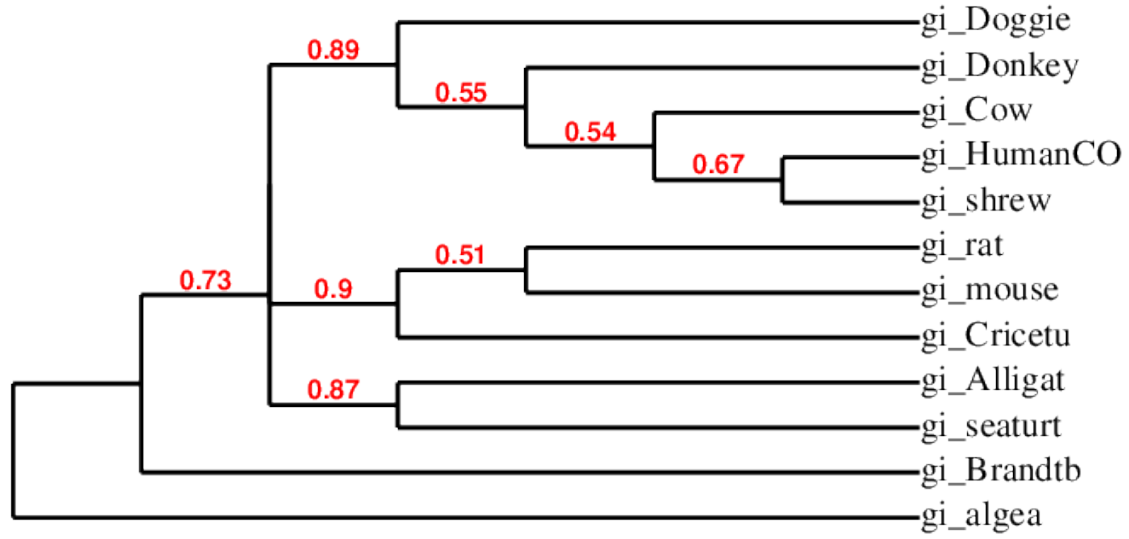


Image 11: Jmol three dimensional protein structure viewer Interface for Phylogeny.fr

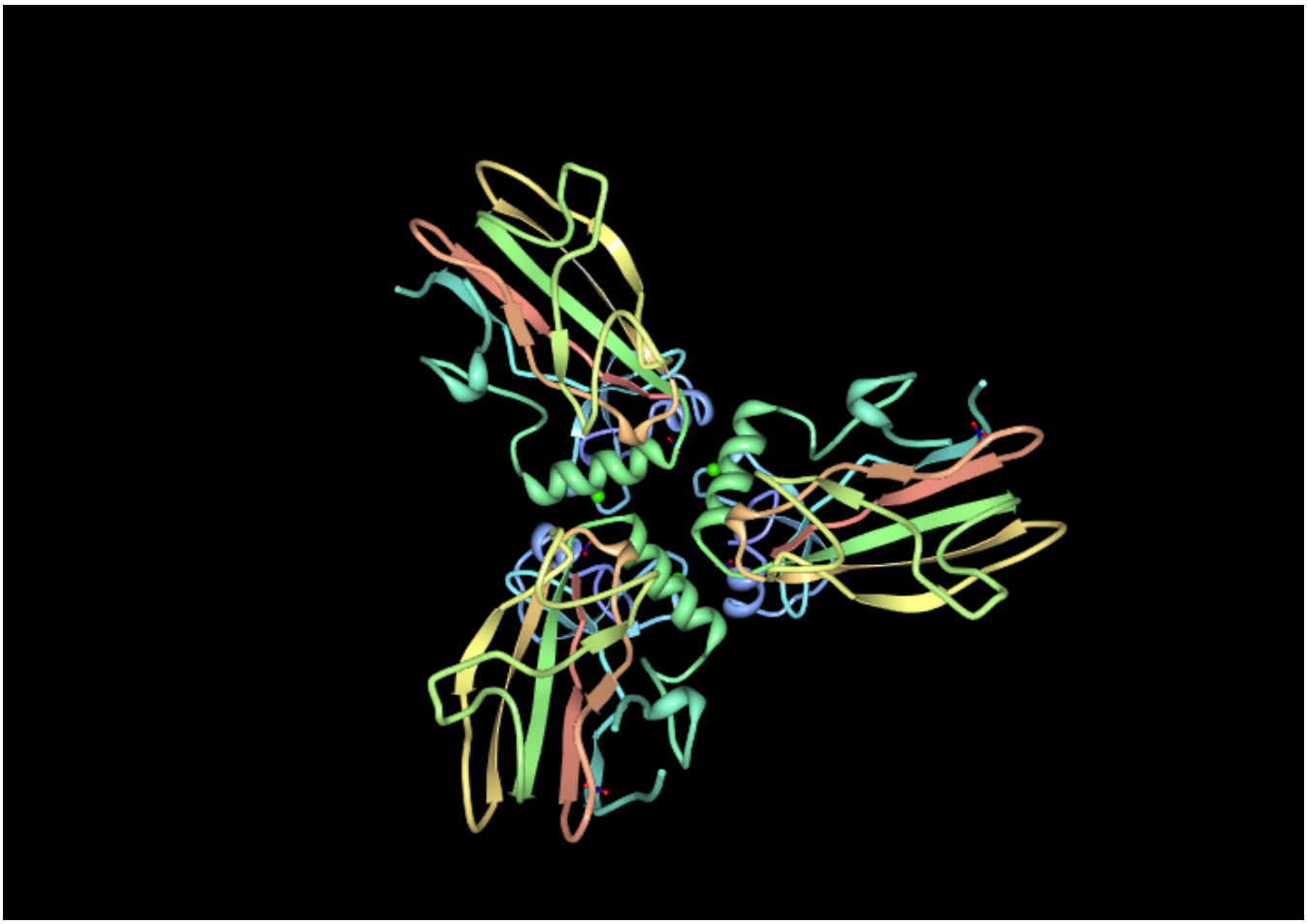


Image 12: Jmol three dimensional protein structure viewer Interface for Phylogeny.fr

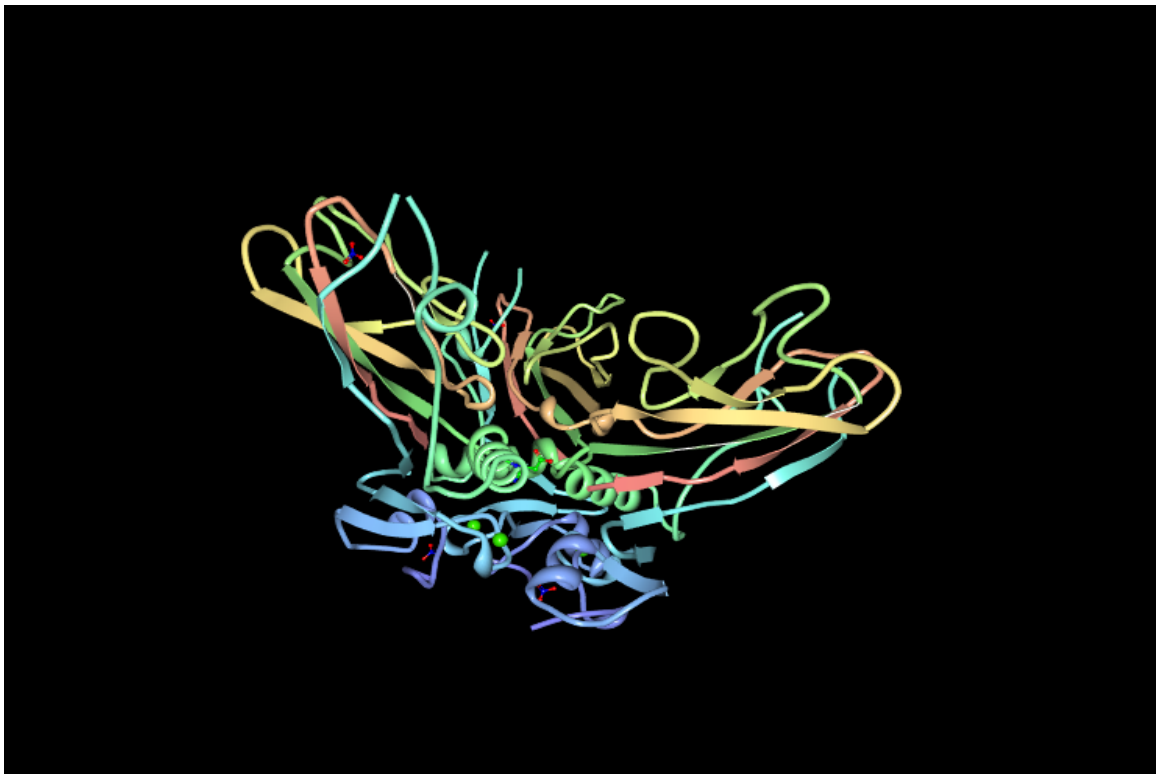


Image 13: Jmol three dimensional protein structure viewer Interface for Phylogeny.fr

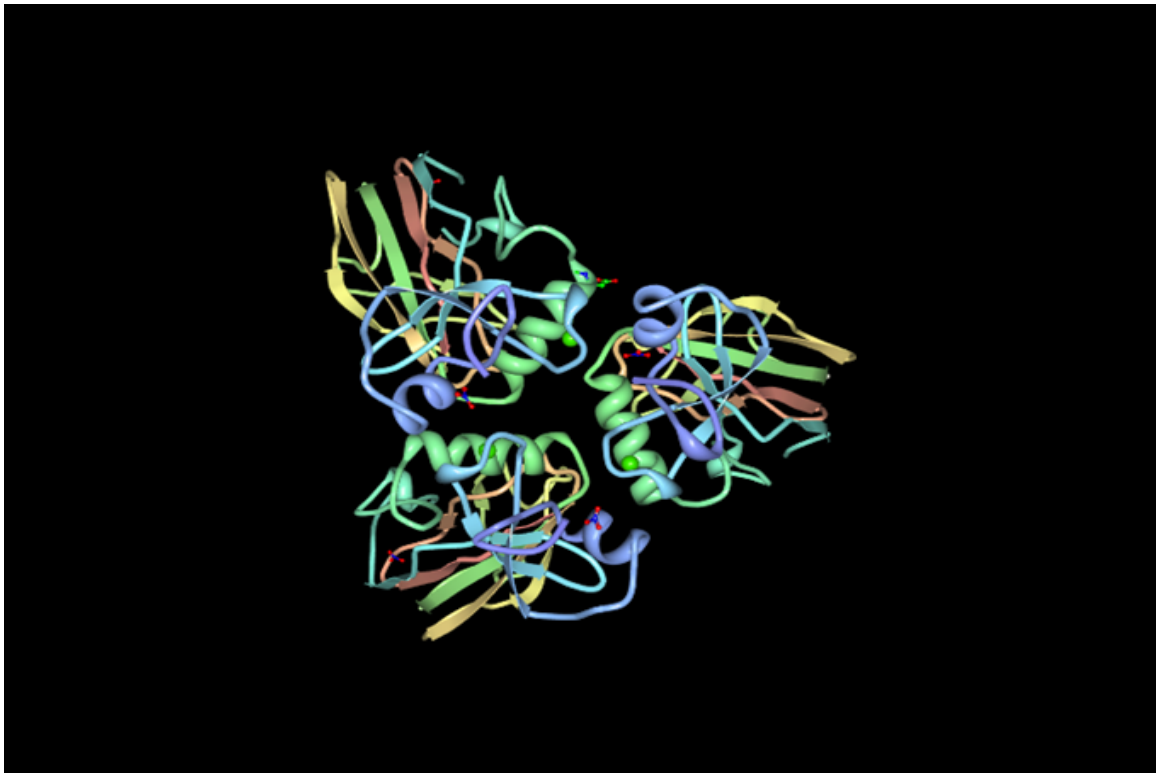


Image 14: Jmol three dimensional protein structure viewer Interface for Phylogeny.fr

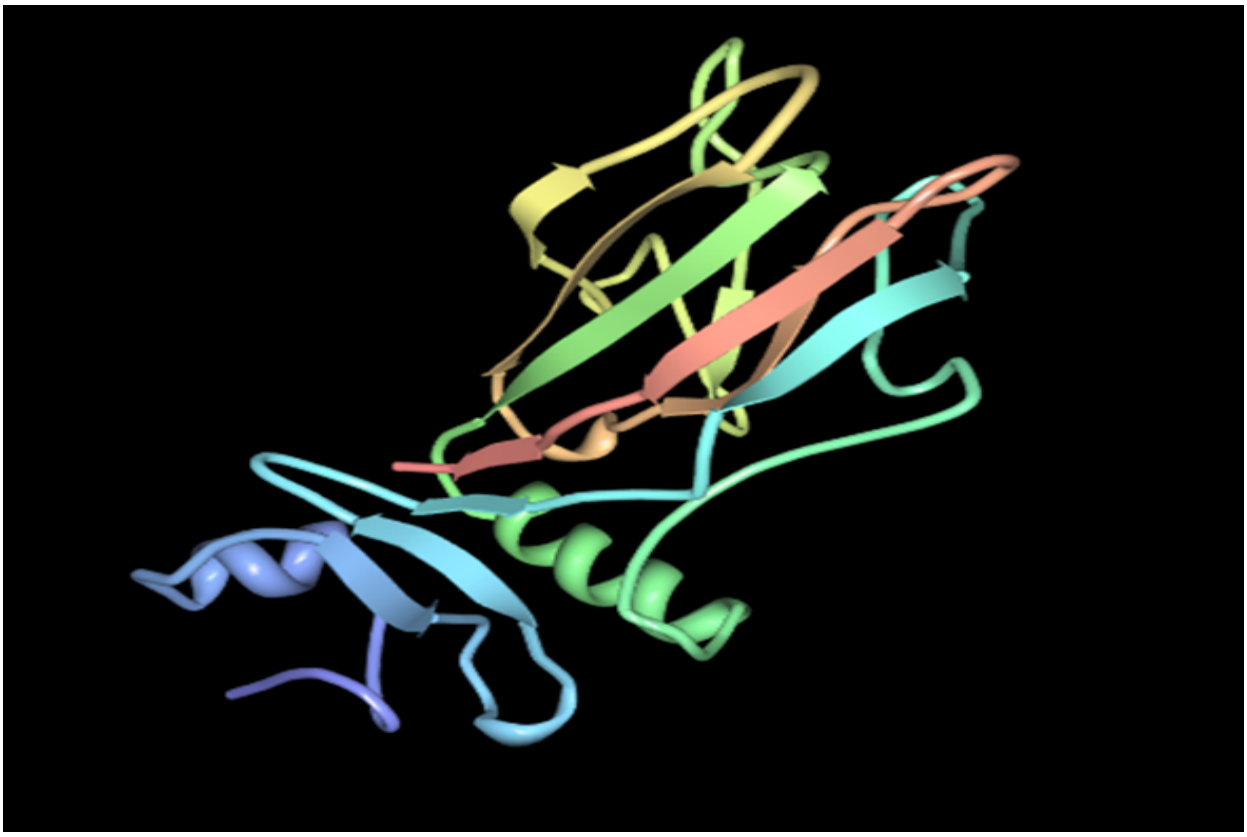


Image 15: Phyre2 Image 1 -API (Interface) after completion of COL1A1 Homology Modeling

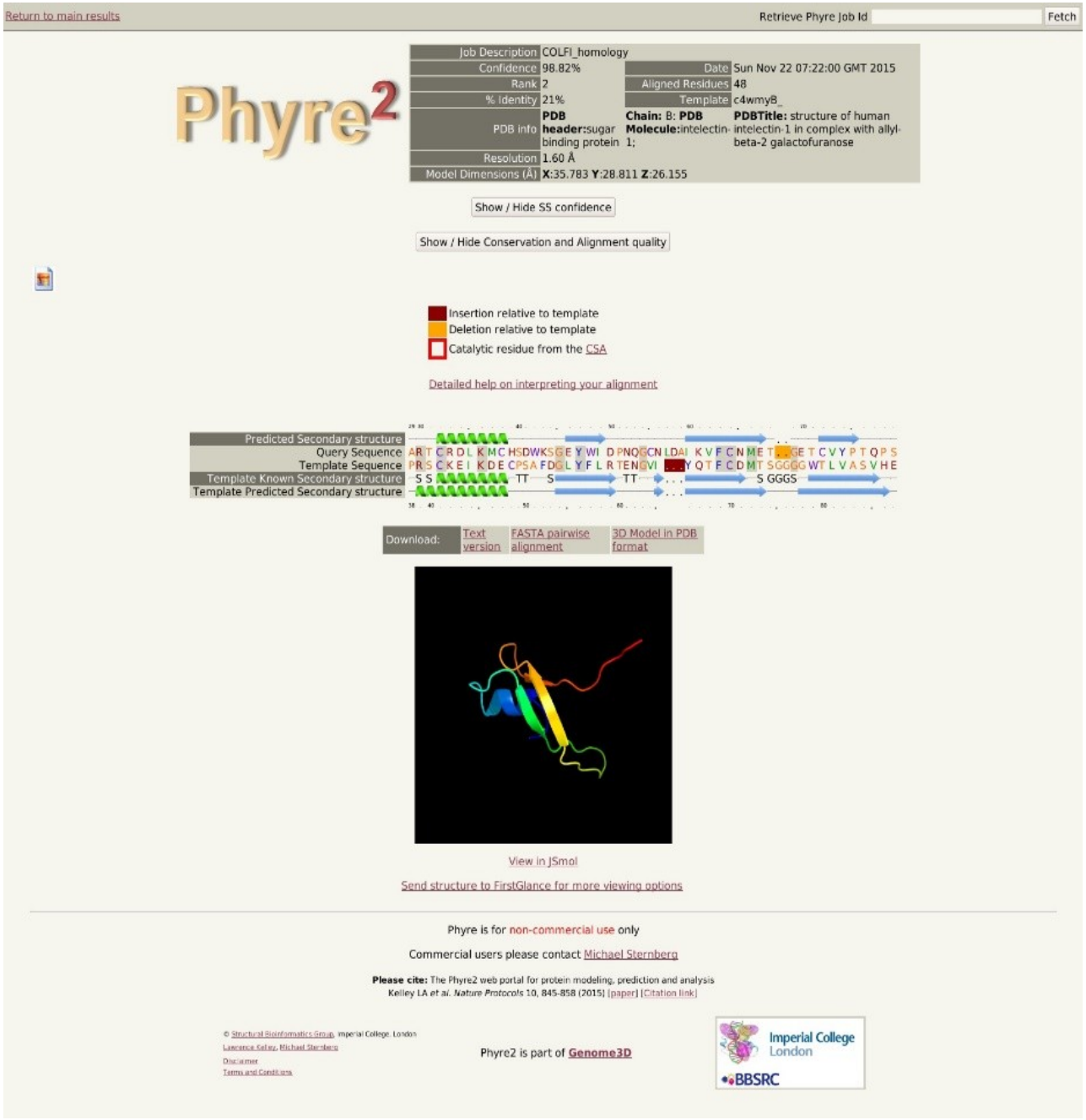


Image 16: Phyre2 Image 2 -API (Interface) after completion of COL1A1 Homology Modeling

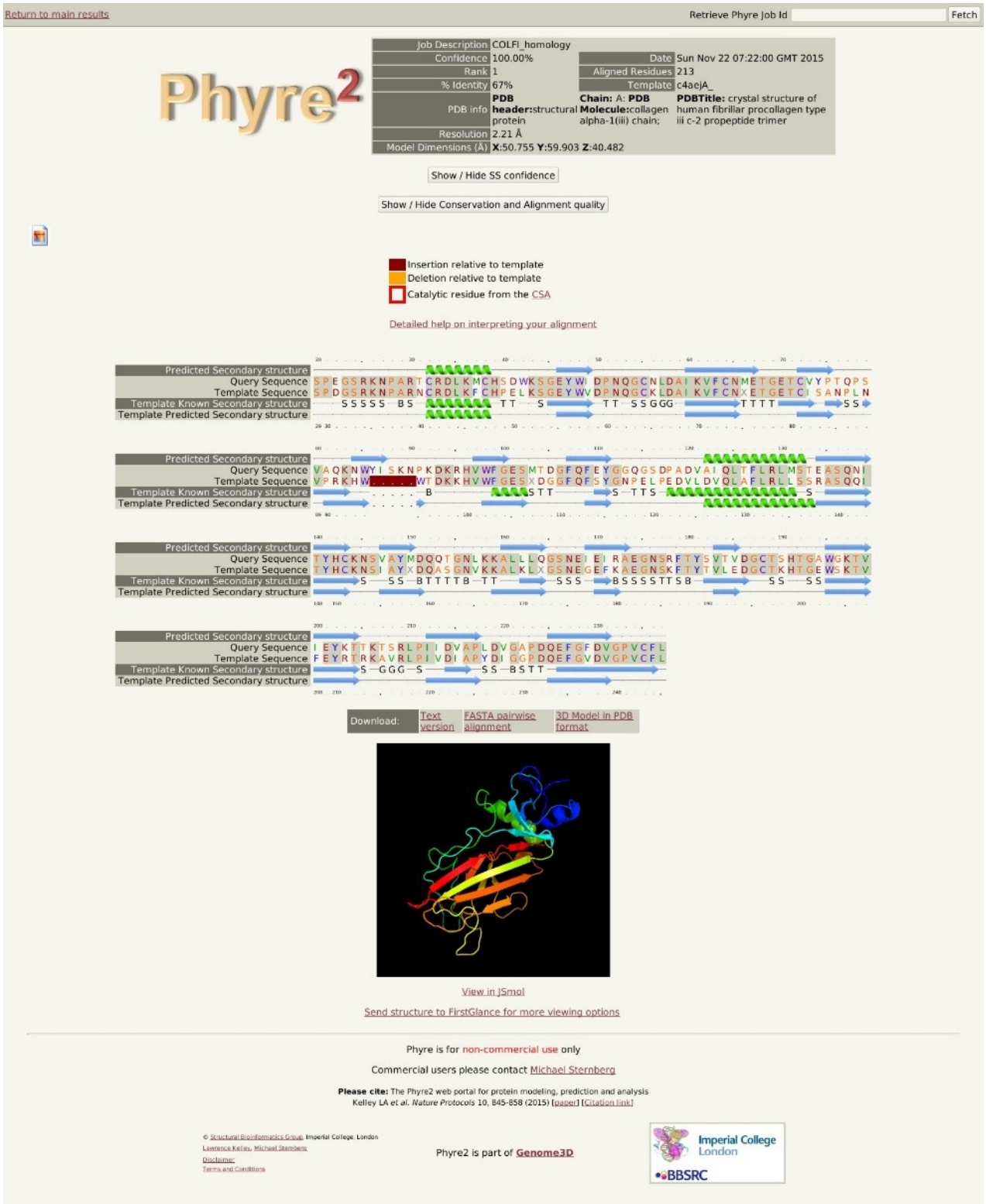


Image 17: StringDB Collagen Type I Alpha-1 Network I Analysis:

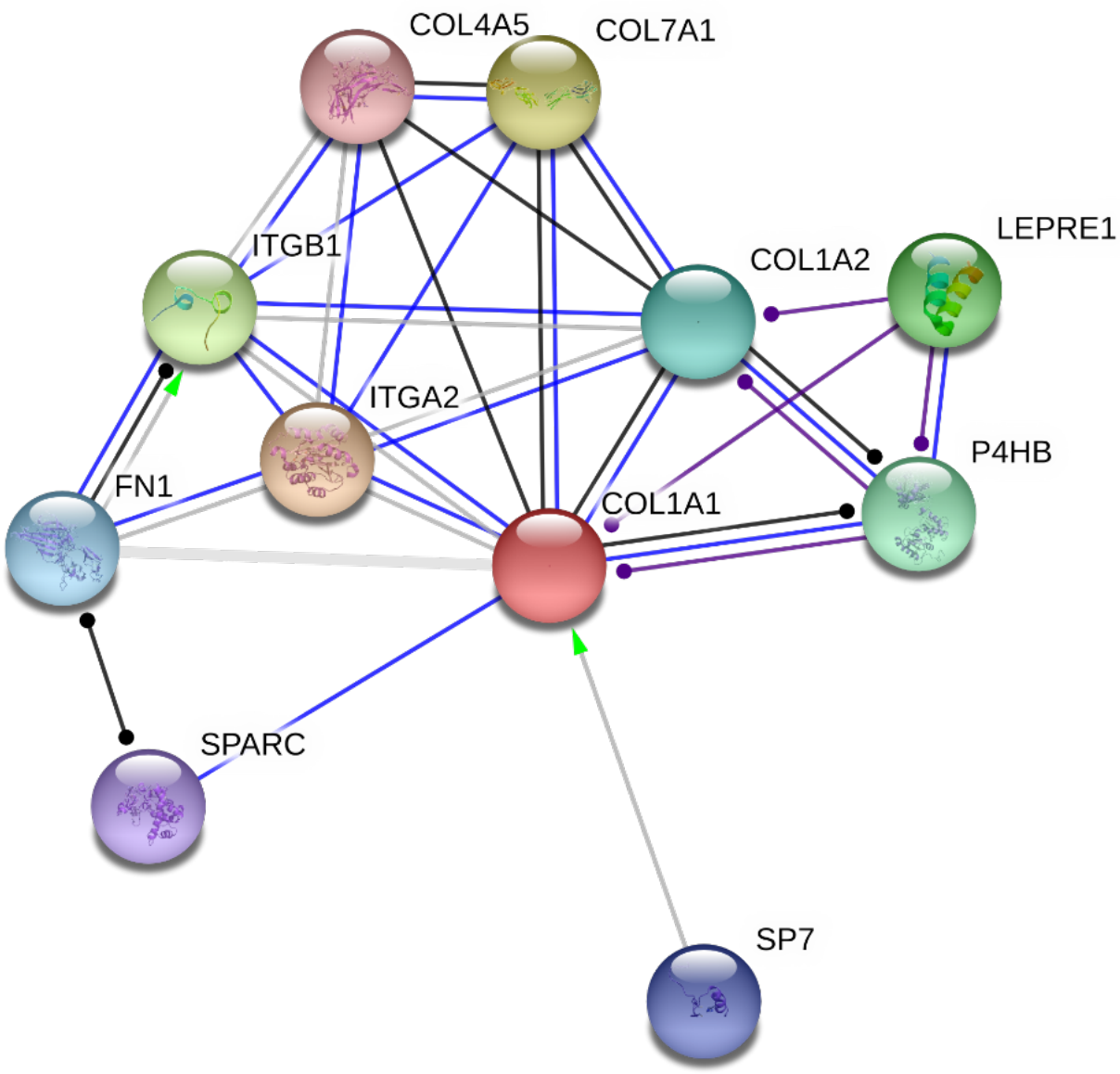


Table 4: String DB's Charaterization of COL1A1-Involved Protein-Protein Interactions:

			Activation	Inhibition	Binding	Phenotype	Catalysis	Post-transl. m	Reaction	Expression	Score
	ITGA2	integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor); Integrin alpha-2/beta-1 is a rece [...] (1181 aa)	•		•						0.985
	COL7A1	collagen, type VII, alpha 1; Stratified squamous epithelial basement membrane protein that form [...] (2944 aa)			•				•		0.977
	ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) (798 aa)	•		•						0.973
	LEPRE1	leucine proline-enriched proteoglycan (leprecan) 1 (736 aa)					•				0.966
	P4HB	prolyl 4-hydroxylase, beta polypeptide; This multifunctional protein catalyzes the formation, b [...] (508 aa)			•		•		•		0.964
	COL1A2	collagen, type I, alpha 2 (1366 aa)			•				•		0.963
	FN1	fibronectin 1 (2477 aa)									0.957
	SP7	Sp7 transcription factor; Transcriptional activator essential for osteoblast differentiation. B [...] (431 aa)	•								0.956
	SPARC	secreted protein, acidic, cysteine-rich (osteonectin); Appears to regulate cell growth through [...] (303 aa)			•						0.951
	COL4A5	collagen, type IV, alpha 5; Type IV collagen is the major structural component of glomerular ba [...] (1691 aa)							•		0.949

Image 18: StringDB Collagen Type I Alpha-1 Coexpression Conservation Analysis of Network I:

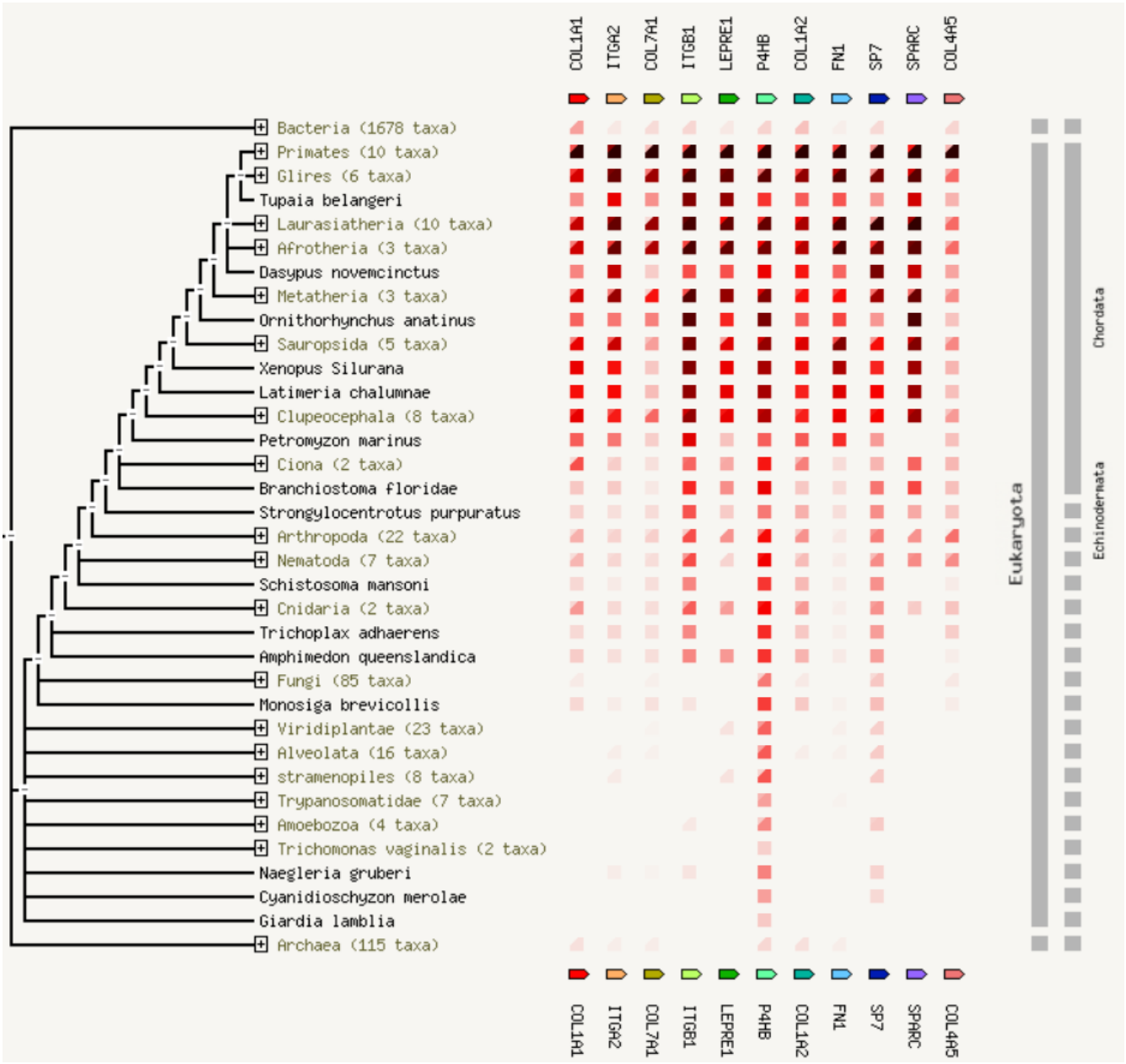


Image 19: StringDB Sources of Collagen Type I Alpha-1 Network I Interaction Partners:

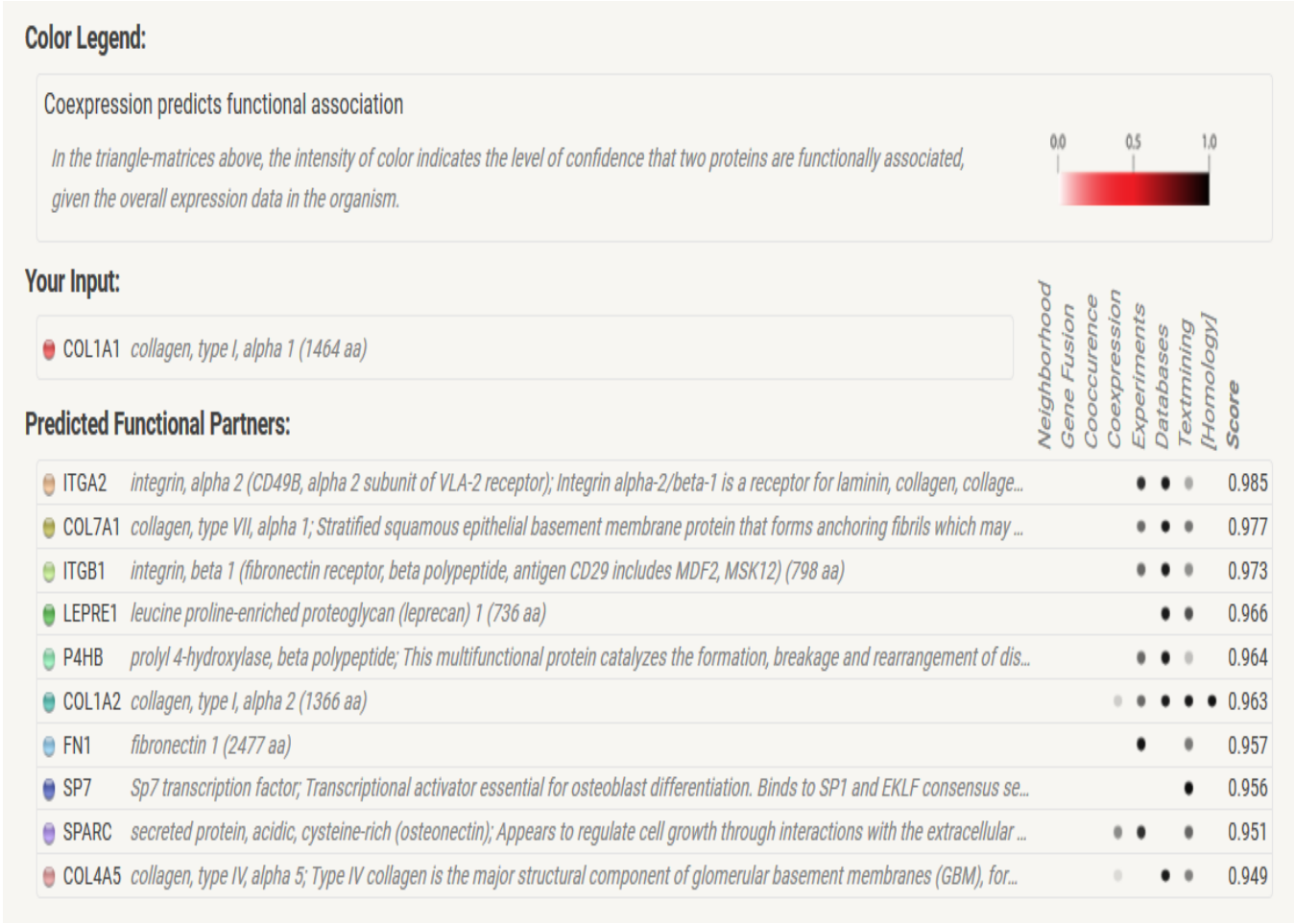


Image 20: StringDB Collagen Type I Alpha-1 and ZNF Network:

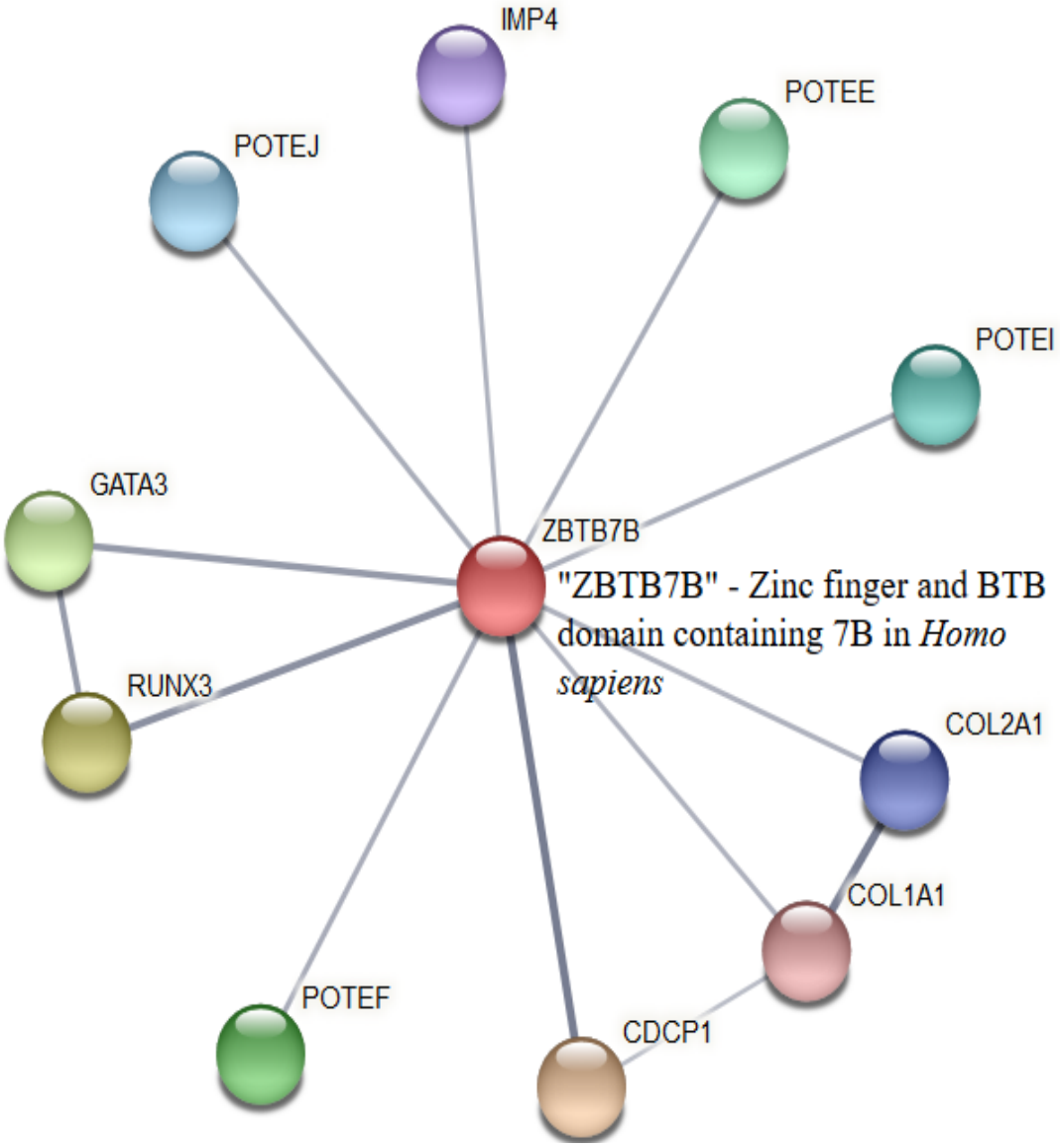


Image 21: StringDB Conservation of COL1A1 Coexpression with ZNF384:

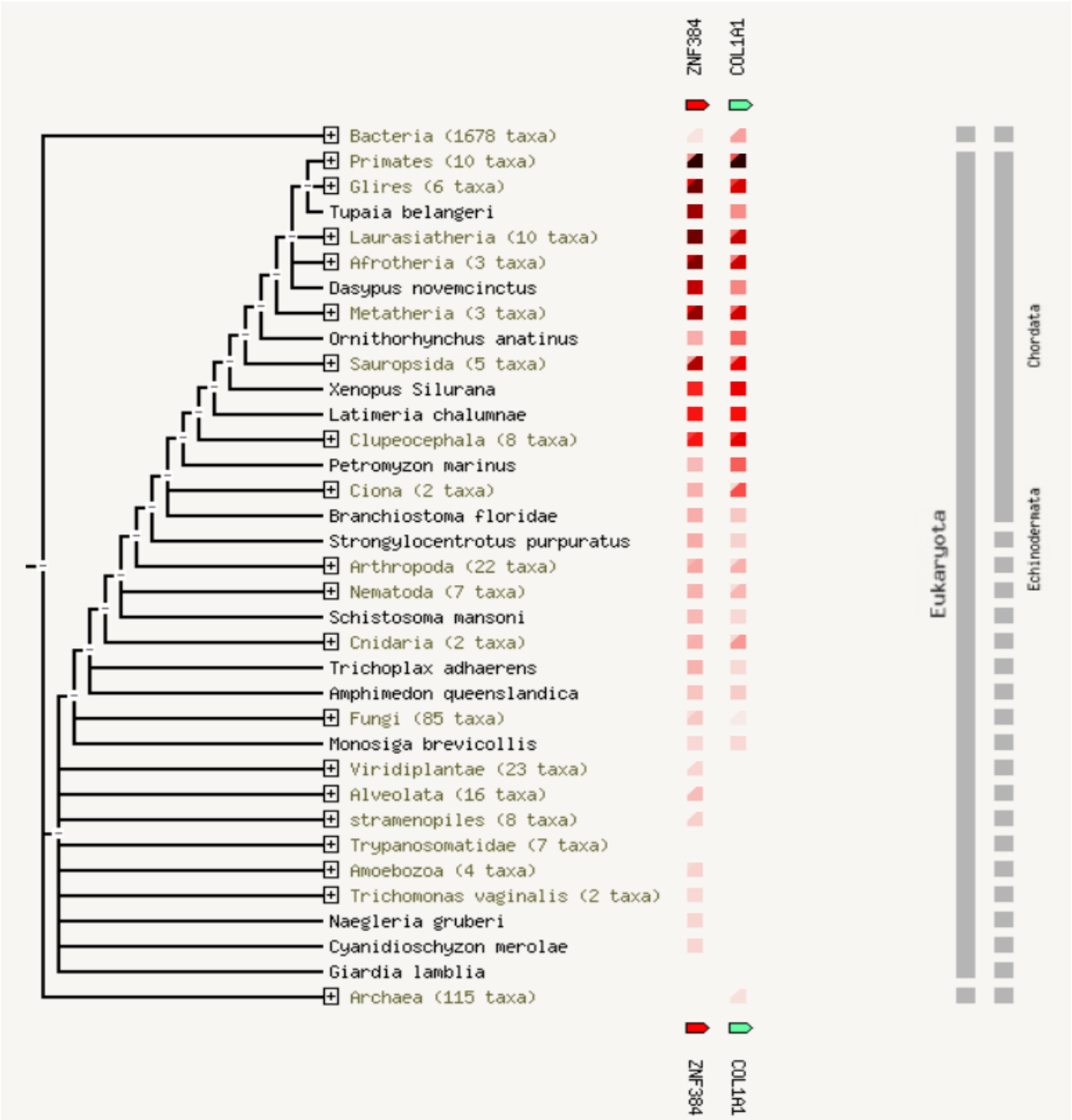


Image 22: GEO Profile 1 Expression Profiling Gastric Cancer

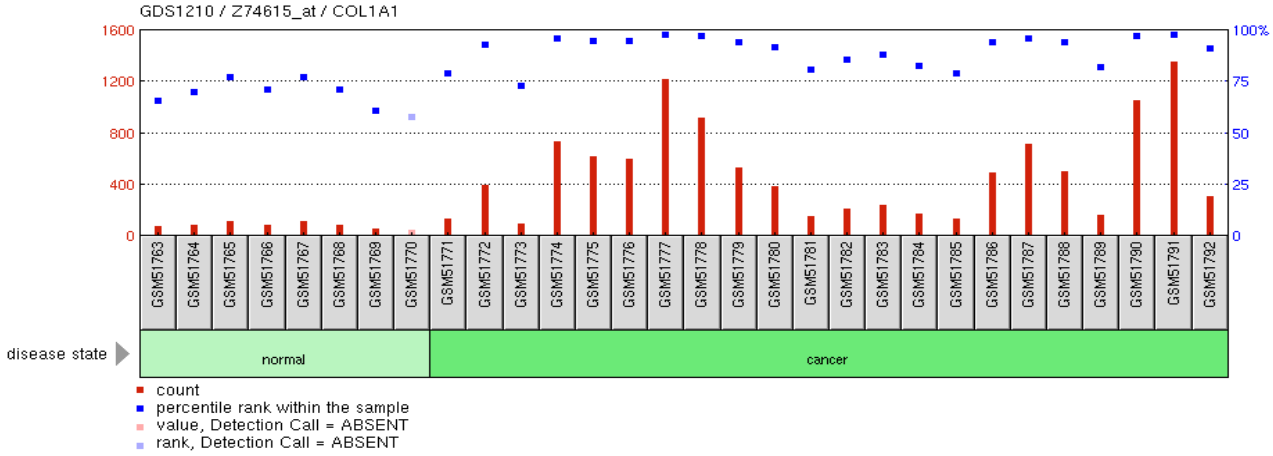


Image 23: GEO Profile 2: Expression Profiling Papillary Thyroid Carcinoma (PTC)

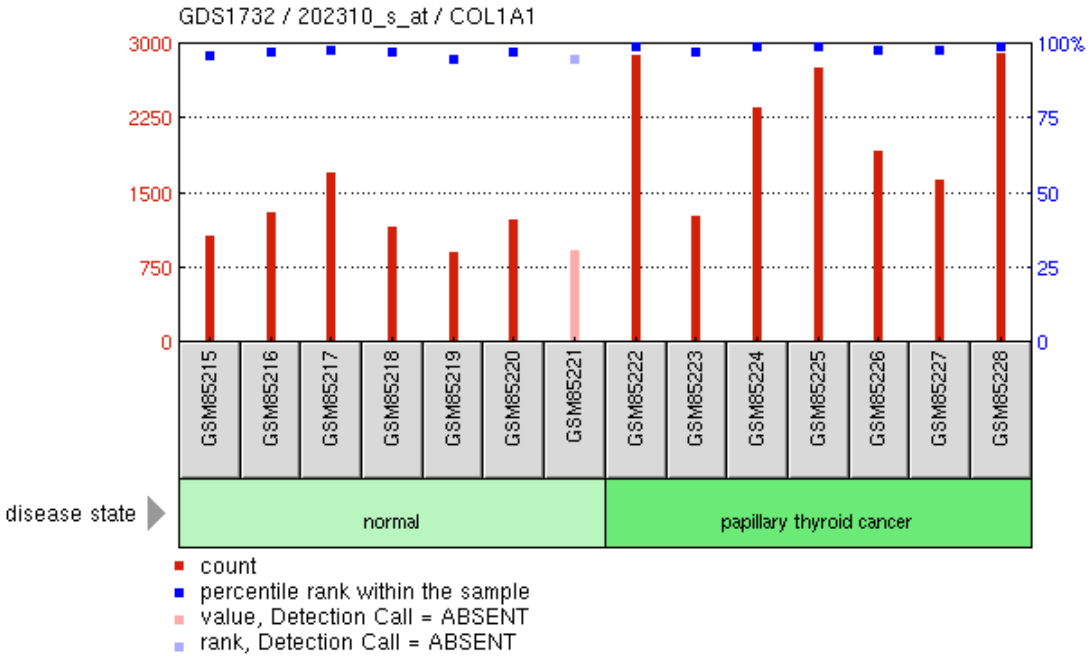


Image 24: GEO Profile 3 - Comparative Panc-1 celll-Line Analysis

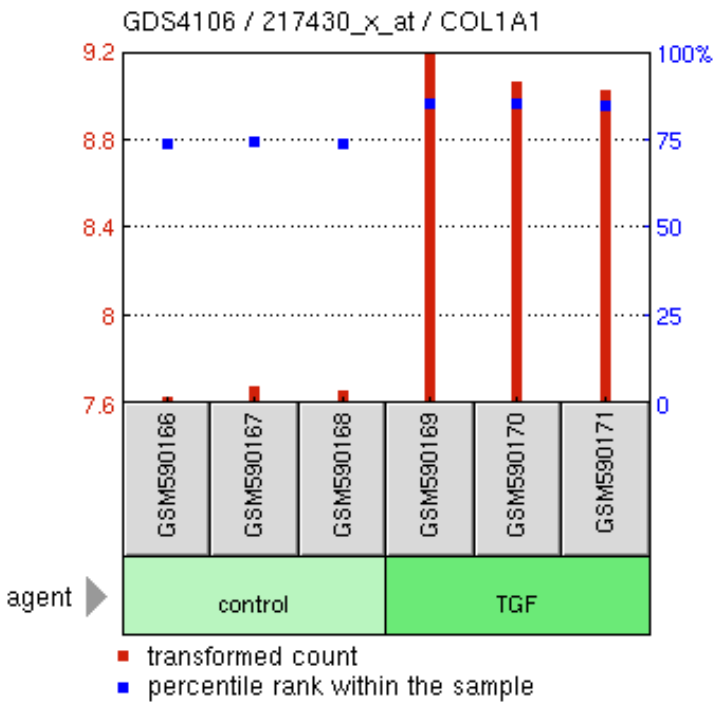


Image 25 - Protein Structure of COL1A1 from SWISS-PROT Modeling: