BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(1)

Analysis of human collagen sequences

Manisha Nassa, Pracheta Anand, Aditi Jain, Aastha Chhabra, Astha Jaiswal, Umang Malhotra, Vibha Rani*

Department of Biotechnology, Jaypee Institute of Information Technology, A-10, Sector-62, NOIDA, 201307, Uttar Pradesh, India; Vibha Rani – Email: vibha.rani@jiit.ac.in; phone: + (91)-120-2594210; *Corresponding author

Received December 07, 2011; Accepted December 16, 2011; Published January 06, 2012

Abstract
The extracellular matrix is fast emerging as important component mediating cell-cell interactions, along with its established role as a scaffold for cell support. Collagen, being the principal component of extracellular matrix, has been implicated in a number of pathological conditions. However, collagens are complex protein structures belonging to a large family consisting of 28 members in humans; hence, there exists a lack of in depth information about their structural features. Annotating and appreciating the functions of these proteins is possible with the help of the numerous biocomputational tools that are currently available. This study reports a comparative analysis and characterization of the alpha-1 chain of human collagen sequences. Physico-chemical, secondary structural, functional and phylogenetic classification was carried out, based on which, collagens 12, 14 and 20, which belong to the FACIT collagen family, have been identified as potential players in diseased conditions, owing to certain atypical properties such as very high aliphatic index, low percentage of glycine and proline residues and their proximity in evolutionary history. These collagen molecules might be important candidates to be investigated further for their role in skeletal disorders.

Keywords: Biocomputational tools, Collagen, Comparative characterization, Extracellular matrix

Background:
The extracellular matrix (ECM) is an intricate network of macromolecules surrounding a substantial volume of cells. It comprises of collagens, proteoglycans, glycoproteins and proteases **[1].** These components are arranged in a highly organized manner and play a significant role not only as a scaffold to the cell but also in multiple processes such as cell migration, cell-cell interaction and cell proliferation **[2].** Collagens are the most abundant component of ECM. They form a triple helical structure with three distinct polypeptide chains, commonly known as the alpha chains. Also, this triple helix is found to possess a peculiar sequence 'Gly-Xaa-Yaa' **[3].** The presence of glycine as every third residue accounts for the stability of the helical structure owing to its property of being the smallest amino acid. Xaa and Yaa can be any amino acid but are mostly occupied by the proline residue. Thus, collagen is known to be a glycine and proline rich entity. Collagen proteins are synthesized as inactive precursor forms known as pro-collagens. The cleavage of pro-peptides present at the N and C terminal by peptidases forms the mature active collagen molecules.

The collagen family is a large and a complex family comprising of 28 genetically distinct members found in humans **[4].** This diverse family includes fiber forming collagens, amorphous collagens, transmembrane collagens and a specialized type of collagen that forms unique structures **[5].** The imperative role of collagen can be ascertained through the wide spectrum of pathological disorders that are associated with it. Mutations in the genes encoding for collagen proteins can lead to a variety of diseases such as Osteogenesis Imperfecta, Ehlers-Danlos syndrome, Spondyloepiphyseal dysplasia, Multiple epiphyseal dysplasia **[6].** Furthermore, variations in the collagen content or a significant remodeling of the collagen network can lead to several dysfunctions like parenchymal disease, hypertensive heart disease, renal fibrosis, tumor and fibrotic diseases **[7].** Computational biology is an emerging field that essentially helps to unveil the hidden information of a protein structure,

both at the genomics and proteomics level. It integrates biology with computational algorithms for better understanding of complex molecules. Such analysis can be enhanced through a combinatorial dry to wet lab approach wherein the propositions made through the biocomputational findings can help in providing a direction for further research during wet lab studies. Thus, it facilitates to appreciate and understand the structural and functional roles of protein molecules, which are the heart of human disorders. A similar approach has been adopted to characterize the family of matrix metalloproteinases (MMP), where *in silico* characterization was followed by experimental confirmation and MMP-7 was subsequently proved to be a potential target in cardiac hypertrophy **[8, 9].** Collagen and its derivatives, such as gelatin, act as substrates of MMPs and are involved in development of many pathological conditions. An extensive analysis of collagen family is hence essential to comprehend the process of matrix remodeling in diseases.

Our study reports a comparative characterization of alpha-1 sequences of human collagen using biocomputational tools. Of the three alpha chains present in collagen, alpha-1 was observed in all 28 human collagens; thus alpha-1 was used for further analysis of the collagen proteins. The physico-chemical, secondary structural, functional and phylogenetic analysis of alpha-1 sequences of human collagen was executed. This research aims to provide an insight to various protein attributes of collagen proteins and characterize this large family. It also intends to propose potential members implicated in disease conditions based on relative examination of the collagen family or the presence of any atypical characteristics in the collagen molecules. This would aid biologists in carrying out further investigations on these complex molecules, for which, the basic structure analysis is of prime importance.

Methodology:

Retrieval of human collagen alpha-1 protein sequences

The complete alpha-1 protein sequences of all 28 members of human collagen family reported till date were derived from UniProtKB/ SWISS-PROT, a curated protein database (http://expasy.org/sprot/) in FASTA format with the help of the accession number provided for each collagen sequence (http://www.uniprot.org/) **[10].** Complete information about the origin, attributes, annotation, ontologies, binary interactions and sequence of proteins was found in this knowledgebase.

Physico-chemical characterization of collagen family

Various features including number of amino acids, molecular weight, theoretical isoelectric point (pI), amino acid composition (%), number of positively (Arg + Lys) and negatively charged (Asp + Glu) residues, extinction co-efficient, instability index, aliphatic index and Grand Average of Hydropathicity (GRAVY) were computed using ExPASy's ProtParam tool using the protein sequence in FASTA format as the input data type (http://expasy.org/tools/protparam.html). Other physico-chemical features including number of codons, bulkiness, polarity, refractivity, recognition factors, hydrophobicity, transmembrane tendency, percent buried residues, percent accessible residues, average area buried, average flexibility and relative mutability were calculated for primary structure characterization by ExPASy's ProtScale tool using the retrieved protein sequence as the input data type (http://expasy.org/tools/protscale.html). A suitable amino acid scale was chosen for computation of each parameter in this sliding windows based tool that gives each amino acid a numerical value known as the amino acid scale.

Secondary structural characterization of collagen family

The secondary structural features of proteins comprising of alpha helix, 3_{10} helix, Pi helix, beta bridge, extended strand, beta turn, bend region, random coil, ambiguous states and other states were predicted using Self-Optimized Prediction Method with Alignment (SOPMA) tool that takes into account the information derived from alignment of protein sequences belonging to the same family (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html) **[11].** The protein sequence in FASTA format was used as the input data type and the number of conformational states was adjusted to four in order to predict Helix, Sheet, Coil and Turn. The other parameters were set as default.

Functional characterization of collagen family

The Motif scan tool was used to scan and identify all the known motifs, their nature and location in the selected alpha-1 protein sequences of the collagen family based on a profile and pattern search (http://myhits.isb-sib.ch/cgi-bin/motif_scan) **[12, 13].** The protein sequence in FASTA format was used as the input data type and scanned against 'PROSITE Patterns', a selected protein profile database out of the eight available.

Phylogenetic classification of collagen family

The human alpha-1 collagen protein sequences were aligned using multiple sequence alignment tool ClustalW2 using the protein sequences in FASTA format as the input data type (http://www.ebi.ac.uk/Tools/msa/clustalw2/) **[14].** The best alignment for a set of input sequences was computed and all the identities, similarities and differences were highlighted. The evolutionary relationships were established by constructing phylograms through retrieval of the alignments using Neighbor Joining (NJ) method.

Discussion:

Functional role of alpha-1 chain of each human collagen was analyzed **(Table 1, see Supplementary material).** MultAlin tool was used to carry out protein sequence alignment wherein a decreased sequence similarity was witnessed with increasing number of input sequences **[15].** It was inferred that the alpha-1 chain showed identities, similarities and differences at various positions along the protein sequence in all 28 members. The computation of amino acid composition of each human alpha-1 collagen sequence using ExPASY's ProtParam tool indicated very high percentages of glycine and proline as compared to other amino acids **(Table 2, see Supplementary material).** Glycine content in all collagens was more than 12%, except Collagen 12, 14 and 20 with a value of 9.2, 10.9 and 11.5% respectively. High percentage of glycine accounts for the stability of collagen triple helical structure, since incorporation of large amino acids can cause steric hindrance **[16].** Furthermore, proline content was more than 10% in most collagens except Collagen 6, 12, 14 and 20 with values 8.7, 8.6, 8.7 and 10%, respectively. Proline residues are equally essential to point outward and stabilize the helix and also to act as structural disruptor of the secondary structural elements **[17].** Thus, this not only helps collagen to act as a structural molecule

but also aids in processes like cell-cell adhesion, migration. Other essential physico-chemical parameters were also calculated using ExPASy's ProtParam and ProtScale tools **(Table 3a and 3b see Supplementary material)**. The pI values for 15 collagens were found to lie in the acidic range, while for the remaining half, alkaline range was observed. Also, analysis of instability index classified most collagens as stable (instability index <40) while Collagen 15, 17, 18, 20 and 26 were declared as unstable collagens. Additionally, Collagen 20 was regarded most thermostable, with highest aliphatic index (79.61), describing the relative volume of protein occupied by aliphatic side chains, closely followed by Collagen 14 (77.67) and 12 (75.45). Furthermore, the GRAVY values, signifying interaction of collagens with water, were observed within a broad range of -0.261 to -0.919, while the hydrophobicity was found to range between -0.3335 of Collagen20 (most hydrophilic) to 0.3335 of Collagen18 (most hydrophobic).

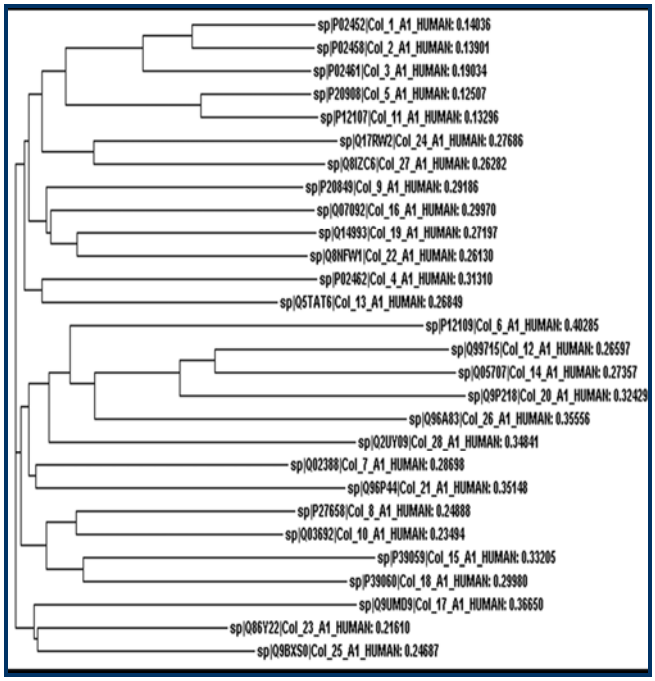

Figure 1: Phylogram of human alpha-1 collagens constructed using Neighbor Joining (NJ) method. Subject sequences were aligned using ClustalW2 multiple alignment tool and phylogenetic trees were constructed thereafter. Clusters with similar relations were identified such as Collagen 12 and 14, with Collagen 20 as their nearest neighbor.

SOPMA analysis of secondary structural features of human collagen protein sequences showed a pre-dominance of random coils while least percentage of β-turns were found. α-helices were found to exceed extended strands in 13 collagens (Collagen 6, 9, 13, 15, 17, 18, 19, 21, 22, 23, 25, 26 and 28) **(Table 4 see Supplementary material).** A high percentage of random coils facilitating self-assembly of the monomer units into well-defined structures can also be linked to high glycine and proline content that are vital to provide the desired flexibility and ability to bond with adjacent units in collagen monomers. This information on packaging of secondary structural elements may assist to derive potential tertiary protein structures and also promote advancements in protein engineering. Motifs of protein can provide significant knowledge about the protein's mechanism of action and nature. Signature motifs were obtained within protein sequences using Motif Scan tool **(Table 5, see supplementary material)**. Collagens 1, 2 and 3 were found to possess the VWFC domain signature, known to participate in oligomerization, hence forming an imperative part of the complex forming proteins **[18, 19].** The presence of this domain can be correlated to the known characteristic of collagen molecules passing through a complex assembly process to form a triple helical structure. Additionally, Collagens 7 and 28 were observed to have pancreatic trypsin inhibitor (Kunitz) family signature. The 28 human alpha-1 collagen protein sequences were aligned based on sequence homology and phylogram was constructed with distance-based NJ method to establish evolutionary relations in the complex collagen family **(Figure 1).** Various clusters with close relationships were identified including Collagen 1 and 2, 12 and 14, 23 and 25 relating closely to Collagen 3, 20 and 17 respectively. These molecules may be analyzed together for possible similar properties owing to their close evolutionary history. Also, Collagen 4, reported majorly in diseases shows similarity to Collagen 13, which may also potentially play a role in pathological conditions. Collagen 9, 12, 14 and 20 are included in the FACIT (Fibril Associated Collagens with Interrupted Triple helices) group of collagen family, where alpha-1 chain of Collagen 9 is recognized for its role in skeletal and rheumatoid disorders, especially in multiple epiphyseal dysplasias **[20].** Owing to their structural similarities, atypical features and proximity in evolutionary history, the other members of this family also present a prime candidature for investigation of their role in skeletal disorders.

Conclusion:

With the availability of a wide variety of computational tools, an in-depth study of the information hidden behind a protein structure is possible. Comparative studies of the members belonging to a protein family and their physico-chemical, secondary structural, functional and phylogenetic classification can help give extensive information of protein's structure, function and its relationship with other members of the family. Characterization of the alpha-1 chain of the vast collagen protein family in humans yielded new insights. Based on this comparative characterization, we hypothesize, Collagen 12, 14 and 20 as a potential protein cluster showing similarity in many properties along with an atypical behavior. These three proteins possess, low glycine and proline, very high aliphatic index and a close evolutionary relation. Since these collagens form a part of the FACIT collagen family, of which collagen 9 is established for its role in skeletal disorders, these collagen molecules might be possible disease candidates. These findings can help biologists working with ECM proteins concentrate their research on collagen proteins proposed as putative players in diseased conditions. Moreover, this study is a model for researchers to fine-tune their specific systems and comprehend their outcomes better.

Acknowledgement:

This work was supported by the research grant awarded to Dr. Vibha Rani by the Department of Science and Technology, Government of India (SR/FT/LS-006/2009: Sept 4, 2009). We acknowledges Jaypee Institute of Information Technology, Deemed to be University, for providing the required support.

References:

[1] Jarvelainen H *et al. PharmacollagenRev*. 2009 **61**:198 [PMID: 19549927]
[2] Bowers SLK *et al. J Mol Cell Cardiol*. 2010 **48**: 474 [PMID: 19729019]
[3] Kadler KE *et al. Biochem J*. 1996 **316**: 1 [PMID: 8645190]
[4] Cen L *et al. Pediat Res*. 2008 **63**: 492 [PMID: 1842729]
[5] Tanzer ML. *J Orthop Sci*. 2006 **11**: 326 [PMID: 16721539]
[6] Bateman JF *et al. Nat Rev Genetics*. 2009 **10**: 173 [PMID: 19204719]
[7] Gonzalez A *et al. Med Clin North Am*. 2004 **88**: 83 [PMID: 14871052]
[8] Jaiswal A *et al. Bioinformation*. 2011 **6**: 23 [PMID: 21464841]
[9] Chhabra A *et al. J Comput Intell Bioinformatics*. 2011 **4**: 1
[10] Bairoch A & Apweiler R. *Nucleic Acids Res*. 1997 **25**: 31[PMID: 9016499]
[11] Geourjon C & Deleage G. *Comput Appl Biosci*. 1995 **11**: 681 [PMID: 8808585]
[12] Pagni M *et al. Nucleic Acids Res*. 2007 **35**: W433- [PMID: 17545200]
[13] Sigrist CJA *et al. Nucleic Acids Res*. 2010 **38**: D161 [PMID: 19858104]
[14] Larkin MA *et al. Bioinformatics*. 2007 **23**: 2947 [PMID: 17846036]
[15] Corpet F. *Nucleic Acids Res*. 1988 **16**: 100881 [PMID: 2849754]
[16] Van der Rest M & Garrone *R. FASEB J*. 1991 **5**: 2814 [PMID: 1916105]
[17] Ramchandran GN *et al. Biochemica et Biophysica Acta*. 1973 **332**: 166 [PMID: 4744330]
[18] Hunt LT & Barker WC. *Biochem Biophys Res Commun*. 1987 **144**: 876 [PMID: 3495268]
[19] Voorberg J *et al. J Cell Biol*. 1991 **113**: 195 [PMID: 2007623]
[20] Czarny-Ratajczak M *et al. Am J Hum Genet*. 2001 **69**: 969 [PMID: 11565064]

Edited by P Kangueane

Citation: **Nassa** *et al*. Bioinformation 8(1): 026-033 (2012)

Supplementary materials:

Table 1: Functional properties of human collagens in diseases:

Collagen	Accession number	Collagen name	Function	Involvement in disease
Col 1	**P02452**	CO1A1_HUMAN Collagen alpha-1(I) chain	Fibrillar forming collagen in bones tendons ligaments	Mutation results in Dwarfism,Ehlers-Danlos syndrome and Osteogenesis imperfecta
Col 2	**P02458**	CO2A1_HUMAN Collagen alpha-1(II) chain	Normal embryonic development of skeleton,linear growth	Cataract,deafness and muation causes stickler syndrome and dwarfism
Col 3	**P02461**	CO3A1_HUMAN Collagen alpha-1(III) chain	Fibrillar forming collagen like type 1 collagen,formation of connective tissue	Muation causes Aortic aneurysm, Ehlers-Danlos syndrome
Col 4	**P02462**	CO4A1_HUMAN Collagen alpha-1(IV) chain	Formation of 'chicken wire' meshwork in glomerulus tissue ,thus involved in filtering of urine in kidney	Alport syndrome
Col 5	**P20908**	CO5A1_HUMAN Collagen alpha-1(V) chain	Fibrillar forming,binds to dna,fibrin,heparin and insulin	Ehlers-Danlos syndrome
Col 6	**P12109**	CO6A1_HUMAN Collagen alpha-1(VI) chain	Cell binding protein	Ulrich myopathy and Bethlem myopathy
Col 7	**Q02388**	CO7A1_HUMAN Collagen alpha-1(VII) chain	Basement membrane organization and adherence	epidermolysis bullosa dystrophica
Col 8	**P27658**	CO8A1_HUMAN Collagen alpha-1(VIII) chain	Proliferation of vscular smooth muscle cells,maintains vessel wall integrity and structure	Posterior polymorphous corneal dystrophy 2
Col 9	**P20849**	CO9A1_HUMAN Collagen alpha-1(IX) chain	Flexible hence connects type II collagen to other cartilage components	Osteoarthritis,muatation may result in epiphyseal dysplasia and stickler syndrome
Col 10	**Q03692**	COAA1_HUMAN Collagen alpha-1(X) chain	-	Schmid type metaphyseal chondrodysplasia
Col 11	**P12107**	COBA1_HUMAN Collagen alpha-1(XI) chain	Important role in fibrillogenesis by controlling growth of collagen II fibrils	Marshall syndrome and stickler syndrome
Col 12	**Q99715**	COCA1_HUMAN Collagen alpha-1(XII) chain	Modifies surrounding matrix by interaction with type I collagen	N.D
Col 13	**Q5TAT6**	CODA1_HUMAN Collagen alpha-1(XIII) chain	Cell-matrix and cell-cell interactions for normal development	-
Col 14	**Q05707**	COEA1_HUMAN Collagen alpha-1(XIV) chain	Interacts with collagen bundles,adhesion	-
Col 15	**P39059**	COFA1_HUMAN Collagen alpha-1(XV) chain	Stabilizes microvessels and muscle cells in heart and skeletal muscle,inhibits angiogenesis	-
Col 16	**Q07092**	COGA1_HUMAN Collagen alpha-1(XVI) chain	Induces integrin mediated interactions:Cell spreading, attachment and alteration of morphlogy of cells	-
Col 17	**Q9UMD9**	COHA1_HUMAN Collagen alpha-1(XVII)	Role in maintaining integrity of hemidesmosome, attachment of basal keratinocytes basement membrane.	-
Col 18	**P39060**	COIA1_HUMAN Collagen alpha-1(XVIII) chain	Determination of renal structure and closure of renal tube,inhibits angiogenesis	-

Col 19	**Q14993**	COJA1_HUMAN Collagen alpha-1(XIX) chain	Esophagus development,organization of pericellular matrix	-
Col 20	**Q9P218**	COKA1_HUMAN Collagen alpha-1(XX) chain	-	
Col 21	**Q96P44**	COLA1_HUMAN Collagen alpha-1(XXI) chain	-	-
Col 22	**Q8NFW1**	COMA1_HUMAN Collagen alpha-1(XXII) chain	Acts as a cell adhesion ligand for skin epithelial cells and fibroblasts	-
Col 23	**Q86Y22**	CONA1_HUMAN Collagen alpha-1(XXIII) chain	-	-
Col 24	**Q17RW2**	COOA1_HUMAN Collagen alpha-1(XXIV) chain	Regulation of type I collagen fibrillogenesis at specific anatomical locations during fetal development	-
Col 25	**Q9BXS0**	COPA1_HUMAN Collagen alpha-1(XXV) chain	Binds heparin,assembles amyloid fibrils into protease resisitant aggregates	-
Col 26	**Q96A83**	EMID2_HUMAN Collagen alpha-1(XXVI) chain	-	-
Col 27	**Q8IZC6**	CORA1_HUMAN Collagen alpha-1(XXVII) chain	calcification of cartilage and the transition of cartilage to bone	-
Col 28	**Q2UY09**	COSA1_HUMAN Collagen alpha-1(XXVIII) chain	May act as a cell binding protein	-

*N.D. - Not defined

Table 2: Amino acid composition of human alpha-1 collagens (in %) :

Collagen	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Col 1	9.5	4.8	1.2	4.5	1.2	3.3	5.1	26.7	0.6	1.6	3.3	3.9	0.9	1.8	19	4.1	3.1	0.4	0.9	3.2
Col 2	9	4.8	2.2	4.2	1.3	4	5.3	27.3	0.5	2.3	3.8	4.5	1.1	1.7	18.2	3.2	3.0	0.5	0.7	2.6
Col 3	7.8	4.1	2.8	3.8	1.5	2.9	5.0	28.2	1.0	2.5	3.3	4.2	1.2	1.6	19.2	5.0	2.1	0.5	0.1	2.5
Col 4	3.5	2.7	1.0	3.5	1.2	4.4	4.2	28.6	1.0	3.5	5.5	5.6	1.9	2.8	19.4	4.3	2.6	0.4	1.1	3.1
Col 5	5	3.8	1.8	5.7	0.7	4.0	6.5	23.3	0.9	2.8	5.3	5.3	1.3	2.1	18.2	3.8	3.8	0.4	2.2	2.9
Col 6	7.5	5.8	2.7	6.9	1.9	4.1	6.6	15.2	1.4	4.1	6.9	5.3	1.1	3.2	8.7	5.3	4	0.4	2.7	6.3
Col 7	6.6	7.4	0.9	4.9	0.6	3.6	6.4	21.3	1.0	1.9	7.2	3.2	0.6	1.3	14.5	5.6	4.5	0.6	1.2	6.8
Col 8	4.4	1.7	1.1	1.5	0.3	5.1	3.5	25.9	1.5	4.4	7.1	6.3	3.1	2.3	22.4	1.5	1.1	0.1	3.0	4.0
Col 9	5.1	5.9	2.2	4.2	1.2	4.2	5.1	23.6	0.9	3.9	6.2	4.6	1.3	2.4	16.6	4.6	2.8	0.7	0.7	4.0
Col 10	5.3	2.8	2.1	1.8	0.1	3.4	3.2	25.7	1.5	4	5.1	5.1	1.6	2.2	21.3	4	3.5	0.3	3.1	3.8
Col 11	5.4	3.8	2.0	5.5	0.6	4.5	6.8	23.4	0.9	3.1	4.4	5.9	1.3	2.6	15.9	4	3.8	0.5	2.0	3.5
Col 12	5.2	5.1	3.9	5.6	0.7	3.4	6.3	9.2	1.0	4.8	6.6	5.1	1.5	3.0	8.6	7.9	8.7	1.0	3.7	8.8
Col 13	6.8	4.9	1.5	3.2	1.1	3.9	6.1	25.7	1.5	2.6	7.4	6.4	1.8	0.7	17.4	3.3	2.4	0.3	0.4	2.4
Col 14	5.2	4.1	3.2	5.2	1.1	4.0	6.6	10.9	1.6	5.5	7.2	4.8	1.8	3.3	8.7	7.4	7.6	1.0	3.0	7.9
Col 15	7.8	3.4	2.7	4.2	0.7	2.9	7.0	15.9	1.8	3.9	8.1	3.7	2.2	3.1	13.8	7.1	5.3	0.7	1.0	4.7
Col 16	5.3	3.9	1.4	3.2	2	4.9	5.7	24.4	1.2	2.6	7.1	5.4	1.6	1.9	17.6	4.7	2.7	0.5	0.8	3.9
Col 17	5.7	4.5	2.1	3.6	0.5	3.2	4.2	18.8	1.5	2.7	8.4	4	2.2	1.7	13.6	11.9	5.6	0.7	2.4	3.9
Col 18	8	5.2	1.5	4.2	1.3	3.8	5.1	17.2	2.1	1.7	6.2	2.4	0.9	3.2	16.9	6.9	3.9	1.2	1.1	5.0
Col 19	4.6	4.1	2.6	4.3	1.2	4	5.9	22.9	1.4	5	10.7	6.7	1.6	2.2	15.1	4.6	2.6	0.6	1.4	2.9
Col 20	8.9	6.6	0.9	3.6	1	4.4	5.7	11.5	2.2	1.8	6.9	3	1.1	2.7	10	8.3	6.7	1.3	2.0	7.6
Col 21	4.2	4	2.7	5	1.5	5.3	5.2	18.9	1.1	5.6	6.2	7.1	1.3	3.0	11.1	5.2	3.6	0.4	2.3	5.5
Col 22	5.8	5.2	1.5	4.3	1.1	3.8	6.4	24.4	1.2	2.8	7.4	5.4	1.2	2.1	16.4	4.1	2.6	0.4	1.0	4.3
Col 23	8	3.9	0.4	5.4	1.1	3.1	6.7	27.2	0.6	1.5	7.5	6.5	0.9	0.4	17.2	3.1	1.5	0.4	0.4	2.8
Col 24	3.4	4.3	3.3	3.5	0.9	5.3	6	21.3	2.2	5.1	6.3	6.1	1.4	2.5	12.1	5.3	4.6	0.3	1.8	4.0
Col 25	5	6.3	1.4	4.3	1.1	4.9	6.9	25.5	1.5	3.1	8.8	7.6	2.6	0.8	16.5	3.1	2.4	0.2	0.6	2.0
Col 26	8.6	5.4	2.3	3.6	2.9	4.1	5.4	14.7	1.6	1.6	7.2	2.7	1.4	0.9	15.4	6.6	5.2	1.1	1.8	4.8
Col 27	5.9	3.6	1.1	3.7	0.8	5.2	3.6	21.3	1.7	2.4	5.9	5.1	2	2.6	16.3	5.7	4.7	0.5	0.9	3.8
Col 28	3.7	3.6	2.4	5.6	1.3	5.2	6.5	18.8	0.6	5.4	5.9	8.1	1.2	3.4	11.5	6.1	4	0.4	1.6	4.7

*Col– Collagens

Table 3(a): Physico-chemical parameters of human alpha-1 collagens:

Collagen	No. of amino acids	Molecular weight	pI	'-' charged residue	'+' charged residue	Extinction Coefficient	Instability index	Aliphatic index	GRAVY
Col 1	1464	138941.5	5.6	141	128	53495	30.43	37.98	-0.788
Col 2	1487	141785.3	6.58	141	139	54525	25.21	40.03	-0.803
Col 3	1466	138564.2	6.21	129	122	62225	30.18	37.31	-0.797
Col 4	1669	1606147.7	8.55	128	138	61070	32.04	47.39	-0.621
Col 5	1838	183559.8	4.94	225	168	98850	33.09	45.35	-0.873
Col 6	1028	108529.4	5.26	139	114	64970	28.52	68.7	-0.525
Col 7	2944	295219.6	5.95	332	310	159140	32.07	61.86	-0.625
Col 8	744	73364	9.62	37	60	38405	36.06	61.21	-0.434
Col 9	921	91869.2	8.94	86	96	42565	32.61	56.13	-0.658
Col 10	680	66157.9	9.68	34	54	42290	25.95	51.94	-0.556
Col 11	1806	181064.8	5.06	222	174	103765	30.81	44.91	-0.859
Col 12	3063	333146.7	5.38	366	313	334620	32.90	75.45	-0.427
Col 13	717	69949.9	9.27	67	81	15970	31.44	52.87	-0.765
Col 14	1796	193515.4	5.16	211	160	179095	37.57	77.67	-0.326
Col 15	1388	141720.1	4.9	155	95	76485	40.19	68	-0.377
Col 16	1604	157751.3	8.14	144	150	65370	35.88	50.73	-0.671
Col 17	1497	150419.3	8.89	117	128	109015	45.25	55.47	-0.573
Col 18	1754	178187.6	5.67	164	133	145185	48.57	61.72	-0.467
Col 19	1142	115220.7	8.57	116	124	63215	30.68	56.68	-0.708
Col 20	1284	135830	8.27	119	123	132990	45.18	79.61	-0.261
Col 21	957	99368.5	8.57	98	106	55655	33.28	69.14	-0.517
Col 22	1626	161145.3	6.88	174	172	57965	34	53.28	-0.715
Col 23	540	51943.9	6.88	65	65	14355	30.81	50.69	-0.829
Col 24	1714	175496.3	8.46	162	170	73075	28.32	64.21	-0.622
Col 25	654	64770.7	8.6	73	78	11835	24.85	47.19	-0.919
Col 26	441	45381.1	7.02	40	40	40170	46.77	63.11	-0.523
Col 27	1860	186892.3	9.83	136	196	81205	37.62	54.15	-0.637
Col 28	1125	116657.1	6.1	136	131	55195	24.18	61.42	-0.66

*Col- Collagens

Table 3(b): Physico-chemical properties of human alpha-1 collagens

Collagen	No. of codons	Bulkiness	Polarity	Refractivity	Recognition factors	Hydrophobicity	Transmembrane tendency	% buried residues	% accessible residues	Average area buried	Average flexibility
Col1	3.889	14.3645	17.2635	14.4125	87.5	-0.2835	-0.6295	6.528	5.622	118.5835	0.4505
Col2	3.611	14.535	17.012	14.417	89.722	0.05	-0.434	6.9445	5.9665	116.9055	0.452
Col3	3.889	13.688	14.3925	14.1385	88.889	-0.222	-0.3985	7.3885	5.65	116.239	0.4505
Col4	3.722	13.466	16.9645	14.739	89.6115	0.0115	-0.564	6.561	5.9555	120.7055	0.4445
Col5	3.889	13.905	16.8185	15.437	88.5555	-0.089	-0.429	6.1335	5.9	117.989	0.452
Col6	3.5555	13.866	17.659	16.544	88.7225	-0.0225	-0.5395	5.95	5.672	116.35	0.453
Col7	3.7775	14.1405	17.502	14.787	90.611	-0.139	-0.6305	6.767	5.9445	118.028	0.4485
Col8	3.722	13.8405	14.092	15.435	88.389	-0.1835	-0.3875	6.5835	5.8725	119.7725	0.4455
Col9	3.5555	14.563	17.537	16.071	87.5555	-0.0055	-0.3405	6	5.422	122.5835	0.444
Col10	3.722	14.577	12.346	14.2365	88.5555	-0.2665	-0.3025	6.25	5.661	119.7725	0.4555
Col11	3.4445	13.735	17.296	16.398	89	-0.55	-0.6195	6.611	6.0665	122.972	0.45
Col12	3.722	14.226	16.8455	15.1145	90	-0.078	-0.5105	6.9225	5.9335	121.2225	0.454
Col13	3.833	13.779	20.1465	15.135	80.0555	-0.2335	-0.55	6.4615	5.9835	118.511	0.4495
Col14	3.5555	14.611	19.96	15.7995	89.3335	-0.283	-0.843	5.622	5.7055	122.35	0.4495
Col15	3.72205	13.5155	16.584	15.1545	87.444	-0.033	-0.513	6.861	5.4945	119.6555	0.443
Col16	3.5	14.316	17.226	15.108	88.2775	0.1775	-0.357	6.1055	5.789	120.672	0.448
Col17	3.889	12.439	17.232	15.038	92.333	-0.1165	-0.3895	7.1615	6.122	116.522	0.454
Col18	3.889	13.5065	17.062	14.587	89.444	0.3335	-0.3765	6.3385	5.5275	122.828	0.4335
Col19	3.5	14.179	17.4405	16.2065	87.4445	-0.622	-0.4485	6.311	5.7165	126.9225	0.4445
Col20	3.7775	14.161	19.7095	15.07	90.6665	-0.3335	-0.5825	6.828	5.9115	124.783	0.4425
Col21	3.5555	14.8025	16.9925	15.4085	88.0555	-0.3335	-0.3235	6.256	5.811	123.1165	0.439
Col22	3.722	14.055	17.253	15.4765	87.056	0.222	-0.32	6.4555	5.911	124.278	0.4435
Col23	3.9445	11.8145	19.881	11.6245	87.222	-0.017	-0.5535	5.63385	6	109.889	0.4535
Col24	3.722	14.5095	17.514	15.4565	89.1115	-0.2555	-0.472	7.1	5.8225	119.835	0.4395
Col25	3.7225	13.5035	17.207	13.0145	87.3885	0.011	-0.534	7.289	5.6555	113.683	0.4545
Col26	3.167	13.7255	14.878	15.4075	87.8885	-0.3055	-0.593	6.767	5.2225	116.1885	0.4455
Col27	3.611	13.9525	17.331	14.9185	90.889	-0.111	-0.3675	6.55	6.011	119.067	0.445
Col28	3.5555	14.7145	19.934	16.194	88.3335	-0.183	-0.5865	6.022	5.7165	127.9725	0.446

*Col- Collagens

Table 4: Secondary structural features of human alpha-1 collagens (in %):

Collagen	α helix	3_{10} helix	Pi helix	β bridge	Extended strand	β turn	Bend region	Random coil	Ambiguous states	Other states
Col1	4.37	0	0	0	8.13	4.3	0	83.2	0	0
Col2	6.27	0	0	0	7.2	4.57	0	81.51	0	0
Col3	4.5	0	0	0	7.44	4.71	0	83.36	0	0
Col4	4.73	0	0	0	7.61	5.69	0	81.97	0	0
Col5	7.94	0	0	0	11.53	4.9	0	75.63	0	0
Col6	24.22	0	0	0	15.86	5.25	0	54.67	0	0
Col7	9.51	0	0	0	15.9	6.15	0	68.44	0	0
Col8	5.65	0	0	0	10.89	9.41	0	74.06	0	0
Col9	9.45	0	0	0	7.17	3.37	0	80.02	0	0
Col10	3.97	0	0	0	12.35	5	0	78.68	0	0
Col11	8.36	0	0	0	12.02	5.59	0	74.03	0	0
Col12	13.12	0	0	0	25.56	4.9	0	56.42	0	0
Col13	14.09	0	0	0	4.6	4.46	0	76.85	0	0
Col14	17.76	0	0	0	24.44	5.57	0	52.23	0	0
Col15	18.44	0	0	0	16.71	6.77	0	58.07	0	0
Col16	7.79	0	0	0	10.1	5.99	0	76.12	0	0
Col17	19.91	0	0	0	11.89	6.95	0	61.26	0	0
Col18	17.56	0	0	0	12.43	6.9	0	63.11	0	0
Col19	13.57	0	0	0	9.54	6.04	0	70.84	0	0
Col20	18.15	0	0	0	21.65	5.69	0	54.52	0	0
Col21	15.26	0	0	0	12.43	4.7	0	67.61	0	0
Col22	11.38	0	0	0	10.7	7.56	0	70.36	0	0
Col23	13.7	0	0	0	0.56	1.3	0	84.44	0	0
Col24	7.41	0	0	0	9.33	2.92	0	80.34	0	0
Col25	12.69	0	0	0	4.43	3.82	0	79.05	0	0
Col26	24.94	0	0	0	11.11	5.9	0	58.05	0	0
Col27	9.46	0	0	0	11.88	5.86	0	72.8	0	0
Col28	19.29	0	0	0	10.58	4.18	0	65.96	0	0

*Col- Collagen

Table 5: Motifs in human alpha-1 collagens**:**

Collagen	Motif found	Motif ID	Description	Start	End	Match Status Significance
Col1	VWFC_1	**PS01208**	VWFC domain signature	58	95	Strong match; not a false positive
Col2	VWFC_1	**PS01208**	VWFC domain signature	52	89	Strong match; not a false positive
Col3	VWFC_1	**PS01208**	VWFC domain signature	50	88	Strong match; not a false positive
Col7	BPTI_KUNITZ_1	**PS00280**	Pancreatic trypsin inhibitor (Kunitz) family signature	2907	2925	Strong match; not a false positive
Col28	BPTI_KUNITZ_1	**PS00280**	Pancreatic trypsin inhibitor (Kunitz) family signature	1100	1118	Strong match; not a false positive

*VWFC: von Willebrand factor (VWF) type C repeat
**No motifs were found in remaining human alpha-1 collagen sequences
*** Col- Collagen