

TRACKING EVOLUTION OF COLLAGEN SUBFAMILY FOR ITS UTMOST POSSIBLE BENEFIT

Esmail I. F. SAAD^{1,2} and Tahani A. A. Almabruk^{3,4}

¹Department of Microbiology, Omar almukhtar university, Al Bayda Libya;

²Department of life science, Libayn academy Aljabal Alkhader branch, Al Bayda Libya

³Department of Computer Science, Omar almukhtar university, Al Bayda Libya;

⁴Department of Computer Science, Libayn academy Aljabal Alkhader branch, Al Bayda Libya

DOI: <https://doi.org/10.58309/aajpas.v1i1.11>

KEYWORDS:

Collagen,
evolution,
evolutionary tree,
selection pressure

ABSTRACT:

The Collagen family is one of the most extensively researched protein families. However, because an evolutionary study has focused mostly on superfamilies, there is a paucity of information about subfamilies. To promote a healthy circulatory system, healthy skin through enhancing elasticity, and a healthy joint system, it is crucial that the subfamily of collagen evolve. Additionally, these kinds of studies support bone health, development against various developmental problems, and placenta growth in expectant women. The current research sought to comprehend collagen evolution on both a comprehensive and reductive level. The full family of human collagen proteins was exposed to BLAST techniques to identify orthologs and paralogs with various criteria, followed by the creation of phylogenetic trees. Our findings unambiguously demonstrated that the most archaic collagen isoforms in higher eukaryotes are COL20A1HP and COL12A1HPa, which later diverged into other collagen family members, possibly as a result of various gene modification processes. Additionally, a site-specific selection pressure study utilizing the selection server revealed that COL4A isoforms are evolving more quickly than other types of collagen.

تتبع تطور عائلة الكولاجين لاستغلال فوائدها

إسماعيل إبراهيم فضيل سعد^{1,2}، تهاني المبروك إكرام المبروك^{3,4}

¹ قسم الأحياء الدقيقة، جامعة عمر المختار، البيضاء - ليبيا

² قسم علوم الحياة، الأكاديمية الليبية الجبل الأخضر، البيضاء - ليبيا

³ قسم الحاسوب، جامعة عمر المختار، البيضاء - ليبيا

⁴ قسم الحاسوب، الأكاديمية الليبية الجبل الأخضر، البيضاء - ليبيا

الكلمات المفتاحية:

الكولاجين،
التطور،
شجرة التطور،
الانتخاب الضاغط.

المستخلص:

تعد عائلة الكولاجين واحدة من أكثر عائلات البروتين التي تمت دراستها على نطاق واسع. ومع ذلك، فقد تم إجراء التحليل التطوري بشكل أساسي على مستوى العائلة العليا الفائقة، مما جعل تفاصيل الأسر الفرعية شحيحة للغاية. وبالتالي، فإن تطور الفصائل الفرعية على التوسع التطوري للكولاجين هو جدير بالدراسة، لما له من فوائد كثيرة منها استكشاف الطريقة المثلى لاستخدامه خاصة في الصناعات دوائية كدعم نظام القلب والأوعية الدموية وتقرحات الجلد. بالإضافة إلى ذلك هذا النوع من الدراسة يمكن أن ينتج عنه فهم وتحديد أفضل أنواع الكولاجين لتحسين المرونة وصحة المفاصل وتعزيز نمو المشيمة عند النساء الحوامل كما أنه يساعد في تعزيز صحة العظام وتطورها ضد أي اضطرابات تنموية أخرى. تهدف الدراسة الحالية إلى فهم تطور الكولاجين على المستويين الشامل وترتيبه الاختزالي. بعد جمع سلال العائلة تعرضت جميع عائلات بروتين الكولاجين البشري لأدوات Blast للعثور على أوجه التشابه بين السلاسل وذلك بمعايير مختلفة متنوعة بتوليد شجرة النشوء والتطور. أظهرت نتائجنا بوضوح أن الأشكال COL20A1HP و COL12A1HPa هي الكولاجين الأكثر بدائية في حقيقيات النوى الأعلى ثم تباعدت إلى أفراد عائلة Collagens الآخرين قد يكون بسبب أحداث تعديل الجينات المختلفة. علاوة على ذلك، تم إجراء تحليل ضغط اختيار خاص بالموقع باستخدام خادم SELECTON، والذي أظهر أن الأشكال COL4A تتطور بشكل أسرع من أنواع الكولاجين الأخرى.

INTRODUCTION

The primary structural protein in the bodies of vertebrates is known as collagen. Collagen is the extracellular protein of the essential connective tissues, and compensates more than 90% of tendons, bones, and skin (Meena, et al., 1999). There are 29 members of the collagen family, and each of them contains at least one triple helix domain. The majority of the collagen that is deposited forms supra-molecular assemblies in the extracellular matrix. There are four types of collagen membrane proteins that are exported from cell surfaces as well as being present in a dissolved form. The mechanical characteristics, structure, and shape of tissues are all influenced by collagen, which acts as a structural component. It interacts with cells from many receptor families and controls their division, migration, and differentiation. Certain collagens have unique biological roles and restricted tissue distribution (Ricard-Blum, 2011). Due to its unique characteristics of low toxicity, good biological properties, and low immunogenicity, collagen is among the most studied proteins and has a wide range of uses in the pharmaceutical, biomedical, cosmetic, leather, and film industries. Despite the fact that collagen is abundant, outbreaks of various diseases in wild animals place our daily use of it in risk. The research for an alternative source, thus began, resulting in the discovery of a vast amount of untapped marine resources, including fish, jellyfish, and various marine mammals (Felician, et al., 2018). In vertebrates, there are 29 different kinds of collagen, each of which has at least 46 distinct polypeptide chains. A right-handed triple helix is created when a left-handed polypropylene II (PPII) helix wraps three parallel strands of the polypeptide around one another. A repeat of the sequence Xaa Yaa Gly, where Xaa and Yaa can be any amino acid, results from the requirement that every third residue in PPII helices be a Gly. All types of collagen have this repetition, although the non-fibrous collagen's triple helix domain has multiple places where it is broken. In the Xaa and Yaa modes of collagen, the amino acids (2S)-proline (Pro, 28%) and (2S,4R)-4-hydroxyproline (HyP, 38%) are frequently

present. The most common triplet in collagen is ProHypGly (10.5%) (Schweitzer et al., 2007; Pevzner et al., 2008 Buckley et al., 2020; Tang et al., 2022). In previous studies have not specifically attempted to analyze the role of Collagen in the hazard of changes phenotype, rather than in its "complication" phenotype. Therefore, This article includes a brief description of the collagen family and we simultaneously evaluated the evolutionary relation and structural analysis of Collagen and their reconstructed haplotypes as possible risk determination using *in-silicon* tools.

MATERIALS AND METHODS

Through creating Boolean queries against the NCBI's Gene database (<http://www.ncbi.nlm.nih.gov/gene/>), the protein architecture (forms and isoforms) of each member of the Collagen family, COL1A to COL 29A, was tracked in Homo sapiens. Furthermore, the databases Genbank (<http://www.ncbi.nlm.nih.gov/nucleotide/>) and genpept (<http://www.ncbi.nlm.nih.gov/protein/>) were used to retrieve the gene sequences, coding sequences (CDS), and protein sequences in Fasta format (Saad, 2017 and Saad and Attitalla 2017). For each member of the Collagen family, a homolog search was done. To obtain just the real positive findings, a rigid parameter was set. The NCBI BLAST tools blastn and blastp were used to conduct the search against a Non-Redundant Database (Altschul et al., 1990). All subject sequences that were artificial constructs and sequences that aligned to queries with e-values greater than zero, identities below 100%, and coverage below 90% were eliminated. Our aim was to determine the extent of evolution in individual genomes as well as the distribution of the Collagen family. To create trees and perform clustering, Topali software (Milne et al., 2009) was utilized. At the protein level, phylogenetic trees were produced by neighbor-joining with a bootstrapping iteration value of 100. I-TASSER server's ab-initio methodology was used to estimate the 3D structure of the chosen isoform of each member of the Human Collagen family in order to take use of protein structures' potential for tracing evolutionary history (Zhang

2008). Due to its high degree of differing from other members of the Collagen subfamily. The structural models were then evaluated using the PROCHECK tool from the Structural Analysis and Verification Server, which uses a Ramachandran plot to evaluate the stereochemical quality of a protein structure (Laskowski et al., 1996). For further analysis, the top models from each unique isoform were considered. The protein sequence and the modeled forms were submitted to the SELECTON server, which calculated the substitution rate ratios of nonsynonymous (K_a) versus synonymous (K_s) mutations, in order to determine whether a positive or negative selection of particular amino acid sites within the full-length sequences of Collagen individuals chosen at random. A K_a/K_s ratio above 1 denotes positive selection, while one below 1 denotes purifying selection (Stem et al., 2007).

RESULTS AND DISCUSSION

The goal of this study was to better understand the Collagen family's possible evolutionary origins. Using Boolean queries like AND and OR, the NCBI gene database was searched for all human Collagen subfamily forms and isoforms Table 1 (see supplementary material). According to Ricard-Blum 2011 and Exposito et al., 2002 research's, the expansion of the Collagen family in complex higher organisms were clearly seen. The rigorous constraints, however, also eliminated misleading negative results, i.e., the Collagen eukaryotic forms significantly diverged lower species and showed no significant resemblance on the protein level for the entire stretch of sequence due to great evolutionary distance and time. Therefore, the BLAST all-to-all technique has been applied to all members of the Collagen family from lower eukaryotes and prokaryotes. Surprisingly, no substantial similarity for the entire length of sequences was observed among lower eukaryotes and prokaryotes, indicating a great evolutionary distance and numerous parallel evolutionary processes among lower organisms. Protein analysis can therefore be used to determine how prokaryotes became eukaryotes. Therefore, protein sequences were used to create phylogenetic trees using a

neighbor-joining method, followed by clustering, in order to determine the evolutionary route from all types of Collagen family (Figure 1).

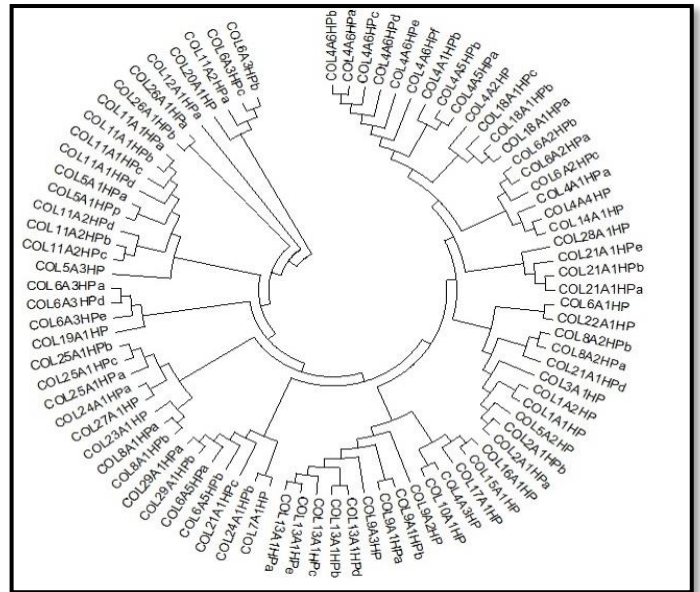


Figure: (1). Phylogenetic tree of Human Collagen on Protein basis

According to the protein tree, it was COL20A1HP and COL12A1HPa that evolved as primitive Collagen and then split into the several Collagen forms as a result of events like gene duplication during evolution. The research clearly demonstrated that the COL20A1HP and COL12A1HPa subfamily undergone further evolution and split into rest of collagen type in eighteen clusters. However, COL4A and its isoforms were found in the recent clusters, i.e. 16th, 17th and 18th cluster / subcluster. SELECTON server was employed to evaluate the various selection pressures on all the sites of the randomly selected type of Collagen. A site-specific selection pressure analysis was performed using the SELECTON server Table 3 (see supplementary material) on all selected protein structures of the longest Collagen forms were modeled by the I-TASSER server using the ab-initio methodology Table 2 (see supplementary material). The study of selection pressures revealed that the majority of the codons in COL4A were under purifying selection, but COL20A1HP and COL12A1HPa showed a neutral selection when taking into account the distribution of the computed K_a/K_s ratio (ω).

CONCLUSION

Collagen, is an essential protein of connective tissues, The study of evolution of this family has immense significance in understanding of multicellularity evolution. From our finding, we speculate and hypothesize that among all the different Collagen sequences, COL20A1HP and COL12A1HPa are found to be the most primitive Collagen types holding the first two clusters that subsequently diverged into other two new Collagen subfamilies. The Collagen IV isoforms are the most recent Collagen sequences at clusters 15th, 16th, 17th, and 18th which showed sudden evolutionary expansion throughout all Collagene .

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990).** Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Buckley M, Harvey V, Orihuela, J, Mychajliw A, Keating J, Milan J, Lawless C, Chamberlain A, Egerton V, Manning P, (2020).** Collagen Sequence Analysis Reveals Evolutionary History of Extinct West Indies *Nesophontes* (Island-Shrews), *Molecular Biology and Evolution*, 37(10), PP 2931–2943, <https://doi.org/10.1093/molbev/msaa137>
- Exposito, J.-Y., Cluzel, C., Garrone, R. and Lethias, C. (2002),** Evolution of collagens. *Anat. Rec.*, 268: 302-316. <https://doi.org/10.1002/ar.10162>
- Felician, Fatuma Felix, Xia, Chunlei, Qi, Weiyan, & Xu, Hanmei. (2018).** Collagen from marine biological sources and medical applications. *Chemistry & biodiversity*, 15(5), e1700557.
- Laskowski, Roman A., Rullmann, J. Antoon C., MacArthur, Malcolm W., Kaptein, Robert, & Thornton, Janet M. (1996).** AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, 8(4), 477-486. doi: 10.1007/BF00228148
- Meena, C, Mengi, SA, & Deshpande, SG. (1999).** *Biomedical and industrial applications of collagen*. Paper presented at the Proceedings of the Indian Academy of Sciences-Chemical Sciences.
- Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F., & Wright, F. (2009).** TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics (Oxford, England)*, 25(1), 126–127. <https://doi.org/10.1093/bioinformatics/btn575>
- Pevzner PA, Kim S, Ng J.(2008).** Comment on “Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry.”. *Science*; 321:1040. [PubMed: 18719266]
- Ricard-Blum S. (2011).** The collagen family. *Cold Spring Harbor perspectives in biology*, 3(1), a004978. <https://doi.org/10.1101/cshperspect.a004978>
- Saad, E. I. (2017).** Drug Design for Cancer-Causing PI3K (P110 α) subunit Mutant Protein. *International Journal of Pharmacy & Life Sciences*, 8(11).
- Saad, E. I. and Attitalla I H. (2017)** Molecular Dynamic Simulation and an Inhibitor Prediction of PI3K (P110 α) Subunit Mutant Protein Structured Model as a Potential Drug Target for Cancer. Vol. 1 No. 1: 7.
- Schweitzer, M. H., Suo, Z., Avci, R., Asara, J. M., Allen, M. A., Arce, F. T., & Horner, J. R. (2007).** Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. *Science (New York, N.Y.)*, 316(5822), 277–280. <https://doi.org/10.1126/science.1138709>
- Stern, A., Doron-Faigenboim, A., Erez, E.,**

Martz, E., Bacharach, E., & Pupko, T. (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic acids research*, 35(Web Server issue), W506–W511. <https://doi.org/10.1093/nar/gkm382>

Tang C., Zhou K., Zhu Y., Zhang W., Xie Z., Wang Z., Zhou H., Yang T., Zhang Q., XuB. (2022). Collagen and its derivatives: From structure and properties to their applications in food industry. *Food Hydrocolloids*, 131, PP.107748. doi: 10.1016/j.foodhyd.2022.107748

Zhang Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, 9, 40. <https://doi.org/10.1186/1471-2105-9-40>

Supplementary material:

Table:(1). The gene-protein architecture (forms and isoforms) of all Collagen subfamily members in *Homosapiens* by generating Boolean queries against NCBI's Gene database

Type of Collagen	Gene Number	Isoforms	Accession number of Gene	Accession number of Protein	Chromosome locations
I	COL1A1	-	NM_000088	NP_000079	Chr 17: 50.18 – 50.2 Mb
	COL1A2	-	NM_000089	NP_000080	Chr 7: 94.39 – 94.43 MB
II	COL2A1	a	NM_001844	NP_001835	Chr 12: 47.97 – 48 MB
		b	NM_033150	NP_149162	Chr 12: 47.97 – 48 MB
III	COL3A1	a	NM_000090	NP_000081	Chr 2: 188.97 – 189.01 Mb
		b	NM_001376916	NP_000081	Chr 2: 188.97 – 189.01 Mb
IV	COL4A1	a	NM_001845	NP_001290039	Chr 13: 110.15 – 110.31 Mb
		b	NM_001303110	NP_001836	
	COL4A2	-	NM_001846	NP_001837	Chr 13: 110.31 – 110.51 Mb
		a	NM_000091	NP_000082	
		b	NM_031362	NP_000082	
		c	NM_031363	NP_000082	Chr 2: 227.16 – 227.31 Mb
		d	NM_031364	NP_000082	
		e	NM_031365	NP_000082	
	f	--	NP_000082		
	COL4A4	-	NM_000092	NP_000083	Chr 2: 227 – 227.16 Mb
	COL4A5	a	NM_000495	NP_000486	
		b	NM_033380	NP_203699	Chr X: 108.44 – 108.7 Mb
COL4A6	c	NM_033381			
	a	NM_001287758	NP_001274687	Chr X: 108.16 – 108.44 Mb	
	b	NM_001287759	NP_001274688	Chr X: 108.16 – 108.44 Mb	
	c	NM_001287760	NP_001274689	Chr X: 108.16 – 108.44 Mb	
	d	NM_001847	NP_001838	Chr X: 108.16 – 108.44 Mb	
	e	NM_033641	NP_378667	Chr X: 108.16 – 108.44 Mb	
COL5A1	a	NM_000093	NP_000084	a/n	
	b	NM_001278074	NP_001265003	Chr 9: 134.64 – 134.84 Mb	
V	COL5A2	-	NP_000384		
		-	NM_000393	NP_031763	n/a
	COL5A3	-	NM_015719	NP_056534	Chr 19: 9.96 – 10.01 Mb
VI	COL6A1	-	NM_001848	NP_001839	Chr 21: 45.98 – 46.01 Mb

		a	NM_001849	NP_001840	Chr 21: 46.1 – 46.13 Mb
	COL6A2	b	NM_058174	NP_478054	Chr 21: 46.1 – 46.13 Mb
		c	NM_058175	NP_478055	Chr 21: 46.1 – 46.13 Mb
		a	NM_004369	NP_004360	Chr 2: 237.32 – 237.41 Mb
		b	NM_057164	NP_476505	Chr 2: 237.32 – 237.41 Mb
	COL6A3	c	NM_057165	NP_476506	Chr 2: 237.32 – 237.41 Mb
		d	NM_057166	NP_476507	Chr 2: 237.32 – 237.41 Mb
		e	NM_057167	NP_476508	Chr 2: 237.32 – 237.41 Mb
	COL6A5	a	NM_001278298	NP_001265227	Chr 3: 130.35 – 130.48 Mb
		b	NM_153264	NP_694996	Chr 3: 130.35 – 130.48 Mb
VII	COL7A1	-	NM_000094	NP_000085	Chr 3: 48.56 – 48.6 Mb
	COL8A1	a	NM_001850	NP_001841	Chr 3: 99.64 – 99.8 Mb
VIII		b	NM_020351	NP_065084	Chr 3: 99.64 – 99.8 Mb
	COLA2	a	NM_001294347	NP_001281276	Chr 1: 36.1 – 36.13 Mb
		b	NM_005202	NP_005193	Chr 1: 36.1 – 36.13 Mb
	COL9A1	a	NM_001851	NP_001842	Chr 6: 70.22 – 70.3 Mb
IX		b	NM_078485	NP_511040	Chr 6: 70.22 – 70.3 Mb
	COL9A2	-	NM_001852	NP_001843	Chr 1: 40.3 – 40.32 Mb
	COL9A3	-	NM_001853	NP_001844	Chr 20: 62.82 – 62.84 Mb
X	COL10A1	-	NM_000493	NP_000484	Chr 6: 116.12 – 116.16 Mb
		a	NM_001190709	NP_001177638	Chr 10: 34.39 – 34.4 Mb
	COL11A1	b	NM_001854	NP_001845	Chr 10: 34.39 – 34.4 Mb
		c	NM_080629	NP_542196	Chr 10: 34.39 – 34.4 Mb
XI		d	NM_080630	NP_542197	Chr 10: 34.39 – 34.4 Mb
		a	NM_001163771	NP_001157243	Chr 6: 33.16 – 33.19 Mb
	COL11A2	b	NM_080679	NP_542410	Chr 6: 33.16 – 33.19 Mb
		c	NM_080680	NP_542411	Chr 6: 33.16 – 33.19 Mb
		d	NM_080681	NP_542412	Chr 6: 33.16 – 33.19 Mb
XII	COL12A1	a	NM_004370	NP_004361	Chr 6: 75.08 – 75.21 Mb
		b	NM_080645	NP_542376	Chr 6: 75.08 – 75.21 Mb

TRACKING EVOLUTION OF COLLAGEN SUBFAMILY FOR ITS UTMOST POSSIBLE BENEFIT

		a	NM_001130103	NP_001123575	Chr 10: 69.8 – 69.96 Mb
		b	NM_005203	NP_001307880	Chr 10: 69.8 – 69.96 Mb
XIII	COL13A1	c	NM_080798	NP_542988	Chr 10: 69.8 – 69.96 Mb
		d	NM_080799	NP_542990	Chr 10: 69.8 – 69.96 Mb
		e	NM_080800	NP_542991	Chr 10: 69.8 – 69.96 Mb
XIV	COL14A1	-	NM_021110	NP_066933	Chr 8: 120.06 – 120.37 Mb
XV	COL15A1	-	NM_001855	NP_001846	Chr 9: 98.94 – 99.07 Mb
XVI	COL16A1	-	NM_001856	NP_001847	Chr 1: 31.65 – 31.7 Mb
XVII	COL17A1	a	NM_130778	NP_000485	Chr 10: 104.03 – 104.09 MB
		b	NM_000494	NP_000485	Chr 10: 104.03 – 104.09 MB
		a	NM_130445	NP_085059	Chr 21: 45.41 – 45.51 Mb
XVIII	COL18A1	b	NM_030582	NP_569711	Chr 21: 45.41 – 45.51 Mb
		c	NM_130444	NP_569712	Chr 21: 45.41 – 45.51 Mb
XIX	COL19A1	-	NM_001858	NP_001849	Chr 6: 69.87 – 70.21 Mb
		a	NM_020882	NP_065933.2	n/a
		b	XM_011528937.1	XP_011527239.1	n/a
XX	COL20A1	c	XM_011528938.1	XP_011527240.1	n/a
		d	XM_011528939.1	XP_011527241.1	n/a
		e	XM_011528940.1	XP_011527242.1	n/a
		a	NM_030820	NP_110447	Chr 6: 56.06 – 56.39 Mb
		b	NM_001318751	NP_001305680	Chr 6: 56.06 – 56.39 Mb
XXI	COL21A1	c	NM_001318752	NP_001305681	Chr 6: 56.06 – 56.39 Mb
		d	NM_001318753	NP_001305682	Chr 6: 56.06 – 56.39 Mb
		e	NM_001318754	NP_001305683	Chr 6: 56.06 – 56.39 Mb
XXII	COL22A1	-	NM_173465	NP_775736	Chr 8: 138.59 – 138.91 Mb
				NP_690850	
XXIII	COL23A1	-	NM_152890	NP_001336884	Chr 5: 178.24 – 178.59 Mb
				NP_690850	
XXIV	COL24A1	-	NM_152890	NP_001336884	n/a
			NM_001256074	NP_001243003	Chr 4: 108.81 – 109.3 Mb
XXV	COL25A1	a	NM_032518	NP_115907	Chr 4: 108.81 – 109.3 Mb

			NM_198721	NP_942014	Chr 4: 108.81 – 109.3 Mb
XXVI	COL26A1(EMID2)	a	NM_133457	NP_001265492	Chr 7: 101.36 – 101.56Mb
		b	NM_001278563	NP_597714	Chr 7: 101.36 – 101.56Mb
XXVII	COL27A1	a	NM_032161	NP_116277	Chr 9: 114.16 – 114.31 Mb
		b	NM_032888	NP_116277	Chr 9: 114.16 – 114.31 Mb
XXVIII	COL28A1	-	NM_001037763	NP_001032852	Chr 7: 7.36 – 7.54 Mb
XXIX	COL29A1	a	NM_001278298	NP_001265227	Chr 3: 130.35 – 130.48 Mb
		b	NM_153264	NP_694996	Chr 3: 130.35 – 130.48 Mb

Table:(2). Modeled 3-D structures of Selected Collagen subfamily members


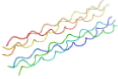
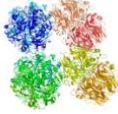
Selected Collagen type	3D structure	Favoured regions %	Additional allowed regions %	Generously allowed regions %	Disallowed regions %
COL20A1		90.1	7.8	2.1	0
COL12A1		80.1	14.9	3.5	1.5
COL4A2		83.0	10.3	4.8	1.9

Table:(3). Selection pressure on different sites of different Collagen subfamily forms

NO	Collagen	Collagen isoforms	Sequence length	0<dN/dS<1	dN/dS=1	dN/dS>1	0<dN/dS<1(%)	dN/dS=1(%)	dN/dS>1(%)
1	COL20A	A1	1284	0	1284	0	0	100	0
2	COL12A	A1	207	0	207	0	0	100	0
3	COL4A	A2	1712	1670	12	30	97.54	0.70	1.75