J. Bio-Sci. 24: 55-65, 2016 ISSN 1023-8654
http://www.banglajol.info/index.php/JBS/index

IN SILICO ANALYSES OF HUMAN COLLAGEN PROTEIN FUNCTION PREDICTION

MS Ahmed[1]*, M Kamruzzaman[2], MM Rana[1], Z Akond[1] and MNH Mollah[1]

[1]*Bioinformatics Lab., Department of Statistics, University of Rajshahi, Bangladesh*
[2]*Institute of Bangladesh Studies, University of Rajshahi, Bangladesh*

Abstract

Collagen is the extracellular matrix protein in the several connective tissues in the human body. It is an important component for mediating cell-cell interactions and pathological conditions in human body. In this study we perform the analysis of physiochemical properties and investigate the functional characteristics of human collagen proteins. Also investigate the functional protein groups by the statistical analysis. The collagen protein family consisting 28 members in human which are involving in the complex structure of protein. The protein function, protein sequence properties, domain composition, phylogenetic and protein-protein interaction (PPI) networks analysis of human collagen alpha-1 protein sequences are implemented by the online bioinformatics tools which are currently available. Based on the PCA analysis amino acid composition, features of collagen protein sequences are divided into two supreme influential functional groups such as collagen 12, 14, 20 formed one group and the rest of others formed another group. The protein-protein interaction network study using STRING showed that top interacting score of functional group proteins 0.952, 0.939 and 0.929. The most common functional domain of collagen proteins are VWC, C4, LamG, VWA, KU, C1Q, TSPN and FN3. Physicochemical, functional and phylogenetic classification can give extensive information of protein's structure and function. The depiction of alpha-1 chains of collagen protein family in human collagen 12, 14 and 20 as a prospective protein cluster. These three proteins are possess, low glycine and proline, very high aliphatic index and a close evolutionary relation in the human skin.

Key words: Collagen protein sequences, k-means clustering, PCA, phylogenetic tree, PPI network, protein domain structure

Introduction

The extracellular matrix (ECM) is consisting of collagens, proteoglycans, glycoproteins and proteases. The extracellular matrix of connective tissues represents a complex alloy of variable members of diverse protein families defining structural integrity and various physiological functions. It is the main component of connective tissue and makes up from 25% to 35% of the whole-body protein content. Collagen Type I protein found in bone, skin, muscles and walls of blood vessels in human body (Järveläinen et al. 2009). Neighboring a substantial volume of cells the ECM is an intricate network of macromolecules. For the multiple processes such as cell migration, cell-cell interaction and cell proliferation these components play vital role (Bowers et al. 2010). The collagen protein is a triple helical structure of polypeptide chains, commonly known as the alpha chains. The common sequence pattern of triple helix is "Gly-X-Y" (Kadler et al. 1996). The stability of the helical structure depends on the presence of glycine as every third residue and being other property of the smallest amino acid. Any amino acid can be taken instead of X and Y but frequently occupied by the proline residue. Every mature active collagen protein molecules were shown that the peptidases form of pro-peptides present at the N and C terminal. The genetically distinct 28 members

*Author for correspondence: shakil.statru@gmail.com found in human biological system. The collagen family is a large and a complex protein family in the proteome (Gelsea et al. 2003). The genomics, proteomics and the computational biology is an evolving field that helps to understand the concealed information of a protein structure (Nassa et al. 2012). In this study reports a qualified portrayals of alpha-1 sequences of human collagen using statistical methods and bio-computational tools. Three alpha chains present in collagen alpha-1 was pragmatic human collagen family. To analysis of physiochemical properties, functional group, phylogenetic classification, domain composition and PPI networks of human collagen using the amino acid sequence features. Objective of this research is to afford an intuition to various proteins features of collagen proteins and represent this large family. The presence of any nonconforming entrances in the collagen molecules for impending adherents to implicate the disease in collagen family (Bateman et al. 2009). Most of the previous studies are based on the wet lab experimental and it is very much time consuming, costly and laborious for identification of functional collagen proteins in the human skin. In this case we have suggested some *in silico* and statistical methods for functional analysis of collagen proteins. It might be reduced time and cost in comparison with experimental methods. This study might be helpful for biologists in stentorian of promote soundings on these complex molecules, interacting proteins, and functional domains for human diseases in collagen family.

Materials and Methods

Collection of human collagen apha-1 sequences

Human collagen family alpha-1 protein sequences of all 28 members were collected in FASTA format using the accession number provided for each collagen sequence from UniProtKB/ SWISS-PROT (http://expasy.org/sprot/) (Bairoch et al. 2000) protein database.

Physiochemical Portrayal of Human Collagen Family

Various features including the 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 by inputting the protein sequence in FASTA format (http://expasy.org/tools/protparam.html).

Protein domain composition of human collagen family

The protein domain composition and post translation sites prediction using SMART (http://smart.embl-heidelberg.de/) (Letunic et al. 2012) bioinformatics tool also used to scan and identify all the known domain. To predict the nature and position in the selected alpha-1 protein sequences of the collagen family based on a profile and pattern search. The input protein sequence in FASTA format was used for a selected protein profile in the database.

Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignment of the human alpha-1 collagen protein sequences were aligned using MEGA5.0 tools (Tamura et al. 2011), the sequence alignment algorithm was used ClustalW of protein sequences in FASTA format as the input data type. For a set of input sequences the best alignment was computed and all the identities. The phylogenetic tree or evolutionary tree was customary by constructing phylograms through recovery of the alignments using Neighbor Joining (NJ) method.

Protein-protein interaction (PPI) networks analysis

The accurate prediction of protein functions is important for interacting residues with each other. This study used STRING (http://string-db.org/) a database of known and predicted protein interactions networks through physical and functional associations (Andrea et al. 2013). The input protein sequence was used in FASTA format for prediction of PPI networks.

Statistical analysis of collagen protein sequences

The analysis of amino acid functional group was used k-means clustering approach. To investigate the collagen protein functional group we used the multivariate statistical techniques principal component analysis (PCA), it is very much popular techniques in bioinformatics data analysis. In this paper, all the statistical analysis likes k-means clustering and PCA were done using R-packages(R 3.2.0) and MS Excel-2010.

Results and Discussion

The portrayals of human collagen alpha-1 extracellular matrix protein are most important for human skin. Collagen protein sequence physiochemical properties analysis was done by the ExPASy ProtParam online tools (Table 1). The highest aliphatic index is 79.61 of Col20 (Fig. 1a) was regarded as the thermostable and Col14 (77.67) and Col12 (75.45). The GRAVY values (Fig.1 b) indicate the range from -2 to +2 of proteins are positively related with the proteins being more hydrophobic (Kyteet al. 1982).

Table 1. Physiochemical properties of collagen protein family.

Collagen Proteins	Accession number	No. of AA	Molecular weight	pI	-ve charged residue	+ve charged residue	Extinction Coefficient	Instability index	Aliphatic index	GRAVY
Coll 1	P02452	1464	138941.5	5.6	141	128	53495	30.43	37.98	-0.788
Coll 2	P02458	1487	141785.3	6.58	141	139	54525	25.21	40.03	-0.803
Coll 3	P02461	1466	138564.2	6.21	129	122	62225	30.18	37.31	-0.797
Coll 4	P02462	1669	1606147.7	8.55	128	138	61070	32.04	47.39	-0.621
Coll 5	P20908	1838	183559.8	4.94	225	168	98850	33.09	45.35	-0.873
Coll 6	P12109	1028	108529.4	5.26	139	114	64970	28.52	68.70	-0.525
Coll 7	Q02388	2944	295219.6	5.95	332	310	159140	32.07	61.86	-0.625
Coll 8	P27658	744	73364	9.62	37	60	38405	36.06	61.21	-0.434
Coll 9	P20849	921	91869.2	8.94	86	96	42565	32.61	56.13	-0.658
Coll 10	Q03692	680	66157.9	9.68	34	54	42290	25.95	51.94	-0.556
Coll 11	P12107	1806	181064.8	5.06	222	174	103765	30.81	44.91	-0.859
Coll 12	Q99715	3063	333146.7	5.38	366	313	334620	32.90	75.45	-0.427
Coll 13	Q5TAT6	717	69949.9	9.27	67	81	15970	31.44	52.87	-0.765
Coll 14	Q05707	1796	193515.4	5.16	211	160	179095	37.57	77.67	-0.326
Coll 15	P39059	1388	141720.1	4.90	155	95	76485	40.19	68.00	-0.377
Coll 16	Q07092	1604	157751.3	8.14	144	150	65370	35.88	50.73	-0.671
Coll 17	Q9UMD9	1497	150419.3	8.89	117	128	109015	45.25	55.47	-0.573
Coll 18	P39060	1754	178187.6	5.67	164	133	145185	48.57	61.72	-0.467
Coll 19	Q14993	1142	115220.7	8.57	116	124	63215	30.68	56.68	-0.708
Coll 20	Q9P218	1284	135830	8.27	119	123	132990	45.18	79.61	-0.261
Coll 21	Q96P44	957	99368.5	8.57	98	106	55655	33.28	69.14	-0.517
Coll 22	Q8NFW1	1626	161145.3	6.88	174	172	57965	34.00	53.28	-0.715
Coll 23	Q86Y22	540	51943.9	6.88	65	65	14355	30.81	50.69	-0.829
Coll 24	Q17RW2	1714	175496.3	8.46	162	170	73075	28.32	64.21	-0.622
Coll 25	Q9BXS0	654	64770.7	8.60	73	78	11835	24.85	47.19	-0.919
Coll 26	Q96A83	441	45381.1	7.02	40	40	40170	46.77	63.11	-0.523
Coll 27	Q8IZC6	1860	186892.3	9.83	136	196	81205	37.62	54.15	-0.637
Coll 28	Q2UY09	1125	116657.1	6.10	136	131	55195	24.18	61.42	-0.66

The collagen 20 GRAVY is -0.261 then we may state that it is most hydrophobic protein than others. From the table collagen 15, 17, 18, 20 and 26 are unstable (instability index >40) and rest of the proteins are stable (instability index <40), the all values of instability index shows in Fig. 1(c).The 14 collagens pI (Fig. 1d) are less than 7, Col26 is approximate equal to 7 and rest of the greater than 7; hence the 14 collagens are acidic, Col26 is neutral and others collagen proteins are basic (Lim 2006).

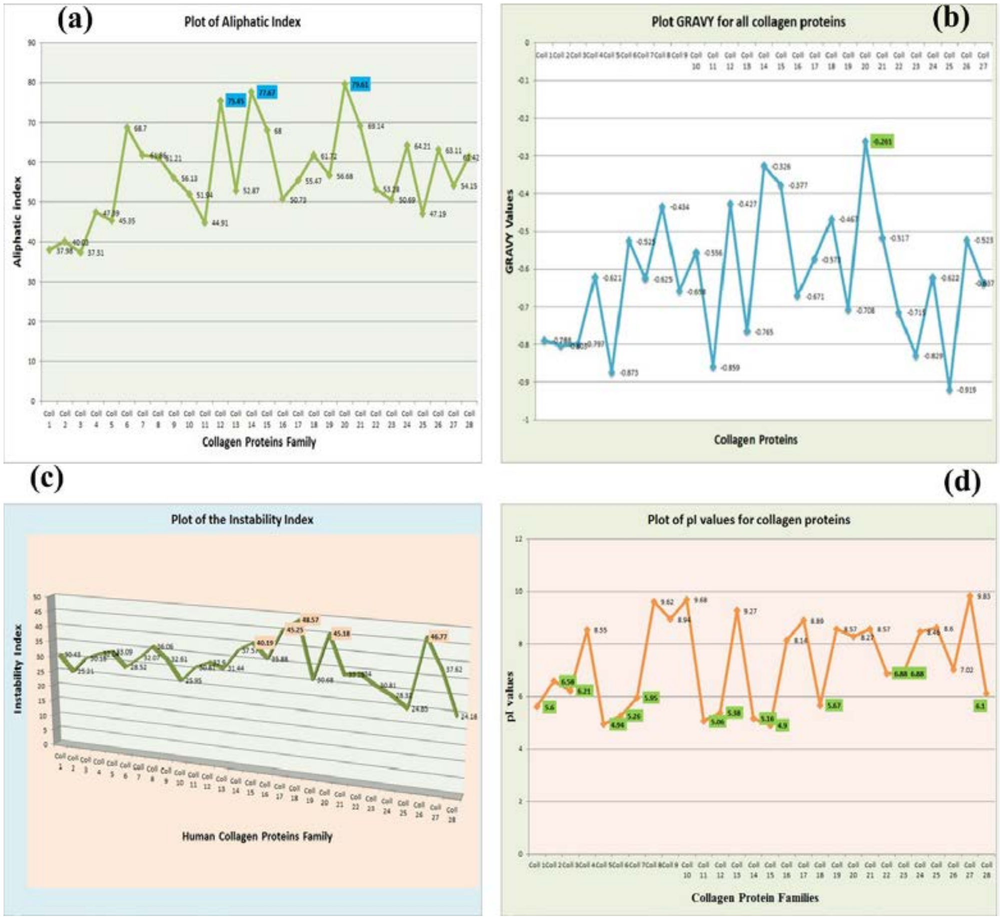


Fig. 1. (a) Aliphatic index, (b) GRAVY values, (c) Instability index and (d) pI values for all 28 human collagen alpha-1 families.

The common domain structure is COLFI (C-termini of Fibrillar collagen) of Coll1, Coll2, Coll3, Coll5, Coll11, Coll14 (Fig. 2) and the other domains are VWC (Von Willebrand factor type C), C4 (C-terminal tandem repeated), LamG (Laminin G), VWA (von Willebrand factor type A), FN3 (Fibronectin type-III), C1Q, TSPN (Thrombospondin N-terminal), FRI (Frizzled) and KU (BPTI/Kunitz family of serine protease inhibitors). From the Table 2 shown that Coll13, Coll15, Coll23, Coll25 and Coll26 proteins has no functional domain out of 28 human collagen protein families in the human skin. The maximum number of domains exists in the Coll12 (E-values: 1.28e-8 to 1.25E-78), Coll14 (E-values: 4.22e-9 to 0.214) and Coll20 (E-values: 19.6e-55 to 2.89e-33) proteins, those three protein domains are more functional activity of the ECM proteins.

Table 2. Human collagen alpha-1 protein domains.

Collagen Proteins	Source Gene	Domain Name	Start	End	E-value
Col 1	ENSG00000108821	VWC	40	95	2.73E-20
		COLFI	1228	1464	1.25E-166
Col 2	ENSG00000139219	VWC	34	89	7.42E-22
		COLFI	1252	1487	6.46E-183
Col 3	ENSG00000168542	VWC	32	88	1.68E-20
		COLFI	1231	1466	1.15E-165
Col 4	ENSG00000187498	C4	1445	1554	3.55E-66
		C4	1555	1668	3.70E-78
Col 5	ENSG00000130635	TSPN	39	230	6.53E-82
		LamG	98	229	6.50E-04
		COLFI	1608	1837	1.06E-155
Col 6	ENSG00000142156	VWA	35	233	4.29E-31
		VWA	613	790	8.00E-31
		VWA	827	1008	2.72E-33
Col 7	ENSG00000114270	VWA	36	216	4.54E-53
		FN3	232	318	5.73E-11
		FN3	327	402	6.54E-06
		FN3	415	493	9.11E-05
		FN3	508	584	1.64E-06
		FN3	598	674	1.94E-08
		FN3	686	764	1.16E-11
		FN3	776	853	6.35E-04
		FN3	867	943	7.45E-10
		FN3	955	1039	5.04E-07
Col 8	ENSG00000144810	C1Q	609	744	3.27E-79
Col 9	ENSG00000112280	TSPN	50	244	2.02E-87
Col 10	ENSG00000123500	C1Q	545	680	6.20E-80
Col 11	ENSG00000060718	TSPN	38	229	1.15E-68
		LamG	97	228	9.48E-06
		COLFI	1576	1805	7.39E-128
Col 12	ENSG00000111799	FN3	25	103	1.28E-08
		VWA	138	317	1.21E-62
		FN3	334	413	1.35E-07
		VWA	438	617	5.62E-58
		FN3	632	710	2.16E-06
		FN3	723	801	1.74E-10
		FN3	814	892	6.59E-11
		FN3	905	984	2.23E-08
		FN3	995	1074	9.54E-08
		FN3	1087	1166	4.09E-07
		VWA	1197	1376	3.46E-58
		FN3	1385	1463	2.46E-10
		FN3	1474	1554	3.29E-11
		FN3	1566	1643	9.83E-10
		FN3	1655	1734	7.63E-07
		FN3	1753	1832	1.09E-11
		FN3	1844	1922	4.09E-07
		FN3	1934	2013	9.69E-09
		FN3	2025	2104	3.73E-10
		FN3	2116	2193	7.57E-11
		FN3	2204	2283	6.35E-04
		VWA	2321	2501	7.09E-55
		TSPN	2520	2712	1.25E-78

Table 2 Contd.

Col 13	ENSG00000197467	-	-	-	-
Col 14	ENSG00000187955	FN3	30	108	4.22E-09
		VWA	156	335	3.36E-56
		FN3	353	433	3.32E-07
		FN3	443	521	6.20E-07
		FN3	535	612	8.83E-12
		FN3	624	703	4.77E-08
		VWA	1030	1210	7.53E-59
		TSPN	1229	1424	2.46E-68
		FN3	735	817	9.25E-06
		FN3	829	908	1.45E-07
		FN3	919	998	2.14E-01
Col 15	ENSG00000204291	-	-	-	-
Col 16	ENSG00000084636	TSPN	50	231	3.82E-74
Col 17	ENSG00000065618	-	-	-	-
Col 18	ENSG00000182871	FRI	333	448	1.25E-25
		TSPN	456	644	6.32E-57
		LamG	505	643	1.11E-01
Col 19	ENSG00000082293	TSPN	50	234	1.01E-73
Col 20	ENSG00000101203	FN3	26	102	1.96E-54
		VWA	177	356	7.66E-51
		FN3	377	457	1.57E-08
		FN3	466	546	1.13E-09
		FN3	557	636	1.55E-07
		FN3	647	726	2.72E-03
		FN3	741	820	4.12E-12
		TSPN	842	1037	2.89E-33
Col 21	ENSG00000124749	VWA	35	212	3.02E-49
		TSPN	230	412	2.18E-19
Col 22	ENSG00000169436	VWA	36	218	3.83E-51
		TSPN	239	427	1.55E-33
Col 23	ENSG00000050767	-	-	-	-
Col 24	ENSG00000171502	TSPN	68	228	1.02E-05
		COLFI	1514	1714	3.85E-35
Col 25	ENSG00000188517	-	-	-	-
Col 26	ENSG00000160963	-	-	-	-
Col 27	ENSG00000196739	TSPN	45	222	1.46E-05
		OLFI	1659	1860	1.41E-42
Col 28	ENSG00000215018	VWA	46	228	3.06E-18
		VWA	796	973	3.02E-40
		KU	1070	1123	1.08E-19

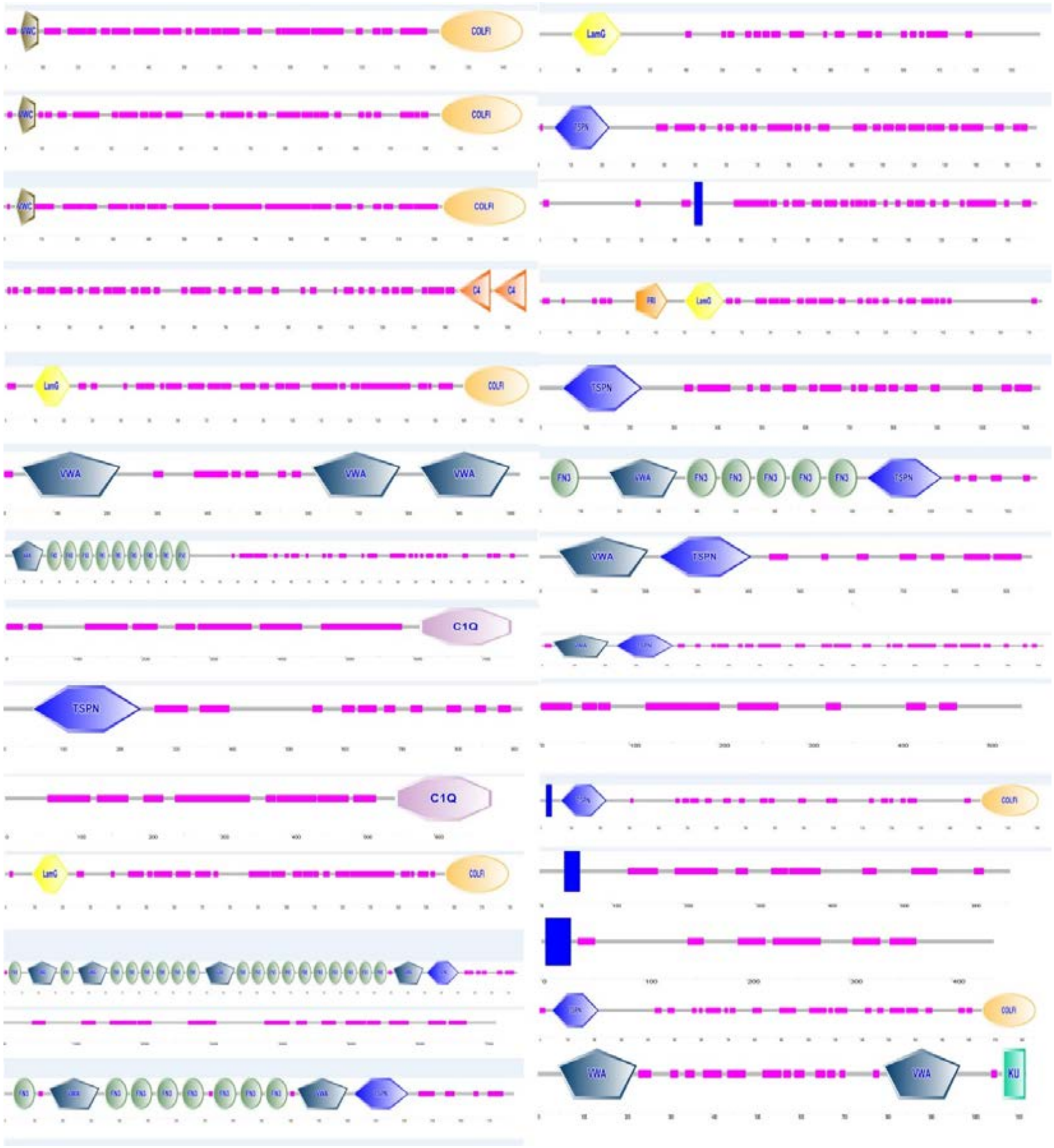


Fig. 2. Protein domain structure of 28 human collagen alpha-1 families.

For the evolutionary study, the phylogenetic analysis shows (Fig. 3a) shows four group of collagen proteins in the 28 family members; the CollXXV and CollXXIII (bright pink), CollXX, CollXII, CollIV (bright green), CollXXVI (red) and rest of the collagen are same groups (blue). In the blue collagen group, there are two different sub-groups CollVI (green) and CollXXVIII (maroon). Hence the CollXXVI is completely separate from the other functional protein groups. Clustering is the unsupervised techniques in machine learning approaches to cluster different groups based on the within groups similarity and between groups dissimilarity.

For the k-means clustering approach first important issues is the selection of number of k; there are several methods exists in the literature for selecting the optimum number of k. Scree plot one of them popular methods for selecting k (Fig. 3b), it was shown two (k = 2) optimum clusters for clustering the amino acid properties. By the k-means clustering shown that in different k = 2, k = 3, k = 4 (Fig. 3c) than the finest two functional properties in the amino acid i.e. positively charged and negatively charged. The principal component analysis (PCA) based clustering approach is the modern multivariate techniques. In this paper we used this techniques for identify the joint functional protein complex in the human collagen protein families (Fig. 3d) using biplot; the standardized PC1 is explained 90.6% and standardized PC2 is 4.9% with compare the total features. Therefore, the two functional protein complexes are in the human collagen alpha-1 proteins. The collagen 12, 14 and 20 are similar protein complex shows the similar properties and rest of the collagens is others group.

The protein-protein interaction network study investigate the jointly and similar functional activity based on the interacting score. The PPI networks analysis (Fig. 4) for identifying the most interacting functional collagen protein groups of 28 human collagen families was done using the STRING database. The top interacting score of Coll12, Coll14 and Coll20 proteins are 0.952, 0.939 and 0.929 respectively (Fig. 5).

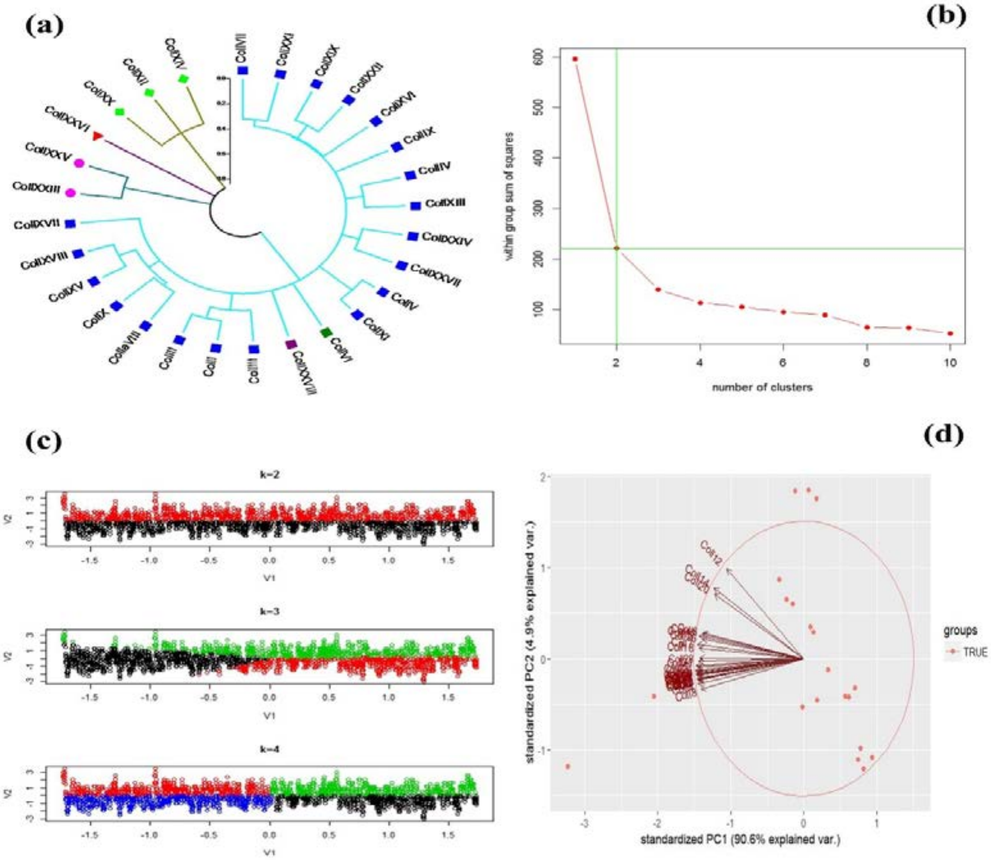


Fig. 3. (a) Phylogenetic tree, (b) Scree plot, (c) K-means clustering and (d) Biplot for the analysis of human collagen family (28) of alpha-1.

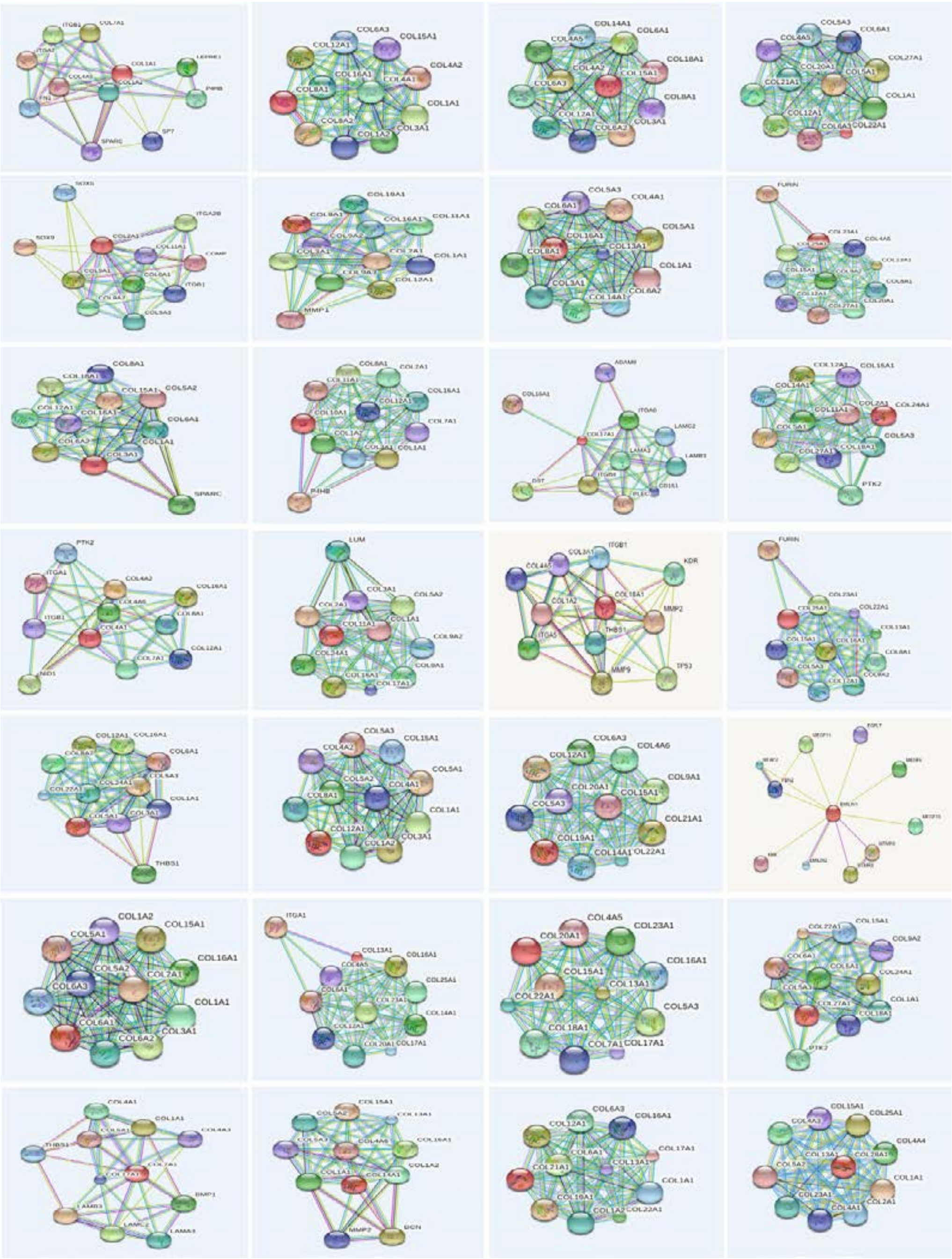
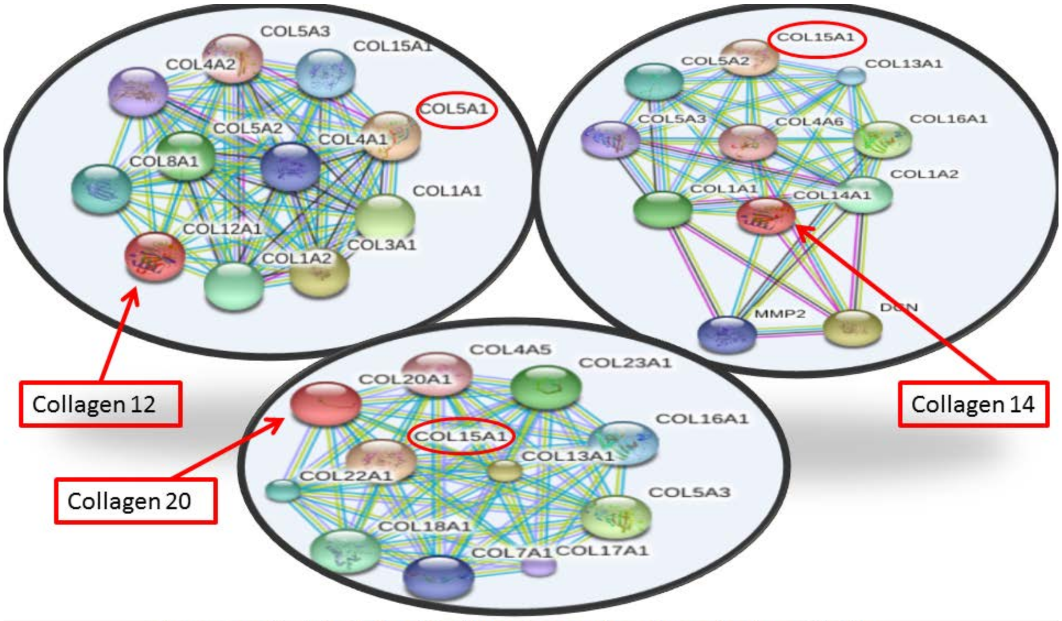
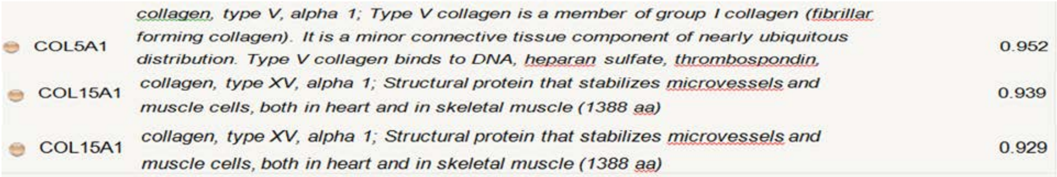


Fig. 4. Protein-protein interaction (PPI) network for finding the most similar functional interacting proteins by the STRING data base of 28 human collagen families.





COL5A1	collagen, type V, alpha 1; Type V collagen is a member of group I collagen (fibrillar forming collagen). It is a minor connective tissue component of nearly ubiquitous distribution. Type V collagen binds to DNA, heparan sulfate, thrombospondin,	0.952
COL15A1	collagen, type XV, alpha 1; Structural protein that stabilizes microvessels and muscle cells, both in heart and in skeletal muscle (1388 aa)	0.939
COL15A1	collagen, type XV, alpha 1; Structural protein that stabilizes microvessels and muscle cells, both in heart and in skeletal muscle (1388 aa)	0.929

Fig. 5. Protein-protein interaction network for most influential functional interacting group proteins by the STRING data base of Coll12, Coll14 and Coll20 based on the human collagen families.

Conclusion

The *in silico* analysis of extracellular matrix proteins is the most important for studying the functional characteristics of human collagen families. The analysis of functional protein domain shows the lots of significant domains are Coll12, Coll14 and Coll20. In this study we used the positively and negatively charged amino acids, that's justify by the screen plot and k-mean clustering approaches. By the principal component analysis approaches it is shown that, the first 2 PC's are explained approximately 95% out of the total variations. The PC's score plot gives us the two most important functional groups, one of them group collagen proteins are collagen 12, 14 and 20 respectively. The above discussion shown that the most important functional collagen proteins are Coll12, Coll14 and Coll20 based on the several analysis tools including statistical techniques. Those collagens are the FACIT (Fibril Associated collagens with Interrupted Triple helices) group of collagen family. This *in silico* study is very much helpful for biologist to analysis of the ECM collagen alpha-1 28 protein families of human skin by the reducing the experimental cost, saving consuming time and laborious work in this field.

References

Andrea F, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Mering CV and Jensen LJ (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration, Nucleic Acids Research 41: 808-815.

Bairoch A and Apweiler R (2000). The SWISS-PROT Protein Sequence Database and Its Supplement Tr EMBL, Nucleic Acids Research 28(1): 45-48.

Bateman JF, Boot-Handford RP and Lamandé SR (2009). Genetic diseases of connective tissues: cellular and extracellular effects of ECM mutations, Nature Reviews Genetics 10(3): 173-83.

Bowers SL, Banerjee I and Baudino TA (2010). The extracellular matrix: at the center of it all, Journal of Molecular and Cellular Cardiology 48(3): 474-82.

Gelsea K, E Poschlb and T Aignera (2003). Collagens-structure, function, and biosynthesis, Advanced Drug Delivery Reviews 55: 1531-1546.

Ivica Letunic and Bork TDP (2012). SMART 7: recent updates to the protein domain annotation resource, Nucleic Acids Research 40: 302-305.

Järveläinen H, Sainio A, Koulu M, Wight TN and Penttinen R (2009). Extracellular matrix molecules: potential targets in pharmacotherapy, Pharmacological Reviews 61(2): 198-223.

Kadler KE, Holmes DF, Trotter JA and Chapman JA (1996). Collagen fibril formation, Journal of Biochemistry 15(1): 1-11.

Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei and Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods, Molecular Biology and Evolution 28(10): 2731-2739.

Kyte J and Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157: 105-132.

Lim KF (2006). Negative pH Does Exist, Journal of Chemical Education 83(10): 1465.

Nassa M, Anand P, Jain A, Chhabra A, Jaiswal A, Malhotra U and Vibha Rani (2012). Analysis of human collagen sequences, Bioinformation 8(1): 026-033.