Predicting bonds between collagens – A multiple alignment approach

Valentin Wesp[1*], Lukas Scholz[1], Janine M. Ziermann-Canabarro[2], Stefan Schuster[1], Heiko Stark[1,3]

[1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany

[2]Department of Anatomy, Howard University College of Medicine, Washington, U.S.A.

[3]Institute of Zoology and Evolutionary Research, Friedrich-Schiller-University Jena, Germany

[*]**Corresponding author:**

Valentin Wesp

Department of Bioinformatics

Matthias-Schleiden-Institut für Genetik, Bioinformatik und Molekulare Botanik

Friedrich-Schiller-University Jena

07743 Jena

Germany

E-Mail: valentin.wesp@uni-jena.de

Abstract

Collagens are structural proteins that are predominantly found in the extracellular matrix, where they are mainly responsible for the stability and structural integrity of various tissues. There are several different types of collagens, some of which differ significantly in form, function, and tissue specificity. Subdivisions into so-called collagen families exist, which are defined on the basis of mainly clinical research. Collagens contain polypeptide strands (α-chains). Their sequences are often analysed in view of clinical aspects, but problems arise with highly homologous sequence segments. To increase the accuracy of collagen classification and prediction of their functions, the structure of these collagens and their expression in different tissues could result in a better focus on sequence segments of interest. Here, we analyse collagen families with different levels of conservation. As a result, families with strong interchain hydrogen bonds can be found, such as fibrillar collagens, network-forming collagens, and FACITs. Moreover, collagen IV α-chains form their own cluster, and other collagen α-chains do not bond at all.

Introduction

Collagen

Collagens are structural proteins of the extracellular matrix (ECM) in multicellular animals and, with a mass fraction of more than 30 %, constitute an essential part of its structure and maintenance [1,2,3,4]. Collagens are found in skin, connective tissue, vascular walls, bone, and cartilage [5,6,7,8,9]. Moreover, the expression of collagens is essential for the formation of the basic structure of tissues during ontogenesis [10]. For example, a somite is a block of condensed mesoderm formed bilaterally along the central axis in vertebrate embryos [11]. Collagens are not only essential for the formation of these blocks but also for the differentiation of the somitic subunits, i.e., dermatome, myotome, and sclerotome [12,13]. The fusion of myoblasts into multinucleated myofibers is another essential task of one of the collagens [14]. There are many different types of collagens that vary widely in their form, function, and tissue specificity. Currently, 28 different types of collagens are known, which can be divided into six different families [15,16,17] (Table 1). A classification is realised by the characteristic properties of the collagen such as its structure, interaction, or site of expression [18]. Due to the variability of characters, various classifications of collagen exist, with none of them being indisputable [19,20,21,22]; Table 1 is thus only one way of classifying them.

Table 1. Classification of the 28 collagens into six different groups based on Gordon & Hahn (2010). Collagens are colour coded according to their group, which is also used further in this manuscript. FACITs = Fibril Associated Collagens with Interrupted Triple helices.

Group	Member
Fibrillar collagens	**I, II, III, V, XI, XXIV, XXVII**
Associated with fibrils (FACITs)	**IX, XII, XIV, XVI, XIX, XX, XXI, XXII**
Network-forming collagens	**IV, VI, VII, VIII, X**
Transmembrane collagens	**XIII, XVII, XXIII, XXV**
Endostatin precursor collagens	**XV, XVIII**
Other collagens	**XXVI, XXVIII**

Most collagens (22) are homotrimers, i.e., they are built from three identical polypeptide units called α-chains which are linked by hydrogen bound [18,23,24,25,26]. However, there are also six heterotrimers (collagens I, IV, V, VI, IX and XI) that consist of two identical α-chains and one that differs from the other two. These trimers are transported from the interior of a cell as procollagens into the extracellular matrix [27,28]. The resulting tropocollagens aggregate with collagen microfibrils in the environment through interchain hydrogen bonds at the telopeptide regions, which are further organised as fibrils [29,30,31. These fibrils are further packed together into fibres, which in turn are bundled into fascicles [32,33].

Collagens are characterised by specific domains that are connected by a repetitive sequence [34]. This repetitive sequence is predominantly involved in structure formation. The domains, however, bind specifically to other domains and can thus form linkages (networks). These linkages show similarities in terms of their sequence. Another characterisation of collagens is based on their abundance of their collagen types and mRNA expression throughout the human body [35,36]. Yet, as shown in Table 2, the mRNA localization of individual collagens is often tissue-specific [35,37]. The understanding of the sequence and structure of collagens is crucial in view of hereditary diseases based on mutations in collagen genes, such as osteogenesis imperfecta, Ehlers-Danlos syndrome, and Stickler syndrome.

Table 2. Characterisation of the different α-chains to the individual tissues and germ layers based on The Human Protein Atlas (*https://www.proteinatlas.org/*, 2023) [35,38]. An α-chain is assigned to a tissue if there is a relatively higher mRNA expression value for this α-chain in this tissue. The colours correspond to the groups in Table 1.

Germ layer	Tissue	Collagen α-chains
Endoderm	Gallbladder	**Iα1, Iα2, IIIα1**
Endoderm	Pancreas	**IIα1**
Endoderm	Colon, Rectum	**VIα1**
Endoderm	Lung	**VIα5**
Endoderm	Parathyroid gland	**VIα6**
Endoderm	Thyroid gland	**IXα3**, **XXIIIα1**
Endoderm	Liver	**XVIIIα1**, **XXVIα1**

Endoderm	Urinary bladder	**IVα6**
Mesoderm	Endometrium	**IVα6**
Mesoderm	Placenta	**IIIα1, XIα1, XXIα1, IVα1, IVα2, XVα1**
Mesoderm	Testis	**XIα2, XXα1, XXVα1**
Mesoderm	Lymphoid tissue	**XIXα1**
Mesoderm	Heart	**XXIIIα1**
Ectoderm	Salivary gland	**XXVIIIα1**
Ectoderm	Brain	**XXIVα1, IXα2, IXα3, XIXα1, XXα1, Xα1, IVα5, XXVα1, XXVIα1**
Ectoderm	Retina	**IIα1, XIα1, IVα3, XXVα1**
Ectoderm	Skin	**VIα5, VIIα1, XVIIα1**

AlphaFold

AlphaFold DB is a protein structure database (DB) and prediction tool created by DeepMind and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) that uses a deep neural network to predict spatial structures for proteins, similar to a graph interference problem [39,40]. It maximises the number of interactions according to a graph-theoretical model considering the similarity across structures. The advantage of AlphaFold is that this, in theory, is possible for all known amino acid sequences.

However, this requires prior measurements of 3D structures to validate these sequences, which is determined by a confidence value. This value indicates for each position in the amino acid sequences how well the prediction matches the measurement. AlphaFold categories these confidence values (CV) into four groups with respect to their accuracy. These groups are 'high accuracy' (CV>90), 'generally good backbone prediction' (70<CV<90), 'low accuracy' (50<CV<70) and 'ribbon-like appearance' (CV<50). AlphaFold suggests that the last group should not be interpreted.

Screening of the collagen sequences revealed that the proportions of repetitive regions were assigned a very low confidence value. These repetitive regions are found mainly

in the inside of the sequences, as they are instrumental in the structure of the collagen. However, they are disruptive in an analysis due to their repetitions, as they occur more frequently than the functional regions of the protein. Yet, the boundary regions seem to be highly conserved, as they return a high confidence value.

Aim of this study

It is surprising that there are only a few complete studies regarding the possible alignment of collagens to each other. Their importance for ontogenesis and structure formation in the body is indisputable [41,42]. In this study, the 44 α-chains of all 28 available human collagens from all six families are analysed by bioinformatics techniques.

The aim of this study is to reconstruct possible hydrogen bonds using multiple sequence alignments of all human collagen α-chains [18] (Table 1) according to their highly conserved regions, i.e., telopeptide regions. The idea behind this is that similar sequences have a higher probability of forming a larger amount of hydrogen bonds, favouring a stable bond. For example, the canonical collagen motif $(Gly\text{-}Pro\text{-}Pro)_n$ easily forms hydrogen bonds to neighbouring strands if they involve the same motif because they match well in space [43,44]. With this knowledge, it should be possible to predict the function of a tissue with a specific expression pattern.

Material & Methods

Sequence preparation

The sequence information of human collagen variants was downloaded from the AlphaFold DB based on AlphaFold with the version from the “14th Critical Assessment of Protein Structure Prediction” [39,40] (Figure 1). In AlphaFold, over 200 million protein structure predictions are stored and freely available.

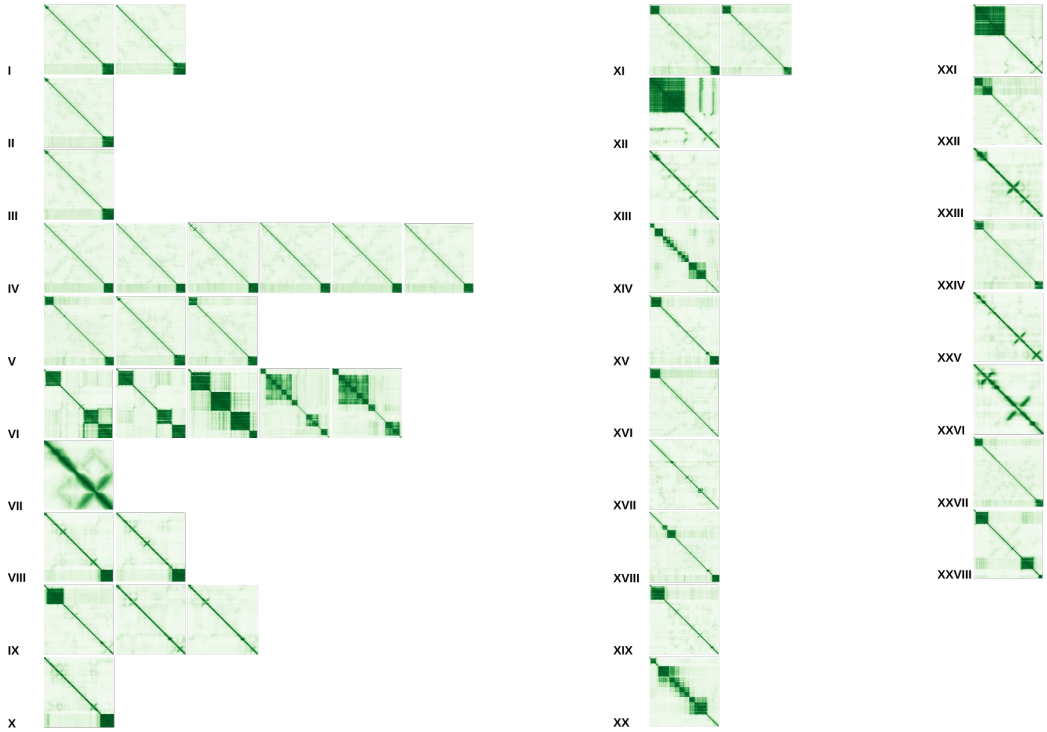

Figure 1. Overview of the predicted aligned confidence of the collagen α-chains of collagens I-XXVIII adapted from AlphaFold [39,40]. The colour indicates the expected position error when the predicted and true structures are aligned. The darker the colour, the lower the alignment error.

The sequences and their position confidence values of the collagen α-chains were downloaded in the form of PDB files. However, the PDB format is atypical for most multiple sequence alignment (MSA) algorithms and was, therefore, reformatted into a FASTA format. To process the following steps, several custom python (v3.10) and R (v4.2.3) scripts were implemented (Supplement 5). First, all amino acids were transferred from the 3-letter code to the 1-letter code based on the International Union

of Pure and Applied Chemistry (IUPAC) nomenclature. Then, high confidence regions from each α-chain were extracted (Figure 2). Subsequences with a length of 200 amino acids and the highest average confidence score were selected and stored. This was repeated maximally ten times for each α-chain. To ensure that parts of each region do not appear twice or more often, after the extraction of each subsequence the confidence values of the particular region of the α-chain were set to $-1*10^9$. Furthermore, for each subsequence, the reversed complement was computed (marked with "rev" in the sequence name). For each confidence score between zero and 95 in steps of five subsequences were only retained if their average confidence score was above the given threshold. All other sequences were removed from the set. The sequences in each set were then merged into a single multifasta file.

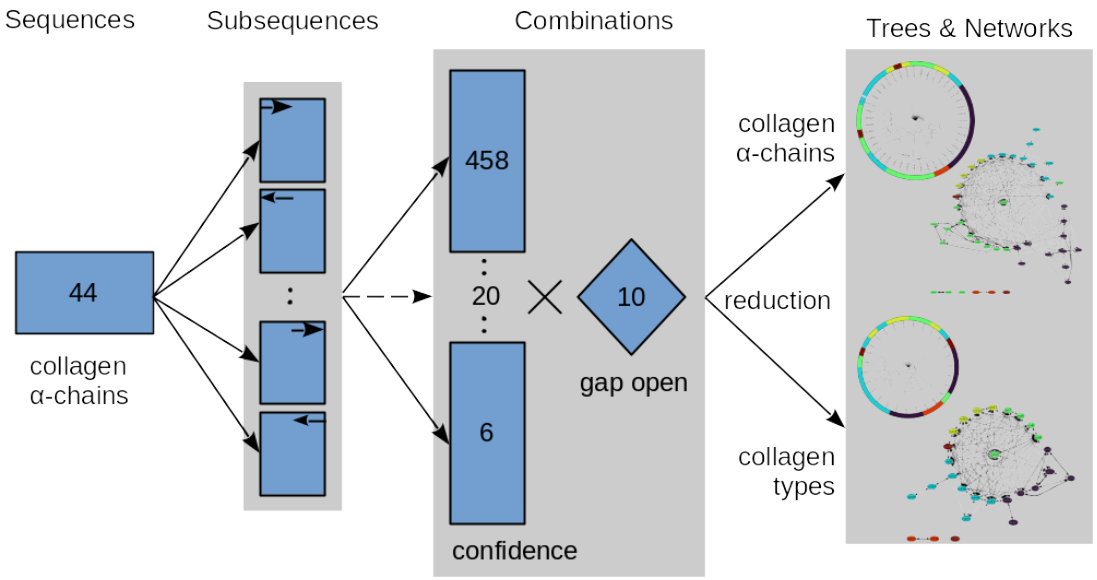

Figure 2. Flowchart of data processing for the 44 collagen α-chains. From left to right, we have collected 44 collagen α-chains. From these, subsequences are extracted based on all confidence values between zero and 95 in steps of five (between 458 subsequences for confidence value zero and six subsequences for confidence value 95). Afterwards, for each confidence value, multiple alignments were produced for all gap open costs between one and ten in steps of one (overall 200 multiple alignments). In the end, all multiple alignments were graphically arranged according to their distance matrix values in the form of phylogenetic trees and networks. Each node was colourized based on the colours in Table 1.

Construction of multiple sequence alignment

To create phylogenetic trees and networks, a MSA is required [45,46]. For the phylogenetic reconstructions, an alignment of sequences is needed to identify conserved regions and differences. The amount of differences is traditionally taken as

indication of distance between compared sequences. For the alignment process, the tool MAFFT (Multiple Alignments using Fast Fourier Transformation, v7.310) was chosen. MAFFT can be executed via a web interface as well as via a downloadable command line program. MAFFT requires a substitution matrix to evaluate different amino acids substitutions in terms of their properties. Since collagens are aligned according to their hydrophobicity, a custom matrix based on the Eisenberg consensus scale which combines several hydrophobicity features into one score was constructed (Supplement 2.1). The difference in hydrophobicity for each amino acid pair was taken as the substitution rate and normalised between zero and 20. The rate values were rounded to the next integer.

The multiple alignment was built heuristically by means of progressive alignments and Fourier transforms [47]. Here, the following parameters were set: **globalpair** (GINSI), **op** <m>, **ep** <n>, **maxiterate** <1000>, **aamatrix** <matrix_file>, **inputorder**, **thread** <6>. The parameter **globalpair** describes that the sequences are globally aligned with each other; **op** is the gap open (indicated by *m*) and **ep** the gap extend costs (indicated by *n*); **maxiterate** specifies the maximum number of iterations the algorithm runs for the progressive alignment process (here 1000); and **aamatrix** specifies the custom hydrophobicity matrix to be used. The **inputorder** option sets the order of the sequences the same as in the input file. Lastly, **thread** specifies the number of CPUs used for MAFFT (here six). To choose which **gap open** and **gap extent** costs produce the best outcome, for each multifasta file MAFFT was run for **gap open** costs between one and ten in steps of one. The **gap extend** costs were set to one tenth of the **gap open** costs. The output alignments were analysed for hydropathy using the client-based tool *AlignmentViewer* (https://alignmentviewer.org/, 01.02.2023) [48].

Distance matrix generation

IQ-TREE (v1.6.12), a fast and effective stochastic algorithm for inferring phylogenetic trees by maximum likelihood, was used to generate the trees [49]. IQ-TREE aims to infer the best protein model to build its trees, but it can also take a custom model as input. Here, the same model as in the MAFFT process was chosen, except the rates were not normalised or rounded (Supplement 2.2). The following parameters were set: **s** <aln_file>, **st** <AA>, **m** <model_file>, **bb** <1000>, **alrt** <1000>, **nt** <6>. **s** is the input alignment file. **st** specifies the input sequence types as amino acids. **m** is the hydrophobicity model file to be used. For the bootstrap procedures, **bb** sets the number

of replicates for the ultrafast bootstrap and **alrt** for the SH-like approximate likelihood ratio test to 1000 for both. Lastly, **nt** specifies the number of CPUs used for IQ-TREE (here six). The output are distance matrices and NEWICK files.
To better compare our results with the literature, a data compression was performed. Therefore, the subsequences were assigned to one of two categories, one containing all α-chains and one containing only the collagen types. For the α-chains (44), the minimum distance between two sets was selected. For the collagen types (28), the minimum distance between the subsequences of all α-chains of the compared collagen types was selected. In a further step, lists were generated from these two compressed datasets listing the number of pairings between each α-chain subset and each collagen subset. For these distances, thresholds were set between 0.5 and 1.5 in steps of 0.1. This was done for each gap open cost and confidence value.
These reduced distance matrices were visualised with the general-purpose diagramming program yEd (v3.23.1) to show the bonds between the various α-chains and collagen types.
Furthermore, the reduced distance matrices are used to construct new trees in Newick format with a R script which are visualised with interactive Tree Of Life (iTOL, 09.02.2023) [50].

Parameter fitting

To find the optimal parameters for the best pairing of subsequences, the parameter space had to be explored (Supplement 1.3.6). An upper bound can be found for the confidence values due to the missing α-chains at higher values.

Reference distance matrix generation

To extrapolate the bonds between the different α-chains, we constructed a reference matrix between all α-chains based on reference databases (Supplement 3). The idea behind this is that even if only neighbour links are considered, a larger network should still result. This was done for GoogleScholar (entire text) and PubMed (only title and abstract) databases. For this purpose, a search query was performed between two collagens (GoogleScholar) or α-chains (Pubmed) for all collagens and α-chains in humans. The number of hits was stored in a matrix. To avoid bias by focusing on

specific collagen types in the literature, these values were logarithmized.

Comparison of trees in literature

Lastly, we compared our constructed trees with those for collagen in the literature (Supplement 4). For this, we took the phylogenetic tree of the first α-chains of all 28 collagens constructed by Nassa et al. (2012) [51]. Since we were not able to find a file associated with the tree, we used TreeSnatcher plus [52] to manually recreate the tree by Nassa et al. (2012) (Supplement 4.1 - 4.6). Afterwards, all edge lengths were adjusted to the longest available edge (COL6a1, 0.40285; Supplement 4.3). The recreated tree by Nassa et al. (2012) was compared with the here constructed phylogenetic trees of all gap open costs and all confidence values using the Environment for Tree Exploration (*ETE*) framework [53]. For comparison, the normalised Robinson-Foulds values were taken and averaged.

Results

Here, we have investigated the interchain hydrogen between different collagen α-chains to build interaction networks. This should help to reveal unknown interactions between different collagen types and make it possible to infer their spatial arrangement, i.e., within the basement membrane. To reduce the complexity of results, we further narrowed down the parameters. The upper bound started with 45 and was increased by five every new run. Starting at confidence 50, the α-chain COL7a1 is filtered out. Then, starting at confidence 60, the α-chains COL9a2, COL9a3 and COL13a1 are removed from the dataset. Naturally, the higher we set the confidence threshold, the more α-chains and thus collagen types were filtered out (Supplement 1.2.3). At the highest confidence threshold of 95, only three out of the 44 α-chains remained, i.e., COL3a1, COL4a1 and COL4a3. In this study, we analysed the alignments for confidence values 45 and 50 for all gap open costs (Figure 3). The remaining alignments, trees, and networks, as well as the rest of the data for the main analysis, can be found in the supplement (Supplement 1). At the extremes of gap opening costs, either no blocks were formed at gap open cost one or saturation occurs at gap open cost six and above. For the above-mentioned reasons, we chose the alignments of gap open cost six and confidence values 45 and 50 for the further analysis.

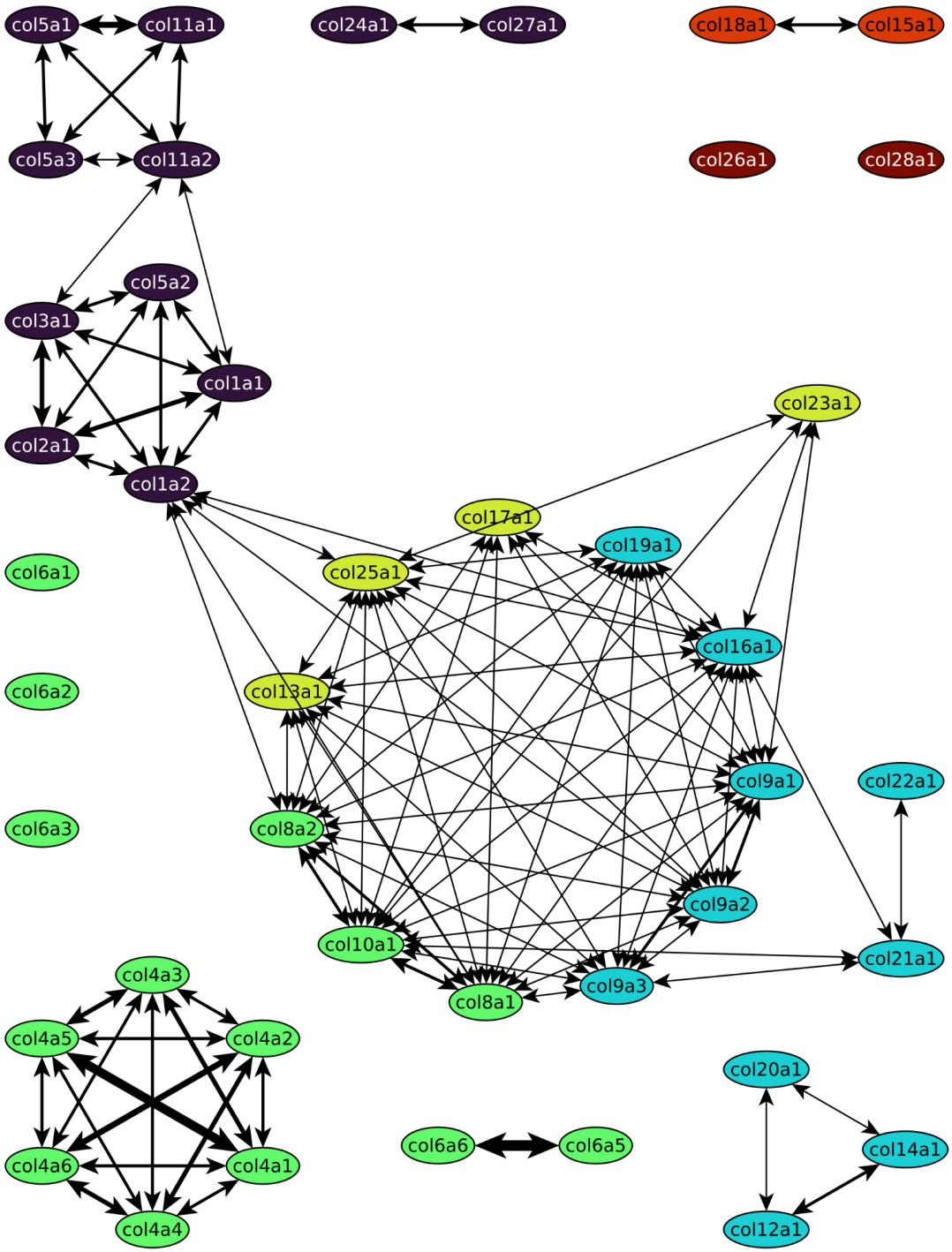

Figure 3. Network representation of the alignment (gap open: 6, confidence: 50, list threshold: 0.9) of the 44 α-chains found in humans. The colours are as follows: Fibrillar collagens: purple, associated with fibrils (FACITs): cyan, network-forming collagens: green, transmembrane collagens: yellow, endostatin precursor collagens: orange, and other collagens: red.

Lists and networks

Our analysis showed that the representation of phylogenetic trees is inadequate

because the parameter settings affected the structure of the trees significantly. Furthermore, in a phylogenetic tree only neighbour relations were shown.

Nevertheless, the phylogenetic trees were necessary for comparison with the literature to eliminate possible sources of error. Listing the possible neighbours for every collagen α-chain as well as representing them as graphs was more suitable to show all bonds of a single α-chain. Throughout the different confidence values and gap open costs, three stable subnetworks were found: (1) called hexagon between the six α-chains of the network-forming collagen COL4; (2) called pentagon between the α-chains COL1a1/2, COL2a1, COL3a1 and COL5a2 of the fibrillar collagens; and (3) called tetragon between the α-chains COL5a1/3 and COL11a1/2 of the fibrillar collagens (Figure 3). A fourth, less stable subnetwork was formed between the α-chains of the FACITs COL12a1, COL14a1 and COL20a1.

Since we considered all collagen α-chains for all tissue types, an adjustment of the threshold was necessary. This means that different thresholds formed different subnetworks for the entire analysis. These subnetworks were able to form functional units, i.e., the basement membrane or more complex organs (tendons). This would need to be adjusted for a specific tissue for the specific collagen α-chains. We focused here on the threshold 0.9. At confidence value 50 for the used threshold of 0.9, the pentagon connected with the tetragon via the α-chains COL1a1 and COL3a1 to the COL11a2 (Figure 3). Additionally, the pentagon connected with a larger unspecified subnetwork through the α-chain COL1a2, more specifically with the two network-forming collagen α-chains COL8a1/2, the two FACITs COL9a2 and COL16a1, and the transmembrane collagen α-chain COL25a1. The α-chain of the FACIT collagen COL16a1 linked the loosely connected neighbouring nodes COL1a2, COL21a1, and COL23a1. The larger subnetwork consisted of three network-forming α-chains, four transmembrane α-chains, and seven FACITs (Figure 3). At a more stringent threshold, the residual network could be further fragmented. At a threshold of 0.7, an hourglass shape was formed between the α-chains of network-forming collagens (COL8a1/2 and COL10a1) and of FACITs (COL9a1/2/3) with COL8a1 as the link (Supplement 1.3.6.: gap open: 6, confidence: 50, list threshold: 0.7). Also at threshold 0.7, five α-chains pairs stood out that merged with other networks at higher thresholds, namely the network-forming collagen α-chains COL6a5 with COL6a6, FACIT α-chains COL12a1 with COL14a1, transmembrane collagen α-chains COL13a1 with COL25a1, endostatin

precursor collagen α-chains COL15a1 with COL18a1, and fibrillar collagen α-chains COL24a1 with COL27a1. At the max threshold of 1.5, most α-chains were merged into a larger subnetwork except for the α-chains of the three collagens COL4, COL15, and COL18. The α-chains of COL4 still formed the hexagon, and COL15 and COL18 still paired with each other, indicating a greater similarity.

Network comparison with PubMed and GoogleScholar

Looking at the reference matrices (Supplement 3.3), the lowest values (and thus the most hits) between collagens (GoogleScholar) or α-chains (PubMed) were on the diagonal, where only single collagens or α-chains were searched for. This is due to the fact that single keywords always have more hits than the conjunction with other keywords. The overall lowest value was for the collagen COL1 with 0.32 (13100 hits) or the α-chain COL1a1 with 0.13 (2264 hits) indicating a high impact in the literature. The second-lowest values were for the cells around the diagonal. These cells mostly represented α-chains of the same or of closely related collagens. For example, the lowest values for the six α-chains of COL4 were between themselves. None of those values were above 0.42 and, therefore, could be found together in the literature. This could indicate that they are closely related or interact with each other. The same correlation could be seen in the networks for the alignments, where the α-chains of COL4 always formed a subnetwork (hexagon) with each other, but never with other α-chains. Similar results could be seen for the fibrillar collagens (pentagon and tetragon). For the pentagon, no α-chain value was above 0.30 except for the α-chain pair COL2a1 - COL5a2 (0.52). For the tetragon, the lowest value was between the α-chain pair COL11a1 - COL11a2 while in the subnetwork it showed a closer relation to COL5a1. Furthermore, in contrast to the subnetwork, the literature showed no (10) or only a minor connection (1.44) between the α-chains of COL11 and the α-chain COL5a3. Therefore, the literature search supported the larger networks, but gave some uncertainty due to bias of studies.

Tree comparison with Nassa et al. (2012)

We compared our phylogenetic trees of the first α-chains of each collagen to that of Nassa et al. (2012). The comparison between the trees showed a difference of 64.67 % on average over all gap open costs and all confidence values compared to the Nassa

et al. (2012) tree (Supplement 4.8). Specifically, gap open 6 and confidence 45 showed a difference of 71.2 % on average. Gap open 6 and confidence 50 showed a difference of 69.2 % on average. Closer inspection of the comparison reveals similar local clusters such as COL1a1, COL2a1, COL3a1, COL5a1, and COL11a1, with only adjacent edges interchanged. In addition, the same cluster containing COL12a1, COL14a1 and COL20a1 was found here and by Nassa et al. (2012). This confirms our findings that in principle a comparison is possible, but due to the complexity of the interactions between collagen α-chains a network approach is more appropriate.

Discussion

Our aim was to reconstruct the interchain hydrogen bonds between the different collagen α-chains based on multiple sequence alignments. Our approach differs from Gordon & Hahn (2010) [15], who classified collagen families based on their function. For reasons of complexity, we focus our analysis on collagens only and exclude other binding proteins. Here, we showed that this classification could be reconstructed again based on the sequences of the α-chains. This supports our hypothesis that the function of the α-chain can be predicted by the sequence itself. When searching the tissue profile databases (https://www.proteinatlas.org/, 2023) [36,38,54], one must distinguish between the collagen protein content and profiles of RNA expression. This is because mass spectroscopy can determine the actual proportion of proteins in a tissue, whereas RNA transcription is only a proxy for protein levels. Since collagens are structural proteins, RNA expression is different between ontogenetic stages and adults because the developmental process is downregulated in adults [55,56]. However, it is to be expected that in tissues with high cell division and, thus, high protein turnover (i.e., skin, hair follicles, mucosa, bone marrow, etc.), RNA expression can reflect protein content.

Fibrillar-forming collagens

In a more detailed analysis of our networks, it can be seen that the fibrillar collagens split into two further clusters. The first fibrillar cluster includes collagens (COL1a1/a2, COL2a1, COL3a1, COL5a2) which are general structure proteins in the body. COL1a1/a2 are present in all tissues except for the eye. The α-chains COL2a1 (eye, connective tissue) and COL3a1 (primary sex organs) are more specific. For COL5a2, no protein expression profile is available, and the RNA expression profile is unspecific [38]. In the second cluster (COL5a1/a3, COL11a1/a2) only COL5a3 has a specific protein expression profile available except for eye and connective tissue. This is in contrast to the expression pattern of COL2a1 which showed expression in those tissues. The RNA expression profiles of these α-chains show high tissue specificity with respect to primary sex organs. The connection between COL5a1 and COL11a1 has a strong bond. Both α-chains are specific for reproductive tissue (cervix and placenta). COL5a1/a2 are associated with several diseases, including Ehlers-Danlos

syndrome, which is characterised by very elastic skin, weakened blood vessels, and joint hypermobility [57,58]. COL24a1 and COL27a1 already form their own network separate from the other fibrillar collagens at low thresholds from the distance matrix and are only incorporated into a large network at high thresholds (Supplement 1.3.6: gap open: 6, confidence: 50, list threshold: 1.3). The triple helices of these two α-chains are shorter than those of the other fibrillar collagens [15]. This may be the reason why they do not interact with other collagen α-chains.

Fibril Associated Collagens with Interrupted Triple helices (FACITs)

Our results show that COL16a1 is a connector between different collagen types, namely fibrillar collagens, transmembrane collagens and FACITs (Figure 3), which also is reflected in its non-specific RNA expression profile [38]. The α-chains of COL9 have a high protein content in the eye and connective tissue and a high RNA expression in the brain and choroid plexus (vascular plexus in the brain ventricles). Defects in the α-chains of COL9 can lead to Stickler syndrome, a disease characterised by ophthalmic, orofacial, articular, and auditory defects [59]. The range of defects highly suggests that the α-chains of COL9 are expressed in the first and second pharyngeal arch derivatives, which include for example the maxilla, mandible, palate, and auditory ossicles [60,61]. For the less stable network between COL12a1, COL14a1 and COL20a1, the former two collagens are present in tissues with either COL1 or COL2 which also explains their strong bond to each other (Gordon & Hahn, 2010). No protein expression was available for COL20a1, but the RNA expression profile showed that it is mainly found in the brain [38], where COL1 is also expressed. However, we could not detect any network neighbours with any α-chains of either COL1 or COL2.

Network-forming collagens

The α-chains of the network-forming collagen COL4 are closely related, which indicates similar functions with few differences. As with the collagens of the second fibrillar cluster, the α-chains of COL4 are also essential structure proteins in the body [62]. For COL4a1/a2/a3 protein expression profiles exist [38]. A defect of the α-chains COL4a3/a4/a5 can lead to Alport syndrome which affects the basement membranes of the kidney, inner ear, and eye [63,64,65,66]. Although part of the bigger subnetwork of FACITs, network-forming collagens and transmembrane collagens, there is a strong

bond between COL8a1/a2 and COL10a1. This can be explained by their high similarity to each other, with COL8a1/a2 having one additional exon compared to COL10a1 [15]. Furthermore, COL8a1 and COL10a1 are associated with age-related macular degeneration [67] whereas mutations in COL8a2 can lead to Fuchs endothelial dystrophy, an impairment of the vision [68]. The α-chains of COL6 do not form any subnetworks of their own, as seen for the α-chains of COL4, except for COL6a5 and COL6a6 where a strong bond can be observed. It is difficult to say why this is the case. Only COL6a1/a2/a3 have known protein content. COL6a2 is expressed throughout most tissues, whereas COL6a1/a3 are mostly present in primary sex organs, muscle tissue and bone marrow [38]. One disease is limb-girdle muscular dystrophy that is caused by a weakness of the dystrophin complex [69]. This weakness is caused by destabilisation of the structural proteins that are supposed to keep the muscular cell intact during contractions. One of these proteins could be COL6a1/a2/a3. For COL6a5/a6, RNA expression profiles show that COL6a5 occurs primarily in lung and skin tissues and COL6a6 is non-specific [38].

Transmembrane collagens

The transmembrane collagens do not show distinct subnetworks on their own (Figure 3). Only COL25a1 has a connection to the pentagon via COL1a2. However, all transmembrane collagens have connections to FACITs and network-forming collagens. COL17a1 is associated with junctional epidermolysis bullosa [70], whereas the literature is insufficient for the other three α-chains.

Endostatin precursor collagens

COL15a1 is structurally homologous to COL18a1 [71] which explains their distinct and stable bond. Looking at the protein content of them, both α-chains are distributed throughout most tissues, although COL18a1 seems to have overall a higher expression [38]. COL18a1 is mostly expressed in the brain and eye, and associated with the Knobloch syndrome, which leads to eye deformations in the development phase, called occipital encephalocele [72]. COL15a1 on the other hand shows its highest expression in muscle and connective tissues [38]. Mouse mutants showed progressive degeneration in skeletal muscles and susceptibility to muscle damage [71]. Other than that, no discernible feature can be extracted from the network.

Collagens without classification

The α-chains of the collagens COL26a1 and COL28a1 show connections to collagens only at high thresholds from the distance matrix. This could indicate non-stable compounds that also bind differently at different confidences. In the Human Protein Atlas, there was an expression pattern only for COL26a1, but not for COL28a1. COL26a1 showed low protein expression in gastrointestinal tract and skin tissues, which both have a high stress environment [38]. Nevertheless, these two α-chains are associated with various diseases, such as aspirin-intolerant asthma or von Willebrand disease [73,74].

Comparison with the literature

Our results regarding the distance matrix are also reflected in the PubMed and GoogleScholar reference matrices. In particular, local clusters (tetragon, pentagon, and hexagon) can be detected. When comparing our trees side-by-side to Nassa et al. (2012), the distances are substantial. However, it should be noted that Nassa and colleagues used the entire sequences of the α-chains, while we analysed partial sequences with high confidence. Common local clusters (i.e., COL12a1, COL14a1, and COL20a1) show up here as well.

Outlook

Since collagens are important structural proteins present in every tissue of multicellular animals, they naturally show serious effects on the body when altered [75]. In this regard, the interchain hydrogen bonds between collagens represent important insights into understanding development and diseases (such as osteogenesis imperfecta). In this context, networks are more interesting for evaluation than phylogenetic trees because they better represent the bonds among collagen α-chains. Our analysis showed that the representation with only collagen types is not sufficient to show the linkages, because the α-chains bind differently within a collagen type. To this end, it is necessary to always examine the α-chains. Moreover, in further studies, it should be possible to verify the overall network and its subnetworks of collagens with immunohistochemistry. Additional analysis could examine the ageing of collagens (in

regard to chemical modifications) and its effects on networks [76]; or also the consideration of further binding proteins between the collagens, such as adhesion proteins like fibronectins, laminins, tenascins, or glycoproteins.

Acknowledgments

We thank all the members of the Department of Bioinformatics that were providing guidance and support during this study. Additionally, we thank A. Berndt for stimulating discussions. This study benefited from research experience of HS gained in studies funded by the Center of Interdisciplinary Prevention of Diseases related to Professional Activities (KIP) funded by the Friedrich-Schiller-University Jena and the 'Berufsgenossenschaft Nahrungsmittel und Gastgewerbe Erfurt (Germany)' (BGN).

Author contributions

All authors contributed to the study conception and design of this manuscript. Data collection and analysis were performed by V.W., L.S. and H.S. The first draft of the manuscript was written by V.W., L.S. and H.S.; S.S. and J.Z-C. contributed extensively with comments and proofreading to all versions of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The datasets generated and analysed during this study along with the code and several supplementary files are available in the Mendeley repository: Wesp, Valentin; Stark, Heiko (2023), “Predicting bonds between collagens – A multiple alignment approach”, Mendeley Data, V1, doi: 10.17632/9pf5n687z4.1

References

1. Huxley-Jones, J., Robertson, D. L. & Boot-Handford, R. P., On the origins of the extracellular matrix in vertebrates. *Matrix Biology* **26**, 2–11 (2007).

2. LeBleu, V. S., MacDonald, B. & Kalluri, R., Structure and function of basement membranes. *Experimental biology and medicine* **232**, 1121–1129 (2007).

3. Patino, M. G., Neiders, M. E., Andreana, S., Noble, B. & Cohen, R. E., Collagen: an overview. *Implant dentistry* **11**, 280–285 (2002).

4. Salamito, M., Nauroy, P. & Ruggiero, F., in *The collagen superfamily and collagenopathies* (Springer, 2021), pp. 1–22.

5. Ahmed, M. S., Kamruzzaman, M., Rana, M. M., Akond, Z. & Mollah, M. N. H., In silico analyses of human collagen protein function prediction. *Journal of Bio-Science* **24**, 55–65 (2016).

6. Bornstein, P. & Sage, H., Structurally distinct collagen types. *Annual review of biochemistry* **49**, 957–1003 (1980).

7. Kucharz, E. J., *The collagens: biochemistry and pathophysiology* (Springer Science & Business Media, 2012).

8. Miller, E. J. & Gay, S., [1] Collagen: An overview. *Methods in enzymology* **82**, 3–32 (1982).

9. Zhao, C., Xiao, Y., Ling, S., Pei, Y. & Ren, J., Structure of collagen. *Fibrous Proteins: Design, Synthesis, and Assembly*, 17–25 (2021).

10. Van Der Rest, M. & Garrone, R., Collagen family of proteins. *The FASEB*

journal **5**, 2814–2823 (1991).

11. Tajbakhsh, S. & Spörle, R., Somite development: constructing the vertebrate body. *Cell* **92**, 9–16 (1998).

12. Duband, J.-L. & Thiery, J. P., Distribution of laminin and collagens during avian neural crest development. *Development* **101**, 461–478 (1987).

13. Leivo, I., Vaheri, A., Timpl, R. & Wartiovaara, J., Appearance and distribution of collagens and laminin in the early mouse embryo. *Developmental biology* **76**, 100–114 (1980).

14. Gonçalves, T. J. M. *et al.*, Collagen XXV promotes myoblast fusion during myogenic differentiation and muscle formation. *Scientific reports* **9**, 5878 (2019).

15. Gordon, M. K. & Hahn, R. A., Collagens. *Cell and tissue research* **339**, 247–257 (2010).

16. Mayne, R. & Brewton, R. G., New members of the collagen superfamily. *Current opinion in cell biology* **5**, 883–890 (1993).

17. Vuorio, E. & De Crombrugghe, B., The family of collagen genes. *Annual review of biochemistry* **59**, 837–872 (1990).

18. Ricard-Blum, S., The collagen family. *Cold Spring Harbor perspectives in biology* **3**, a004978 (2011).

19. Exposito, J.-Y., Valcourt, U., Cluzel, C. & Lethias, C., The fibrillar collagen family. *International journal of molecular sciences* **11**, 407–426 (2010).

20. Franzke, C.-W., Tasanen, K., Schumann, H. & Bruckner-Tuderman, L., Collagenous transmembrane proteins: collagen XVII as a prototype. *Matrix*

biology **22**, 299–309 (2003).

21. Knupp, C. & Squire, J. M., Molecular packing in network-forming collagens. *TheScientificWorldJOURNAL* **3**, 558–577 (2003).

22. Ricard-Blum, S., Ruggiero, F. & van der Rest, M., The collagen superfamily. *Collagen: Primer in Structure, Processing and Assembly*, 35–84 (2005).

23. Francomano, C. A., Key role for a minor collagen. *Nature genetics* **9**, 6–8 (1995).

24. Gelse, K., Pöschl, E. & Aigner, T., Collagens—structure, function, and biosynthesis. *Advanced drug delivery reviews* **55**, 1531–1546 (2003).

25. Hulmes, D. J. S., in *Collagen: structure and mechanics* (Springer, 2008), pp. 15–47.

26. Linsenmayer, T. F., in *Cell Biology of Extracellular Matrix: Second Edition* (Springer, 1991), pp. 7–44.

27. Connizzo, B. K., Yannascoli, S. M. & Soslowsky, L. J., Structure–function relationships of postnatal tendon development: A parallel to healing. *Matrix Biology* **32**, 106–116 (2013).

28. McAlinden, A. *et al.*, α-Helical coiled-coil oligomerization domains are almost ubiquitous in the collagen superfamily. *Journal of Biological Chemistry* **278**, 42200–42207 (2003).

29. Birk, D. E. & Bruckner, P., Collagen suprastructures. *Collagen: primer in structure, processing and assembly*, 185–205 (2005).

30. Fratzl, P., in *Collagen: structure and mechanics* (Springer, 2008), pp. 1–13.

31. Goh, K. L., Listrat, A. & Béchet, D., Hierarchical mechanics of connective tissues: Integrating insights from nano to macroscopic studies. *Journal of biomedical nanotechnology* **10**, 2464–2507 (2014).

32. Heino, J., The collagen family members as cell adhesion proteins. *Bioessays* **29**, 1001–1010 (2007).

33. Nimni, M. E. & Harkness, R. D., in *Collagen* (CRC Press, 2018), pp. 1–78.

34. Brodsky, B. & Persikov, A. V., Molecular structure of the collagen triple helix. *Advances in protein chemistry* **70**, 301–339 (2005).

35. Kim, M.-S. *et al.*, A draft map of the human proteome. *Nature* **509**, 575–581 (2014).

36. Uhlén, M. *et al.*, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

37. Oh, S. P., May Griffith, C., Hay, E. D. & Olsen, B. R., Tissue-specific expression of type XII collagen during mouse embryonic development. *Developmental dynamics* **196**, 37–46 (1993).

38. Pontén, F., Jirström, K. & Uhlen, M., The Human Protein Atlas—a tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **216**, 387–393 (2008).

39. Jumper, J. *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

40. Varadi, M. *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **50**, D439–D444 (2022).

41. Marro, J., Pfefferli, C., de Preux Charles, A.-S., Bise, T. & Jaźwińska, A., Collagen XII contributes to epicardial and connective tissues in the zebrafish heart during ontogenesis and regeneration. *PLoS One* **11**, e0165497 (2016).

42. Mereness, J. A. & Mariani, T. J., The critical role of collagen VI in lung development and chronic lung disease. *Matrix Biology Plus* **10**, 100058 (2021).

43. Brown, J. C. & Timpl, R., The collagen superfamily. *International archives of allergy and immunology* **107**, 484–490 (1995).

44. Beck, K. & Brodsky, B., Supercoiled protein motifs: the collagen triple-helix and the α-helical coiled coil. *Journal of structural biology* **122**, 17–29 (1998).

45. Kapli, P., Yang, Z. & Telford, M. J., Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* **21**, 428–444 (2020).

46. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J., Ancestral sequence reconstruction for protein engineers. *Current opinion in structural biology* **69**, 131–141 (2021).

47. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**, 3059–3066 (2002).

48. Reguant, R. *et al.*, AlignmentViewer: sequence analysis of large protein families. *F1000Research* **9** (2020).

49. Minh, B. Q. *et al.*, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution* **37**, 1530–1534 (2020).

50. Letunic, I. & Bork, P., Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research* **47**, W256–W259 (2019).

51. Nassa, M. *et al.*, Analysis of human collagen sequences. *Bioinformation* **8**, 26 (2012).

52. Laubach, T., von Haeseler, A. & Lercher, M. J., TreeSnatcher plus: capturing phylogenetic trees from images. *BMC bioinformatics* **13**, 1–5 (2012).

53. Huerta-Cepas, J., Serra, F. & Bork, P., ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution* **33**, 1635–1638 (2016).

54. Karlsson, M. *et al.*, A single–cell type transcriptomics map of human tissues. *Science advances* **7**, eabh2169 (2021).

55. Tonelotto, V. *et al.*, Spatio-temporal expression and distribution of collagen VI during zebrafish development. *Scientific Reports* **9**, 19851 (2019).

56. Imamura, Y., Scott, I. C. & Greenspan, D. S., The Pro-α3 (V) Collagen Chain: complete primary structure, expression domains in adult and developing tissues, and comparison to the structures and expression domains of the other types V and XI procollagen chains. *Journal of Biological Chemistry* **275**, 8749–8759 (2000).

57. Steinmann, B., Royce, P. M. & Superti-Furga, A., The Ehlers-Danlos syndrome. *Connective tissue and its heritable disorders: molecular, genetic, and medical aspects*, 431–523 (2002).

58. Malfait, F., Wenstrup, R. J. & De Paepe, A., Clinical and genetic aspects of Ehlers-Danlos syndrome, classic type. *Genetics in medicine* **12**, 597–605

(2010).

59. Robin, N. H., Moran, R. T. & Ala-Kokko, L., webpage on the Internet. *Stickler Syndrome*, 1993–2015 (2000).

60. Trainor, P. A. & Krumlauf, R., Hox genes, neural crest cells and branchial arch patterning. *Current opinion in cell biology* **13**, 698–705 (2001).

61. Liu, Q., Gibson, M. P., Sun, H. & Qin, C., Dentin sialophosphoprotein (DSPP) plays an essential role in the postnatal development and maintenance of mouse mandibular condylar cartilage. *Journal of Histochemistry & Cytochemistry* **61**, 749–758 (2013).

62. Khoshnoodi, J., Pedchenko, V. & Hudson, B. G., Mammalian collagen IV. *Microscopy research and technique* **71**, 357–370 (2008).

63. De Gregorio, V., Caparali, B., Shojaei, A., Ricardo, S. & Barua, M., Alport Syndrome: Clinical Spectrum and Therapeutic Advances. *Kidney Medicine*, 100631 (2023).

64. Imafuku, A. *et al.*, Autosomal dominant form of type IV collagen nephropathy exists among patients with hereditary nephritis difficult to diagnose clinicopathologically. *Nephrology* **23**, 940–947 (2018).

65. Shulman, C. *et al.*, Type IV collagen variants in CKD: Performance of computational predictions for identifying pathogenic variants. *Kidney Medicine* **3**, 257–266 (2021).

66. Deltas, C., POS-435 NEXT GENERATION SEQUENCING IDENTIFIES CANDIDATE GENETIC MODIFIERS POTENTIALLY EXACERBATING KIDNEY DISEASE IN COL4A3/A4 HETEROZYGOUS PATIENTS. *Kidney International Reports* **7**, S194 (2022).

67. Cascella, R. *et al.*, Towards the application of precision medicine in Age-Related Macular Degeneration. *Progress in Retinal and Eye Research* **63**, 132–146 (2018).

68. Zhang, J. & Patel, D. V., The pathophysiology of Fuchs' endothelial dystrophy–a review of molecular and cellular insights. *Experimental eye research* **130**, 97–105 (2015).

69. Bushby, K. M. D., Collins, J. & Hicks, D., Collagen type VI myopathies. *Progress in Heritable Soft Connective Tissue Diseases*, 185–199 (2014).

70. Pulkkinen, L. & Uitto, J., Mutation analysis and molecular genetics of epidermolysis bullosa. *Matrix biology* **18**, 29–42 (1999).

71. Marneros, A. G. & Olsen, B. R., The role of collagen-derived proteolytic fragments in angiogenesis. *Matrix Biology* **20**, 337–345 (2001).

72. Seppinen, L. & Pihlajaniemi, T., The multiple functions of collagen XVIII in development and disease. *Matrix Biology* **30**, 83–92 (2011).

73. Pasaje, C. F. A. *et al.*, A possible association of EMID2 polymorphisms with aspirin hypersensitivity in asthma. *Immunogenetics* **63**, 13–21 (2011).

74. Veit, G. *et al.*, Collagen XXVIII, a novel von Willebrand factor A domain-containing protein with many imperfections in the collagenous domain. *Journal of Biological Chemistry* **281**, 3494–3504 (2006).

75. Iozzo, R. V. & Gubbiotti, M. A., Extracellular matrix: the driving force of mammalian diseases. *Matrix Biology* **71**, 1–9 (2018).

76. Fichtner, M., Schuster, S. & Stark, H., Determination of scoring functions for protein damage susceptibility. *Biosystems* **187**, 104035 (2020).