Predicting bonds between collagens – An 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 types of collagens, some of which differ significantly in form, function, and tissue specificity. Subdivisions into so-called collagen families exist, which are defined based on mainly clinical research. Collagens contain polypeptide strands (α-chains). Their sequences are often analysed because 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, clusters with high interconnectivity can be found, such as the fibrillar collagens, the COL4 network-forming collagens, and the COL9 FACITs. Furthermore, a large cluster between network-forming, FACIT, and COL28a1 α-chains is formed with COL6a3 as a major hub node.

Keywords

α-chains, AlphaFold, Collagen, Data Matrix, High Confidence Region, Network, Sequence Alignment

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 (Huxley-Jones et al., 2007; LeBleu et al., 2007; Patino et al., 2002; Salamito et al., 2021). Collagens are found in skin, connective tissue, vascular walls, bone, and cartilage (Ahmed et al., 2018; Bornstein & Sage, 1980; Kucharz, 1992; Miller & Gay, 1982; Zhao et al., 2021). Moreover, the expression of collagens is essential for the formation of the basic structure of tissues during ontogenesis (Van Der Rest & Garrone, 1991). For example, a somite is a block of condensed mesoderm formed bilaterally along the central axis in vertebrate embryos (Tajbakhsh & Spörle, 1998). 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 (Duband & Thiery, 1987; Leivo et al., 1980). The fusion of myoblasts into multinucleated myofibers is another essential task of one of the collagens (Gonçalves et al., 2019).

Many types of collagens 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 (Table 1; Gordon & Hahn, 2010; Mayne & Brewton, 1993; Mayne & Burgeson, 1987; Vuorio & De Crombrugghe, 1990). A classification is realised by the characteristic properties of the collagen such as its structure, interaction, or site of expression (Ricard-Blum, 2010). Due to the variability of characters, various classifications of collagen exist, with none of them being indisputable (Exposito et al., 2010; Franzke et al., 2003; Knupp & Squire, 2003; Ricard-Blum et al., 2005); 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**

These 28 Collagens are further differentiated into 44 α-chains. In consideration of the collagen groups in Table 1, these α-chains are characterised by specific post-translational modifications (PTM) that are connected by a repetitive sequence (Brodsky & Persikov, 2005), called a chain. This chain is predominantly involved in structure formation. The first PTM, which appears in all α-chains except for the transmembrane ones, is the signal sequence that is important for the translocation of the α-chain to the endoplasmic reticulum that is afterward removed (Connizzo et al., 2013; McAlinden et al., 2003). In nine fibrillar and two network-forming α-chains (COL1a1, COL1a2, COL2a1, COL3a1, COL4a1, COL4a2, COL5a1, COL5a2, COL11a1, COL11a2, COL27a1), a propeptide PTM is located next to the signal PTM and/or at the end of all fibrillar α-chains (The Uniprot Consortium, 2023). These propeptides are essential for the assembly of collagen trimers and are cleaved in fibrillar collagens after helix formation (Exposito et al., 2010). Most of the collagens are homotrimers, i.e., they are built from three identical α-chains (Francomano, 1995; Brinckmann, 2005; Ricard-Blum, 2010; Gelse et al., 2003, Hulmes, 2008; Linsenmayer, 1991). However, some heterotrimers consist of two identical α-chains. The resulting trimers aggregate with collagen microfibrils in the environment, which are further organised as fibrils (Birk & Bruckner, 2005; Fratzl, 2008; Goh et al., 2014). These fibrils are further packed into fibres, which are bundled into fascicles (Heino, 2007; Nimni & Harkness, 2018).

Another characterisation of collagens is based on the abundance of their collagen types and mRNA expression throughout the human body (Kim et al., 2014; Uhlén et

al., 2015). The mRNA localization of individual collagens is often tissue-specific (Oh et al., 1993; Kim et al., 2014).
The understanding of the sequence and structure of collagens is crucial because of hereditary diseases based on mutations in collagen genes, such as osteogenesis imperfecta, Ehlers-Danlos syndrome, and Stickler syndrome.

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). It uses a deep neural network to predict spatial structures for proteins, similar to a graph interference problem (Jumper et al., 2021; Varadi et al., 2022). 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 concerning 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 characterised by the motif $(Gly\text{-}Pro\text{-}Pro)_n$ and are found mainly in the inside of the sequences, as they are instrumental in the structure of the collagen (Brown & Timpl, 1995; Beck & Brodsky, 1998). However, they are disruptive in analysis due to their repetitions, as they occur more frequently than the functional regions of the protein. This is in contrast to boundary regions, as they return a high confidence value in most collagens, which indicates known knowledge about the 3D structure of other proteins.

The aim of this study

Surprisingly, there are only a few complete studies regarding the possible alignment of collagens to each other (Nassa et al., 2012). Their importance for ontogenesis and structure formation in the body is indisputable (Marro et al., 2016; Mereness & Mariani, 2021). In this study, the 44 α-chains of all 28 available human collagens from all six families are analysed by bioinformatics techniques.

One question is why the regions of high confidence are mostly restricted to the beginning and end of α-chain sequences. A second question is whether there is a correlation between the high confidence and a possible alignment between different α-chains and their collagens. This study aims to investigate these two questions by aligning these short high-confidence regions based on their hydrophobicity.

Material & Methods

Sequence preparation

The sequence information of human collagen variants is retrieved from the AlphaFold DB based on AlphaFold with the version from the "14th Critical Assessment of Protein Structure Prediction" (Fig. 1; Jumper et al., 2021; Varadi et al., 2022, v2.0). In AlphaFold, over 200 million protein structure predictions are stored and freely available.

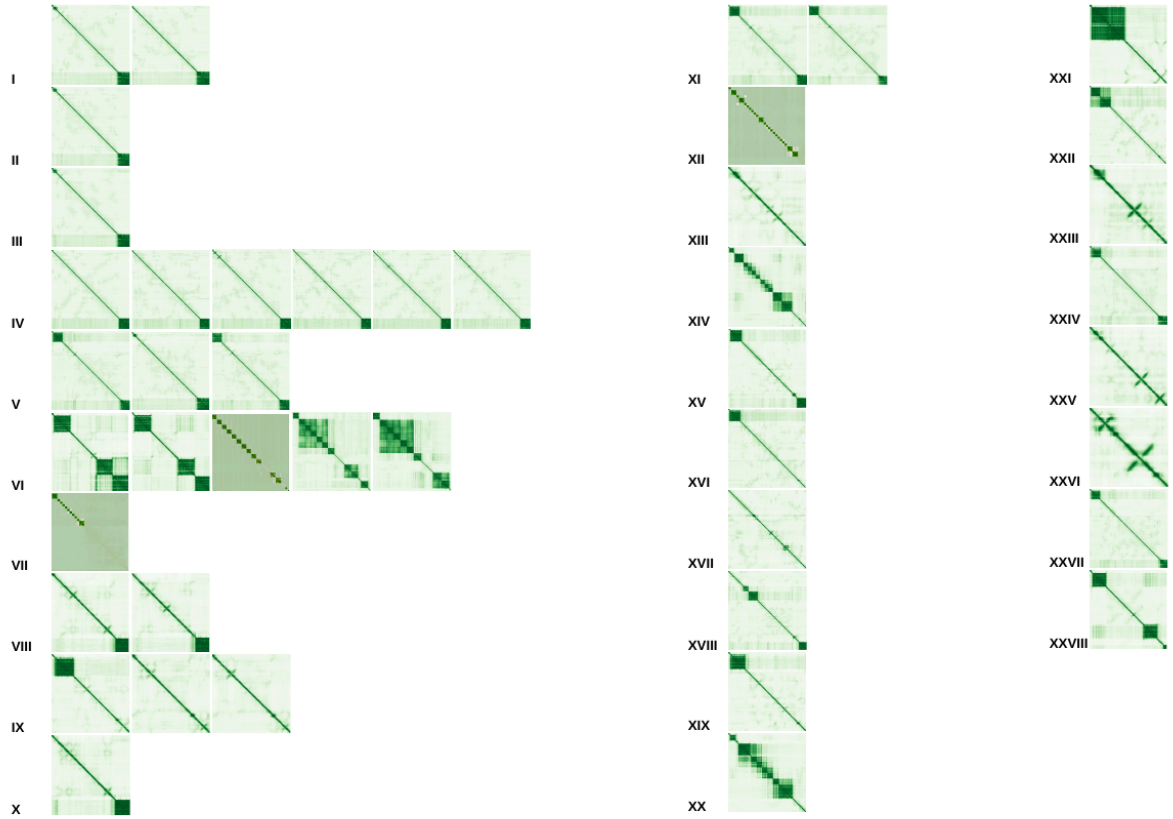

Figure 1: Overview of the predicted aligned confidence of the collagen α-chains of collagens I-XXVIII adapted from AlphaFold (Jumper et al., 2021; Varadi et al., 2022). The colour indicates the expected position error when the predicted and true structures are aligned. The darker the colour of green, the lower the alignment error. The dark-shaded boxes were predicted with a local ColabAlphaFold instance.

The α-chains are selected based on their review state. If this is not available, the reference proteome state is favoured. The sequences and their position confidence

values of the collagen α-chains are downloaded in the form of PDB files. An exception is made for Col6a3, COL7a1, and Col12a1 for which longer sequences are available in the UniProt database (The Uniprot Consortium, 2023). For those α-chains, we download the sequences in FASTA format and predict their confidence values with a local ColabAlphaFold instance (v.1.5.2, https://github.com/YoshitakaMo/localcolabfold). Additionally, we retrieve the PTMs for each α-chain. The IDs and PTM position of the chosen α-chains can be found in Supplement 1.

However, the PDB format is atypical for most sequence analysis tools and is, therefore, reformatted into a FASTA format based on the confidence values of each position. To process the following steps, several custom python (v3.10.12) scripts are implemented (Supp. 7). First, all amino acids are 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 (based on their average confidence value) from each α-chain are extracted (Fig. 2). This is done separately for different subsequence lengths (from ten AA to 500 AA in steps of ten) and confidence thresholds (from zero to 95 in steps of five). 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 are set to -1×10^{9}. Furthermore, for each subsequence, the reversed complement is computed (marked with “rev” in the sequence name). This is done for three conditions based on the PTM positions of each α-chain: *whole sequence* (entire α-chain sequence), *without signal* (the signal sequence is removed from the α-chain sequence), and *only chain* (the signal and propeptide sequences are removed from the α-chain sequence).

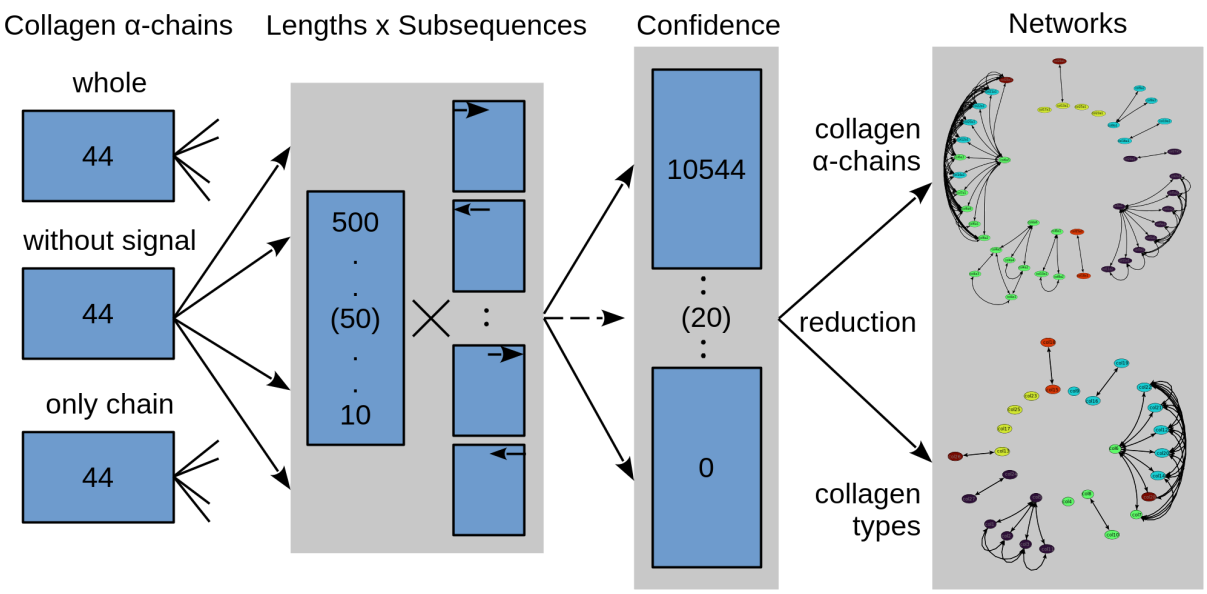

Figure 2: Flowchart of data processing for the 44 collagen α-chains. From left to right, we have collected 44 collagen α-chains with three conditions: *whole sequence* (entire α-chain sequence), *without signal* (the signal sequence is removed from the α-chain sequence), and *only chain* (the signal and propeptide sequences are removed from the α-chain sequence). From these, subsequences are extracted based on lengths (between ten and 500 in steps of ten) and confidence values (between zero and 95 in steps of five). Afterward, for each set, each sequence was globally aligned to each other and matrices were constructed based on the calculated similarities. In the end, all similarity matrices were visualised in the form of networks. Each node was colourized based on the colours in Table 1.

Construction of similarity matrices

To create the distance similarity matrices on which we base our networks, each subsequence in a set is globally aligned to all other subsequences based on their hydrophobicity. The substitution matrix is based on the Eisenberg consensus scale, which combines several hydrophobicity features into one score (Supp. 2). The difference in hydrophobicity for each amino acid pair is taken as the substitution rate and normalised between zero and 20. The rate values are rounded to the next integer. Gap open cost is set to ten and the gap extend cost to 0.5 (EMBOSS default). Additionally, to compare each set with the other, the calculated similarities are also normalised to the length of the aligned subsequences (resulting in a maximum score

of 20 if two identical sequences are compared).

Compression of similarity matrices for further analysing

Afterward, data compressions on the similarity matrices are performed to better analyse our results. The subsequences are assigned to one of two categories, one containing all α-chains and one containing only the collagen types. For the α-chains (44), the maximum similarity between two α-chains over all their subsequence comparisons is selected. For the collagen types (28), the maximum similarity between the subsequences of all α-chains of the compared collagen types is selected. In a further step, lists are generated from these two compressed datasets listing the number of pairings between each α-chain subset and each collagen subset. For these similarities, score thresholds are set between 12 and 20 in steps of 0.25. This is done for each subsequence length and confidence value.
For each reduced similarity matrix, a network (in GML format) is built with the α-chains/collagens as nodes. The similarity scores are used as edges between each α-chain and based on the score thresholds using a custom PASCAL program (v1.1.24). The networks are visualised with the general-purpose diagramming program yEd (v3.23.1).

Generation of literature matrix for validation

To extrapolate the connection between the different α-chains, we construct a reference matrix between all α-chains based on reference databases. The idea behind this is that even if only neighbour links are considered, a larger network should still result. This is done for Google Scholar (entire text) and PubMed (only title and abstract) databases. For this purpose, a search query is performed between two collagens (Google Scholar) or α-chains (Pubmed) for all collagens and α-chains in humans. The number of hits is stored in a matrix. To avoid bias by focusing on specific collagen types in the literature, these values are logarithmized (to the base 100) and normalised between zero and 20. AS with the alignment networks, different score thresholds for the networks were used (between one and 20 in steps of one).

Results

Here, we have investigated the possible interactions between different collagen α-chains to build networks. This should help reveal unknown interactions between different collagen types and make it possible to infer their spatial arrangement through simulations, i.e., within the basement membrane.

Identification of relevant parameters

Overall, we calculated 192,000 different networks (three sequence types [*whole sequence*, *without signal*, *only chain*], 50 subsequence lengths, 20 confidence thresholds, 32 network thresholds, and two experiments [α-chains, collagens]). Therefore, to reduce the complexity of the results, we further narrowed the parameters. According to AlphaFold, a confidence threshold of at least 70 is required to have a "good backbone prediction". In the case of the *whole sequence* and *without signal* sequence types, the longest possible subsequence length for confidence 70, for which all α-chains (at least one sequence) are present, is at length 30. Additionally, we get the same parameters when we calculate an optimum over the entire set of matrices (50 subsequences and 20 confidence thresholds) between the number of subsequences and the standard deviation of the similarity scores for the *whole sequence* and *without signal* datasets. For simplicity' sake, we take the same parameters for the *only chain* dataset. The high-confidence sequences and results for the *whole sequence*, *without signal* and *only chain* datasets, can be found in Supplements 3, 4, and 5 (in that order). Table 2 shows the three sequence types at which confidence thresholds α-chains are filtered out. For the *only chain* sequence type, even at low subsequence lengths of ten and a confidence threshold of 70 most fibrillar α-chains (COL1a1, COL1a2, COL2a1, COL3a1, COL11a1, COL27a1) are filtered out.

Table 2: Summary of the excluded α-chains based on subsequence lengths and confidence thresholds. The first line shows all α-chains missing for the next highest subsequence length (40) and the third line for the next highest confidence threshold (75). The second line shows the α-chains missing for our chosen parameters.

	whole sequence	***without signal***	***only chain***
Subseq. length 40 /confidence 70	COL9a2, COL9a3	COL9a2, COL9a3	COL1a1, COL1a2, COL2a1, COL3a1, COL9a2, COL9a3, COL11a1, COL27a1
Subseq. length 30 /confidence 70			COL1a1, COL1a2, COL2a1, COL3a1, COL11a1, COL27a1
Subseq. length 30 /confidence 75	COL9a2, COL9a3	COL9a2, COL9a3	COL1a1, COL1a2, COL2a1, COL3a1, COL9a2, COL9a3, COL11a1, COL27a1

The remaining score matrices and networks, as well as the rest of the data for the main analysis, can be found on Mendeley Data (doi: 10.17632/9pf5n687z4.2). For the reasons mentioned above, we chose the subsequence length of 30 and the confidence threshold of 70 for further analysis.

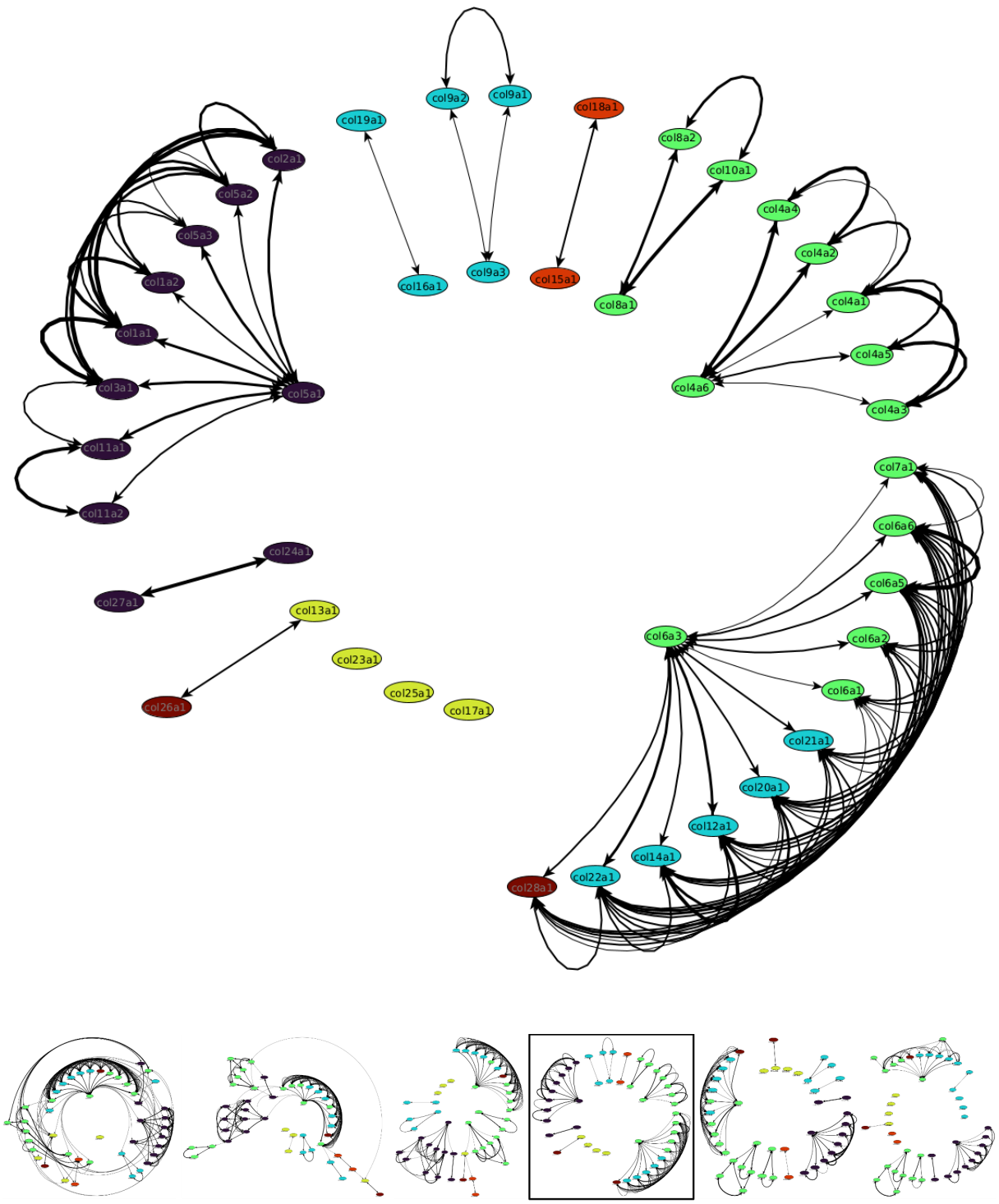

Figure 3: Above: Network representation of the similarity matrix (sequence type: *without signal*, subsequence length: 30, confidence: 70, network threshold: 17.5) of the 44 α-chains found in humans. Below: From left to right, the network presentation of the similarity matrix for the network thresholds 16.75, 17.0, 17.25, 17.5 (selected, see above), 17.75, and 18.0. 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

With the selected subsequence length and confidence threshold, a network is obtained in which α-chain/collagen is connected, and only the thickness of the edges indicates the similarity. For clarity, we further filtered out thin edges (comparatively low similarity) using various threshold values (from 12 to 20 in steps 0.25) to emphasise local subnetworks. From these, the first (lowest threshold) network with the most subnetworks was chosen. In all three sequence types, it was the network with the threshold of 17.5 (*whole sequence* and *without signal*: nine, *only chain*: eight). As written above, for the network of the *only chain* type most fibrillar α-chains are filtered out due to their repetitive chain sequences, leading to poor confidence and similarity values. This is not the case for the other two sequence types. The difference between the *whole sequence* and *without signal* datasets are negligible. Therefore, we focus on the network of the *without signal* type.

As can be seen in Fig. 3, nine subnetworks are formed: (1) the fibrillar subnetwork (consisting of all fibrillar α-chains except for COL24a1 and COL27a1), (2) the COL4 subnetwork (consisting of all six COL4 alpha chains), (3) the COL9 subnetwork (consisting of all three COL9 α-chains), (4) the COL8/10 subnetwork (consisting of the two α-chains of COL8 and the COL10a1 α-chain), (5-8) four pair subnetworks (COL16a1 and COL19a1, COL15a1 and COL18a1, COL13a1 and COL26a1, COL24a1 and COL27a1), three unconnected transmembrane α-chains (COL17a1, COL23a1 and COL25a1), and (9) the largest subnetwork consisting of the remaining α-chains.

To show the interconnection between the individual subnetworks, it is necessary to lower the threshold. With this, the following new connections are formed (Fig. 3): The pair subnetworks are incorporated into the larger subnetworks. COL24/27 is connected to the fibrillar subnetwork through COL1a2, COL13/26 to the COL4 subnetwork through COL4a4, and COL15/18 to COL13/26 through COL13a1. At even lower thresholds, almost all subnetworks are connected. The fibrillar network is connected to the COL4 subnetwork through COL4a3 and COL4a4, the COL8/10 subnetwork through COL8a2, and to the largest subnetwork through COL6a5 and

COL7a1. The subnetworks COL15/18 and COL16/19 now connect to the largest subnetwork through COL6a6. A new pair subnetwork between COL23a1 and COL25a1 is formed. Interestingly, the COL9 subnetwork is very stable.
With even lower thresholds, the networks become multidimensional and are difficult to visualise.

Network comparison with PubMed and Google Scholar

Looking at the reference matrices (Supp. 6), the lowest values (and thus the most hits) between collagens (Google Scholar) or α-chains (PubMed) were on the diagonal, where only single collagens or α-chains were searched for. This is because 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 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 with each other, but only with other α-chains at lower thresholds. Similar results could be seen for the fibrillar collagens. In general, the values were around 0.5 or lower. However, exceptions exist due to the few literature available for the specific α-chains. For example, COL5a3 has values above one or even a value of ten (meaning an occurrence of zero) with COL1a2. The same can also be seen for COL24a1 and COL27a1. The lowest value was at 0.72 between COL27a1 and COL1/COL5a1 while most other values were at ten. At the same time, our network shows the same pattern by separating the COL24/27a1 pair from the large fibrillar network at higher thresholds. In contrast to the alignment analysis, the PubMed networks tend to form a single large cluster, even at high thresholds (Fig. 4). Further analyses were only done for PubMed, as the query limit for Google Scholar was exceeded for α-chains.

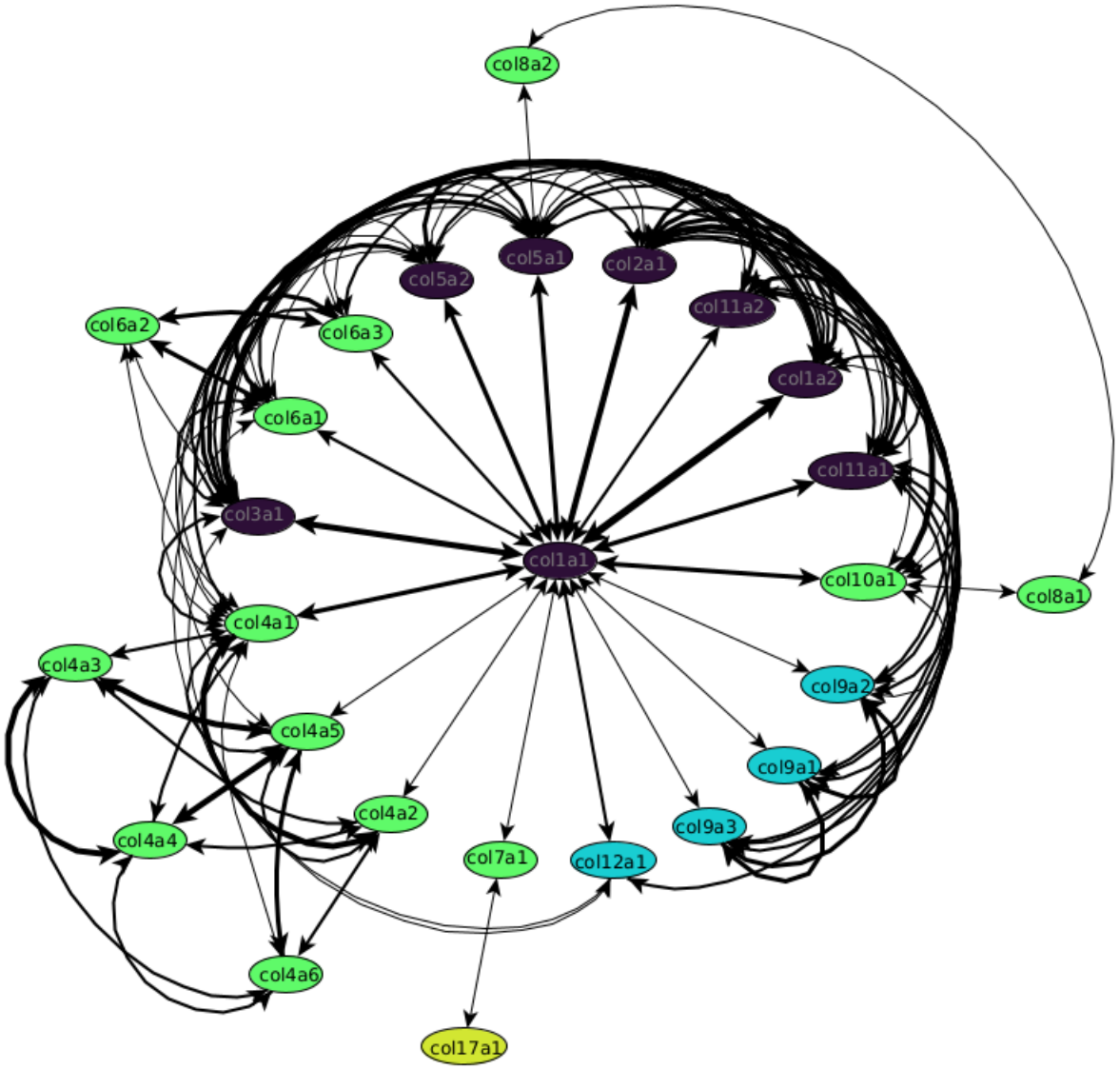

Figure 4: Network representation of the reference matrix of the 26 α-chains with significant entries found in PubMed. The other 18 α-chains had no connections and were not shown. The colours are as follows: Fibrillar collagens: purple, associated with fibrils (FACITs): cyan, network-forming collagens: green, transmembrane collagens: yellow, endostatin precursor collagens: orange, and other collagens: red.

Discussion

We aimed to reconstruct the network between the different collagen α-chains based on pairwise sequence alignments of high-confidence regions. Our approach differs from Gordon & Hahn (2010), 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 generally reconstructed again based on the sequences of the α-chains. Overall, the difference between the *whole sequence* and *without signal* datasets is negligible, since the signal sequence is extremely short. In the following, we will discuss the individual clusters for the *without signal* and *only chain* datasets.

Fibrillar-forming collagen clusters

What is striking is that the propeptides are essential for this analysis. Without them, fibrillar collagens are filtered out due to the poor confidence values of the chain sequences.

In a more detailed analysis of our networks, it can be seen that the fibrillar collagens form two distinct clusters in the *without signal* dataset. Interestingly, COL5a1 is the major hub node in the larger cluster, which can be explained by its role as a regulator for the formation of a uniformly small corneal fibril diameter with COL1 (Mak et al., 2016). COL5a1/a2 are also associated with several diseases, including Ehlers-Danlos syndrome, which is characterised by very elastic skin, weakened blood vessels, and joint hypermobility (Imamura et al., 2000; Steinmann et al., 2002; Malfait et al., 2010). COL24a1 and COL27a1 form their network separate from the other fibrillar collagens starting at a threshold of 17.5 (Fig. 3). The triple helices of these two α-chains are shorter than those of the other fibrillar collagens (Gordon & Hahn, 2010). This may be why they do not interact with other collagen α-chains at high thresholds.

Network-forming collagen clusters

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 fibrillar cluster, the α-chains of COL4 are also essential structure proteins in the body

(Khoshnoodi et al., 2008). At a lower threshold of 17.0, a connection between COL4 and COL7 can be seen as mentioned in the literature (Roig-Rosello and Rousselle, 2020; Brittingham et al., 2006). 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 (Gregorio et al., 2023; Imafuku et al., 2018; Shulman et al., 2021; Deltas, 2022).
The second cluster is formed by COL8a1, COL8a2 and COL10a2. This can be explained by their high similarity to each other, with COL8a1/a2 having one additional exon compared to COL10a1 (Gordon & Hahn, 2010). Furthermore, COL8a1 and COL10a1 are associated with age-related macular degeneration (Cascella et al., 2018) whereas mutations in COL8a2 can lead to Fuchs endothelial dystrophy, an impairment of vision (Zhang & Patel, 2015). The remaining network-forming collagens are part of the largest network, where COL6a3 is a major hub node.

Largest cluster

The largest cluster is not homogeneous like the previous two clusters, consisting of three different types of collagens (network-forming, FACITs, and COL28a1). All α-chains are highly connected to each other. As mentioned before, COL6a3 is a major hub node in this cluster. Generally, COL6 forms the most connections between other collagens. At lower network thresholds, the α-chains of COL6 seem to be a major hub node for the whole network (Fig. 3). This could be explained by the fact that COL6 is important for the function and stability of the cell membrane by interacting with other collagens such as COL1, COL2, COL4 and COL14 (Cescon et al., 2015; Tonelotto et al., 2019). Such connections can also be seen in our network for different thresholds, at lower thresholds for COL1/2 (17.0) and COL4 (16.75) and for COL14 at even higher thresholds of 18.0 (Supp. 4).
This network seems to contain important collagens for developmental biology. One disease associated with this is limb-girdle muscular dystrophy, which is caused by a weakness of the dystrophin complex (Bushby et al., 2014). 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 those of COL6.

COL9 cluster

All three COL9 α-chains are relatively short (the longest chain, COL9a1, is only 921 AA long) and highly similar to each other. Known is that defects in these α-chains can lead to Stickler syndrome, a disease characterised by ophthalmic, orofacial, articular, and auditory defects (Robin et al., 2000). 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 (Trainor & Krumlauf, 2001, Liu et al., 2013). This could also explain why the α-chains of COL9 seem to only connect to other α-chains at low network thresholds (Fig. 3).

Pair clusters

One cluster consisting of only two α-chains is COL15/18. COL15a1 is structurally homologous to COL18a1 (Marneros & Olsen, 2001) which explains their distinct and stable bond. 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 (Seppinen & Pihlajaniemi, 2011). Mouse mutants deficient in COL15a1 showed progressive degeneration in skeletal muscles and susceptibility to muscle damage (Marneros & Olsen, 2001). Other than that, no discernable feature can be extracted from the network.
For the other two pair clusters COL13/26 and COL16/19 the literature search shows no entries concerning interactions. Therefore, no information could be extracted.

Comparison with the literature

Our results regarding the similarity matrix are also reflected in the PubMed reference matrices. In particular, local clusters can be detected. COL1a1 forms the main hub node, with the remaining fibrillar collagens having the strongest connection to each other. At the same, the COL4 and to a lesser degree the COL6, COL8/10 and COL9 clusters also emerge as subclusters in the network. However, it should be noted that this is not surprising, since α-chains and collagens of the same type are usually studied together. The same can be said for COL1a1 functioning as the main hub

node. COL1 is one of the most abundant structural proteins in vertebrates and, therefore, often used in conjunction with other α-chains in studies (Stover and Verrelli, 2010). Lastly, it should be noted that the number of references decreases considerably at higher collagen numbers. For the α-chains not shown, future studies will result in higher literature data and may remedy this problem.
Overall, the resulting reference network was as expected, but could still be used to validate parts of our alignment network.

Outlook

Since collagens are important structural proteins present in every tissue of multicellular animals, they naturally show serious effects on the body when altered (Iozzo & Gubbiotti, 2018). 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 (regarding chemical modifications) and its effects on networks (Fichtner et al., 2020); or also the consideration of further binding proteins between the collagens, such as adhesion proteins like fibronectins, laminins, tenascins, or glycoproteins, or in the far future for the whole human proteome.

Acknowledgments

We thank all the members of the Department of Bioinformatics who were providing guidance and support during this study. Additionally, we thank A. Berndt for stimulating discussions. This study benefited from the 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

V.W., L.S., and H.S. conceived the study. V.W. and H.S. supervised and performed the calculations. V.W. and H.S. analysed network data. V.H. and H.S. drafted the manuscript and all figures. All authors contributed to the interpretation of the results and revised the manuscript.

Competing interests

The author(s) declare no competing interests.

References

Ahmed, M., Kamruzzaman, M., Rana, M., Akond, Z., & Mollah, M. (2018). In silico analyses of human collagen protein function prediction. Journal of Bio-Science, 24(July), 55–65. https://doi.org/10.3329/jbs.v24i0.37487

Beck, K., & Brodsky, B. (1998). Supercoiled protein motifs: The collagen triple-helix and the α- helical coiled coil. Journal of Structural Biology, 122(1–2), 17–29. https://doi.org/10.1006/jsbi.1998.3965

Birk, D. E., & Bruckner, P. (2005). Collagen suprastructures. Topics in Current Chemistry, 247, 185–205. https://doi.org/10.1007/b103823

Bornstein, P., & Sage, H. (1980). Structurally distinct collagen types. Annual Review of Biochemistry, 49, 957–1003. https://doi.org/10.1146/annurev.bi.49.070180.004521

Brinckmann, J. (2005). Collagens at a glance. Topics in Current Chemistry, 247, 1–6. https://doi.org/10.1007/b103817

Brittingham, R., Uitto, J., & Fertala, A. (2006). High-affinity binding of the NC1 domain of collagen VII to laminin 5 and collagen IV. Biochemical and biophysical research communications, 343(3), 692-699.

Brodsky, B., & Persikov, A. V. (2005). Molecular structure of the collagen triple helix. Advances in protein chemistry, 70, 301-339. https://doi.org/10.1016/S0065-3233(05)70009-7

Brown, J. C., & Timpl, R. (1995). The collagen superfamily. International Archives of Allergy and Immunology, 107(4), 484–490. https://doi.org/10.1159/000237090

Bushby, K. M. D., Collins, J., & Hicks, D. (2014). Collagen Type VI Myopathies (pp. 185–199). https://doi.org/10.1007/978-94-007-7893-1_12

Cascella, R., Strafella, C., Caputo, V., Errichiello, V., Zampatti, S., Milano, F., … Giardina, E. (2018). Towards the application of precision medicine in Age-Related Macular Degeneration. Progress in Retinal and Eye Research, 63, 132–146. https://doi.org/10.1016/j.preteyeres.2017.11.004

Cescon, M., Gattazzo, F., Chen, P., & Bonaldo, P. (2015). Collagen VI at a glance. Journal of cell science, 128(19), 3525-3531. https://doi.org/10.1242/jcs.169748

Connizzo, B. K., Yannascoli, S. M., & Soslowsky, L. J. (2013). Structure–function relationships of postnatal tendon development: A parallel to healing. Matrix Biology, 32(2), 106–116. https://doi.org/10.1016/j.matbio.2013.01.007

Deltas, C. (2022). Pos-435 Next Generation Sequencing Identifies Candidate Genetic Modifiers Potentially Exacerbating Kidney Disease in Col4a3/a4 Heterozygous Patients. Kidney International Reports, 7(2), S194. https://doi.org/10.1016/j.ekir.2022.01.462

Duband, J. L., & Thiery, J. P. (1987). Distribution of laminin and collagens during avian neural crest development. Development, 101(3), 461-478. https://doi.org/10.1242/dev.101.3.461

Exposito, J. Y., Valcourt, U., Cluzel, C., & Lethias, C. (2010). The fibrillar collagen family. International Journal of Molecular Sciences, 11(2), 407–426. https://doi.org/10.3390/ijms11020407

Fichtner, M., Schuster, S., & Stark, H. (2020). Determination of scoring functions for protein damage susceptibility. Biosystems, 187, 104035. https://doi.org/10.1016/j.bbagen.2016.08.008

Francomano, C. A. (1995). Key role for a minor collagen. Nature Genetics, 9(1), 6–8. https://doi.org/10.1038/ng0195-6

Franzke, C. W., Tasanen, K., Schumann, H., & Bruckner-Tuderman, L. (2003). Collagenous transmembrane proteins: Collagen XVII as a prototype. Matrix Biology, 22(4), 299–309. https://doi.org/10.1016/S0945-053X(03)00051-9

Fratzl, P. (2008). Collagen: Structure and mechanics, an introduction. Collagen: Structure and Mechanics, 1–13. https://doi.org/10.1007/978-0-387-73906-9_1

Gelse, K., Pöschl, E., & Aigner, T. (2003). Collagens - Structure, function, and biosynthesis. Advanced Drug Delivery Reviews, 55(12), 1531–1546.

https://doi.org/10.1016/j.addr.2003.08.002

Goh, K. L., Listrat, A., & Béchet, D. (2014). Hierarchical Mechanics of Connective Tissues: Integrating Insights from Nano to Macroscopic Studies. Journal of Biomedical Nanotechnology, 10(10), 2464–2507. https://doi.org/10.1166/jbn.2014.1960

Gonçalves, T. J., Boutillon, F., Lefebvre, S., Goffin, V., Iwatsubo, T., Wakabayashi, T., ... & Armand, A. S. (2019). Collagen XXV promotes myoblast fusion during myogenic differentiation and muscle formation. Scientific reports, 9(1), 5878. https://doi.org/10.1038/s41598-019-42296-6

Gordon, M. K., & Hahn, R. A. (2010). Collagens. Cell and Tissue Research, 247–257. https://doi.org/10.1007/s00441-009-0844-4

Gregorio, V. De, Caparali, E. B., Shojaei, A., Ricardo, S., & Barua, M. (2023). Alport Syndrome: Clinical Spectrum and Therapeutic Advances. Kidney Medicine, 5(5), 100631. https://doi.org/10.1016/j.xkme.2023.100631

Heino, J. (2007). The collagen family members as cell adhesion proteins. BioEssays, 29(10), 1001–1010. https://doi.org/10.1002/bies.20636Hulmes, D. J. S. (2008). Collagen diversity, synthesis and assembly. Collagen: Structure and Mechanics, 15–47. https://doi.org/10.1007/978-0-387-73906-9_2

Huxley - Jones, J., Robertson, D. L., & Boot-Handford, R. P. (2007). On the origins of the extracellular matrix in vertebrates. Matrix Biology, 26(1), 2–11. https://doi.org/10.1016/j.matbio.2006.09.008

Imafuku, A., Nozu, K., Sawa, N., Hasegawa, E., Hiramatsu, R., Kawada, M., … Ubara, Y. (2018). Autosomal dominant form of type IV collagen nephropathy exists among patients with hereditary nephritis difficult to diagnose clinicopathological. Nephrology, 23(10), 940–947. https://doi.org/10.1111/nep.13115

Imamura, Y., Scott, I. C., & Greenspan, D. S. (2000). 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(12), 8749-8759. https://doi.org/10.1074/jbc.275.12.8749

Iozzo, R. V., & Gubbiotti, M. A. (2018). Extracellular matrix: The driving force of mammalian diseases. Matrix Biology, 71–72, 1–9. https://doi.org/10.1016/j.matbio.2018.03.023

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Khoshnoodi, J., Pedchenko, V., & Hudson, B. G. (2008). Mammalian collagen IV. Microscopy Research and Technique, 71(5), 357–370. https://doi.org/10.1002/jemt.20564

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., … Pandey, A. (2014). A draft map of the human proteome. Nature, 509(7502), 575–581. https://doi.org/10.1038/nature13302

Knupp, C., & Squire, J. M. (2003). Molecular Packing in Network-Forming Collagens. The Scientific World JOURNAL, 3, 558–577. https://doi.org/10.1100/tsw.2003.40

Kucharz, E. J. (1992). The Collagens: Biochemistry and Pathophysiology. FEBS Letters (Vol. 316). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-76197-3

LeBleu, V. S., MacDonald, B., & Kalluri, R. (2007). Structure and function of basement membranes. Experimental Biology and Medicine, 232(9), 1121–1129. https://doi.org/10.3181/0703-MR-72

Leivo, I., Vaheri, A., Timpl, R., & Wartiovaara, J. (1980). Appearance and distribution of collagens and laminin in the early mouse embryo. Developmental biology, 76(1), 100-114. https://doi.org/10.1016/0012-1606(80)90365-6

Linsenmayer, T. F. (1991). Collagen. In Cell Biology of Extracellular Matrix (pp. 7–44). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-3770-0_2

Liu, Q., Gibson, M. P., Sun, H., & Qin, C. (2013). Dentin sialophosphoprotein (DSPP) plays an essential role in the postnatal development and maintenance of mouse mandibular condylar cartilage. Journal of Histochemistry & Cytochemistry, 61(10), 749-758. https://doi.org/10.1369/0022155413502056

Mak, K.M., Png, C.Y.M. and Lee, D.J. (2016), Type V Collagen in Health, Disease, and Fibrosis. Anat. Rec., 299: 613-629. https://doi.org/10.1002/ar.23330

Malfait, F., Wenstrup, R. J., & De Paepe, A. (2010). Clinical and genetic aspects of Ehlers-Danlos syndrome, classic type. Genetics in Medicine, 12(10), 597–605. https://doi.org/10.1097/GIM.0b013e3181eed412

Marneros, A. G., & Olsen, B. R. (2001). The role of collagen-derived proteolytic fragments in angiogenesis. Matrix Biology, 20(5–6), 337–345. https://doi.org/10.1016/S0945-053X(01)00151-2

Marro, J., Pfefferli, C., De Charles, A. S. P., Bise, T., & Jaźwiñska, A. (2016). Collagen XII contributes to epicardial and connective tissues in the zebrafish heart during ontogenesis and regeneration. PLoS ONE, 11(10), e0165497. https://doi.org/10.1371/journal.pone.0165497

Mayne, R., & Brewton, R. G. (1993). New members of the collagen superfamily. Current Opinion in Cell Biology, 5(5), 883–890. https://doi.org/10.1016/0955-0674(93)90039-S

Mayne, R., & Burgeson, R. E. (1987). Structure and Function of Collagen Types. Structure and Function of Collagen Types (Vol. III). Elsevier. https://doi.org/10.1016/B978-0-12-481280-2.X5001-6

McAlinden, A., Smith, T. A., Sandell, L. J., Ficheux, D., Parry, D. A. D., & Hulmes, D. J. S. (2003). α-Helical Coiled-coil Oligomerization Domains Are Almost Ubiquitous in the Collagen Superfamily. Journal of Biological Chemistry, 278(43), 42200–42207. https://doi.org/10.1074/jbc.M302429200

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

https://doi.org/10.1016/j.mbplus.2021.100058

Miller, E. J., Gay, S. (1982). Collagen. An Overview. In: Structural and Contractile Proteins, Part A: Extracellular Matrix, 3–32. Academic Press.

Nassa, M., Anand, P., Jain, A., Chhabra, A., Jaiswal, A., Malhotra, U., & Rani, V. (2012). Analysis of human collagen sequences. Bioinformation, 8(1), 26–33. https://doi.org/10.6026/97320630008026

Nimni, M. E., & Harkness, R. D. (2018). Molecular Structure and Functions of Collagen. In Collagen (pp. 1–78). https://doi.org/10.1201/9781351070799-1

Oh, S. P., May Griffith, C., Hay, E. D., & Olsen, B. R. (1993). Tissue-specific expression of type XII collagen during mouse embryonic development. Developmental Dynamics, 196(1), 37–46. https://doi.org/10.1002/aja.1001960105

Patino, M. G., Neiders, M. E., Andreana, S., Noble, B., & Cohen, R. E. (2002). Collagen: An Overview. Implant Dentistry, 11(3), 280–285. https://doi.org/10.1097/00008505-200207000-00014

Ricard-Blum, S. (2010). The Collagen Family. Cold Spring Harbor Perspectives in Biology, 3(1), 1–19. https://doi.org/10.1101/cshperspect.a004978

Ricard-Blum, S., Ruggiero, F., & van der Rest, M. (2005). The Collagen Superfamily: from the extracellular matrix to the cell membrane. In Pathologie Biologie (Vol. 53, pp. 35–84). https://doi.org/10.1007/b103819

Robin, N. H., Moran, R. T., & Ala-Kokko, L. (2000). Stickler Syndrome. In M. P. Adam (Eds.) et. al., GeneReviews®. University of Washington, Seattle.

Roig-Rosello E, Rousselle P. The Human Epidermal Basement Membrane: A Shaped and Cell Instructive Platform That Aging Slowly Alters. Biomolecules. 2020; 10(12):1607. https://doi.org/10.3390/biom10121607

Salamito, M., Nauroy, P., & Ruggiero, F. (2021). The Collagen Superfamily: Everything You Always Wanted to Know. In F. Ruggiero (Ed.), The Collagen Superfamily and Collagenopathies (8th ed., pp. 1–22). Springer.

https://doi.org/10.1007/978-3-030-67592-9_1

Seppinen, L., & Pihlajaniemi, T. (2011). The multiple functions of collagen XVIII in development and disease. Matrix Biology, 30(2), 83–92. https://doi.org/10.1016/j.matbio.2010.11.001

Shulman, C., Liang, E., Kamura, M., Udwan, K., Yao, T., Cattran, D., … Barua, M. (2021). Type IV Collagen Variants in CKD: Performance of Computational Predictions for Identifying Pathogenic Variants. Kidney Medicine, 3(2), 257–266. https://doi.org/10.1016/j.xkme.2020.12.007

Steinmann, B., Royce, P. M., & Superti-Furga, A. (n.d.). The Ehlers-Danlos Syndrome. In Connective Tissue and Its Heritable Disorders (pp. 431–523). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471221929.ch9

Stover, D. A., & Verrelli, B. C. (2011). Comparative vertebrate evolutionary analyses of type I collagen: potential of COL1a1 gene structure and intron variation for common bone-related diseases. Molecular biology and evolution, 28(1), 533-542.

Tajbakhsh, S., & Spörle, R. (1998). Somite development: constructing the vertebrate body. Cell, 92(1), 9-16. https://doi.org/10.1016/S0092-8674(00)80894-6

Tonelotto, V., Trapani, V., Bretaud, S., Heumüller, S. E., Wagener, R., Ruggiero, F., & Bonaldo, P. (2019). Spatio-temporal expression and distribution of collagen VI during zebrafish development. Scientific Reports, 9(1), 19851. https://doi.org/10.1038/s41598-019-56445-4

Trainor, P. A., & Krumlauf, R. (2001). Hox genes, neural crest cells and branchial arch patterning. Current opinion in cell biology, 13(6), 698-705. https://doi.org/10.1016/S0955-0674(00)00273-8

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Pontén, F. (2015). Tissue-based map of the human proteome. Science, 347(6220). https://doi.org/10.1126/science.1260419

Van Der Rest, M., & Garrone, R. (1991). Collagen family of proteins. The FASEB

Journal, 5(13), 2814–2823. https://doi.org/10.1096/fasebj.5.13.1916105

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., … Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Vuorio, E., & De Crombrugghe, B. (1990). The family of collagen genes. Annual Review of Biochemistry, 59, 837–872. https://doi.org/10.1146/annurev.bi.59.070190.004201

Zhang, J., & Patel, D. V. (2015). The pathophysiology of Fuchs' endothelial dystrophy – A review of molecular and cellular insights. Experimental Eye Research, 130, 97–105. https://doi.org/10.1016/j.exer.2014.10.023

Zhao, C., Xiao, Y., Ling, S., Pei, Y., & Ren, J. (2021). Structure of Collagen. In S. Ling (Ed.), Fibrous Proteins (pp. 17–25). New York. https://doi.org/10.1007/978-1-0716-1574-4_2

Mendeley repository

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 (2024), "Predicting bonds between collagens – An alignment approach", Mendeley Data, V2, doi: 10.17632/9pf5n687z4.2