Collagen formation

Canty-Laird, EG., Jupe, S., Kalamajski, S., Raleigh, S., Ricard-Blum, S.

European Bioinformatics Institute, New York University Langone Medical Center, Ontario Institute for Cancer Research, Oregon Health and Science University.

The contents of this document may be freely copied and distributed in any media, provided the authors, plus the institutions, are credited, as stated under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) License. For more information see our license.

This is just an excerpt of a full-length report for this pathway. To access the complete report, please download it at the Reactome Textbook.

18/11/2022
Introduction

Reactome is open-source, open access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many bioinformatics databases. A system of evidence tracking ensures that all assertions are backed up by the primary literature. Reactome is used by clinicians, geneticists, genomics researchers, and molecular biologists to interpret the results of high-throughput experimental studies, by bioinformaticians seeking to develop novel algorithms for mining knowledge from genomic studies, and by systems biologists building predictive models of normal and disease variant pathways.

The development of Reactome is supported by grants from the US National Institutes of Health (P41 HG003751), University of Toronto (CFREF Medicine by Design), European Union (EU STRP, EMI-CD), and the European Molecular Biology Laboratory (EBI Industry program).

Literature references


Reactome database release: 82

This document contains 3 pathways (see Table of Contents)
Collagen is a family of at least 29 structural proteins derived from over 40 human genes (Myllyharju & Kivirikko 2004). It is the main component of connective tissue, and the most abundant protein in mammals making up about 25% to 35% of whole-body protein content. A defining feature of collagens is the formation of trimeric left-handed polyproline II-type helical collagenous regions. The packing within these regions is made possible by the presence of the smallest amino acid, glycine, at every third residue, resulting in a repeating motif Gly-X-Y where X is often proline (Pro) and Y often 4-hydroxyproline (4Hyp). Gly-Pro-Hyp is the most common triplet in collagen (Ramshaw et al. 1998). Collagen peptide chains also have non-collagenous domains, with collagen subclasses having common chain structures. Collagen fibrils are mostly found in fibrous tissues such as tendon, ligament and skin. Other forms of collagen are abundant in cornea, cartilage, bone, blood vessels, the gut, and intervertebral disc. In muscle tissue, collagen is a major component of the endomysium, constituting up to 6% of muscle mass. Gelatin, used in food and industry, is collagen that has been irreversibly hydrolyzed.

On the basis of their fibre architecture in tissues, the genetically distinct collagens have been divided into subgroups. Group 1 collagens have uninterrupted triple-helical domains of about 300 nm, forming large extracellular fibrils. They are referred to as the fibril-forming collagens, consisting of collagens types I, II, III, V, XI, XXIV and XXVII. Group 2 collagens are types IV and VII, which have extended triple helices (>350 nm) with imperfections in the Gly-X-Y repeat sequences. Group 3 are the short-chain collagens. These have two subgroups. Group 3A have continuous triple-helical domains (type VI, VIII and X). Group 3B have interrupted triple-helical domains, referred to as the fibril-associated collagens with interrupted triple helices (FACIT collagens, Shaw & Olsen 1991). FACITs include collagen IX, XII, XIV, XVI, XIX, XX, XXI, XXII and XXVI plus the transmembrane collagens (XIII, XVII, XXIII and XXV) and the multiple triple helix domains and interruptions (Multiplexin) collagens XV and XVIII (Myllyharju & Kivirikko 2004). The non-collagenous domains of collagens have regulatory functions; several are biologically active when cleaved from the main peptide chain. Fibrillar collagen peptides all have a large triple helical domain (COL1) bordered by N and C terminal extensions, called the N- and C-propeptides, which are cleaved prior to formation of the collagen fibril. The intact form is referred to as a collagen propeptide, not procollagen, which is used to refer to the trimeric triple-helical precursor of collagen before the
propeptides are removed. The C-propeptide, also called the NC1 domain, directs chain association during assembly of the procollagen molecule from its three constituent alpha chains (Hulmes 2002).

Fibril forming collagens are the most familiar and best studied subgroup. Collagen fibres are aggregates or bundles of collagen fibrils, which are themselves polymers of tropocollagen complexes, each consisting of three polypeptide chains known as alpha chains. Tropocollagens are considered the subunit of larger collagen structures. They are approximately 300 nm long and 1.5 nm in diameter, with a left-handed triple-helical structure, which becomes twisted into a right-handed coiled-coil 'super helix' in the collagen fibril. Tropocollagens in the extracellular space polymerize spontaneously with regularly staggered ends (Hulmes 2002). In fibrillar collagens the molecules are staggered by about 67 nm, a unit known as D that changes depending upon the hydration state. Each D-period contains slightly more than four collagen molecules so that every D-period repeat of the microfibril has a region containing five molecules in cross-section, called the 'overlap', and a region containing only four molecules, called the 'gap'. The triple-helices are arranged in a hexagonal or quasi-hexagonal array in cross-section, in both the gap and overlap regions (Orgel et al. 2006). Collagen molecules cross-link covalently to each other via lysine and hydroxylysine side chains. These cross-links are unusual, occuring only in collagen and elastin, a related protein.

The macromolecular structures of collagen are diverse. Several group 3 collagens associate with larger collagen fibers, serving as molecular bridges which stabilize the organization of the extracellular matrix. Type IV collagen is arranged in an interlacing network within the dermal-epidermal junction and vascular basement membranes. Type VI collagen forms distinct microfibrils called beaded filaments. Type VII collagen forms anchoring fibrils. Type VIII and X collagens form hexagonal networks. Type XVII collagen is a component of hemidesmosomes where it is complexed with alpha6Beta4 integrin, plectin, and laminin-332 (de Pereda et al. 2009). Type XXIX collagen has been recently reported to be a putative epidermal collagen with highest expression in suprabasal layers (Soderhall et al. 2007). Collagen fibrils/aggregates arranged in varying combinations and concentrations in different tissues provide specific tissue properties. In bone, collagen triple helices lie in a parallel, staggered array with 40 nm gaps between the ends of the tropocollagen subunits, which probably serve as nucleation sites for the deposition of crystals of the mineral component, hydroxyapatite (Ca10(PO4)6(OH)2) with some phosphate. Collagen structure affects cell-cell and cell-matrix communication, tissue construction in growth and repair, and is changed in development and disease (Sweeney et al. 2006, Twardowski et al. 2007). A single collagen fibril can be heterogeneous along its axis, with significantly different mechanical properties in the gap and overlap regions, correlating with the different molecular organizations in these regions (Minary-Jolandan & Yu 2009).

**Literature references**


**Editions**

<table>
<thead>
<tr>
<th>Date</th>
<th>Action</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-08-05</td>
<td>Authored</td>
<td>Jupe, S.</td>
</tr>
<tr>
<td>2012-04-11</td>
<td>Edited</td>
<td>Jupe, S.</td>
</tr>
<tr>
<td>2012-05-24</td>
<td>Reviewed</td>
<td>Canty-Laird, EG.</td>
</tr>
</tbody>
</table>

*https://reactome.org*
Collagen biosynthesis and modifying enzymes

Location: Collagen formation

Stable identifier: R-HSA-1650814

The biosynthesis of collagen is a multistep process. Collagen propeptides are cotranslationally translocated into the ER lumen. Propeptides undergo a number of post-translational modifications. Proline and lysine residues may be hydroxylated by prolyl 3-, prolyl 4- and lysyl hydroxylases. 4-hydroxyproline is essential for intramolecular hydrogen bonding and stability of the triple helical collagenous domain. In fibril forming collagens approximately 50% of prolines are 4-hydroxylated; the extent of this and of 3-hydroxyproline and lysine hydroxylation varies between tissues and collagen types (Kivirikko et al. 1972, 1992). Hydroxylysine molecules can form cross-links between collagen molecules in fibrils, and are sites for glycosyl- and galactosylation. Collagen peptides all have non-collagenous domains; collagens within the subclasses have common chain structures. These non-collagenous domains have regulatory functions; some are biologically active when cleaved from the main peptide chain. Fibrillar collagens all have a large triple helical domain (COL1) bordered by N and C terminal extensions, called the N and C propeptides, which are cleaved prior to formation of the collagen fibril. The C propeptide, also called the NC1 domain, is highly conserved. It directs chain association during intracellular assembly of the procollagen molecule from three collagen propeptide alpha chains (Hulmes 2002). The N-propeptide has a short linker (NC2) connecting the main triple helix to a short minor one (COL2) and a globular N-terminal region NC3. NC3 domains are variable both in size and the domains they contain.

Collagen propeptides typically undergo a number of post-translational modifications. Proline and lysine residues are hydroxylated by prolyl 3-, prolyl 4- and lysyl hydroxylases. 4-hydroxyproline is essential for intramolecular hydrogen bonding and stability of the triple helical collagenous domain. Prolyl 4-hydroxylase may also have a role in alpha chain association as no association of the C-propeptides of type XII collagen was seen in the presence of prolyl 4-hydroxylase inhibitors (Mazzorana et al. 1993, 1996). In fibril forming collagens approximately 50% of prolines are 4-hydroxylated; the extent of this is species
dependent, lower hydroxylation correlating with lower ambient temperature and thermal stability (Cohen-Solal et al. 1986, Notbohm et al. 1992). Similarly the extent of 3-hydroxyproline and lysine hydroxylation varies between tissues and collagen types (Kivirikko et al. 1992). Hydroxylsine molecules can form cross-links between collagen molecules in fibrils, and are sites for glycosyl- and galactosylation.

Collagen molecules fold and assemble through a series of distinct intermediates (Bulleid 1996). Individual collagen polypeptide chains are translocated co-translationally across the membrane of the endoplasmic reticulum (ER). Intra-chain disulfide bonds are formed within the N-propeptide, and hydroxylation of proline and lysine residues occurs within the triple helical domain (Kivirikko et al. 1992). When the peptide chain is fully translocated into the ER lumen the C-propeptide folds, the conformation being stabilized by intra-chain disulfide bonds (Doege and Fessler 1986). Pro alpha-chains associate via the C-propeptides (Byers et al. 1975, Bachinger et al. 1978), or NC2 domains for FACIT family collagens (Boudko et al. 2008) to form an initial trimer which can be stabilized by the formation of inter-chain disulfide bonds (Schofield et al. 1974, Olsen et al. 1976), though these are not a prerequisite for further folding (Bulleid et al. 1996). The triple helix then nucleates and folds in a C- to N- direction. The association of the individual chains and subsequent triple helix formation are distinct steps (Bachinger et al. 1980). The N-propeptides associate and in some cases form inter-chain disulfide bonds (Bruckner et al., 1978). Procollagen is released via carriers into the extracellular space (Canty & Kadler 2005). Fibrillar procollagens undergo removal of the C- and N-propeptides by procollagen C and N proteinases respectively, both Zn2+ dependent metalloproteinases. Propeptide processing is a required step for normal collagen I and III fibril formation, but collagens can retain some or all of their non-collagenous propeptides. Retained collagen type V and XI N-propeptides contribute to the control of fibril growth by sterically limiting lateral molecule addition (Fichard et al. 1995). Processed fibrillar procollagen is termed tropocollagen, which is considered to be the unit of higher order fibrils and fibres. Tropocollagens of the fibril forming collagens I, II, III, V and XI spontaneously aggregate in vitro in a manner that has been compared with crystallization, commencing with a nucleation event followed by subsequent organized aggregation (Silver et al. 1992, Prockop & Fertala 1998). Fibril formation is stabilized by lysyl oxidase catalyzed crosslinks between adjacent molecules (Siegel & Fu 1976).

**Literature references**


**Editions**

<table>
<thead>
<tr>
<th>Date</th>
<th>Action</th>
<th>Author/Reviewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-07-20</td>
<td>Authored</td>
<td>Jupe, S.</td>
</tr>
<tr>
<td>2012-05-14</td>
<td>Edited</td>
<td>Jupe, S.</td>
</tr>
<tr>
<td>2012-05-24</td>
<td>Reviewed</td>
<td>Canty-Laird, EG.</td>
</tr>
</tbody>
</table>

[https://reactome.org](https://reactome.org)
Collagen trimers in triple-helical form, referred to as procollagen or collagen molecules, are exported from the ER and trafficked through the Golgi network before secretion into the extracellular space. For fibrillar collagens namely types I, II, III, V, XI, XXIV and XXVII (Gordon & Hahn 2010, Ricard-Blum 2011) secretion is concomitant with processing of the N and C terminal collagen propeptides. These processed molecules are known as tropocollagens, considered to be the units of higher order collagen structures. They form within the extracellular space via a process that can proceed spontaneously, but in the cellular environment is regulated by many collagen binding proteins such as the FACIT (Fibril Associated Collagens with Interrupted Triple helices) family collagens and Small Leucine-Rich Proteoglycans (SLRPs). The architecture formed ultimately depends on the collagen subtype and the cellular conditions. Structures include the well-known fibrils and fibres formed by the major structural collagens type I and II plus several different types of supramolecular assembly (Bruckner 2010). The mechanical and physical properties of tissues depend on the spatial arrangement and composition of these collagen-containing structures (Kadler et al. 1996, Shoulders & Raines 2009, Birk & Bruckner 2011).

Fibrillar collagen structures are frequently heterotypic, composed of a major collagen type in association with smaller amounts of other types, e.g. type I collagen fibrils are associated with types III and V, while type II fibrils frequently contain types IX and XI (Wess 2005). Fibres composed exclusively of a single collagen type probably do not exist, as type I and II fibrils require collagens V and XI respectively as nucleators (Kadler et al. 2008, Wensstrup et al. 2011). Much of the structural understanding of collagen fibrils has been obtained with fibril-forming collagens, particularly type I, but some central features are believed to apply to at least the other fibrillar collagen subtypes (Wess 2005). Fibril diameter and length varies considerably, depending on the tissue and collagen types (Fang et al. 2012). The reasons for this are poorly understood (Wess 2005).

Some tissues such as skin have fibres that are approximately the same diameter while others such as ten-
don or cartilage have a bimodal distribution of thick and thin fibrils. Mature type I collagen fibrils in tendon are up to 1 cm in length, with a diameter of approx. 500 nm. An individual fibrillar collagen triple helix is less than 1.5 nm in diameter and around 300 nm long; collagen molecules must assemble to give rise to the higher-order fibril structure, a process known as fibrillogenesis, prevented by the presence of C-terminal propeptides (Kadler et al. 1987). In electron micrographs, fibrils have a banded appearance, due to regular gaps where fewer collagen molecules overlap, which occur because the fibrils are aligned in a quarter-stagger arrangement (Hodge & Petruska 1963). Collagen microfibrils are believed to have a quasi-hexagonal unit cell, with tropocollagen arranged to form supertwisted, right-handed microfibrils that interdigitate with neighbouring microfibrils, leading to a spiral-like structure for the mature collagen fibril (Orgel et al. 2006, Holmes & Kadler 2006).

Neighbouring tropocollagen monomers interact with each other and are cross-linked covalently by lysyl oxidase (Orgel et al. 2000, Maki 2006). Mature collagen fibrils are stabilized by lysyl oxidase-mediated cross-links. Hydroxylysyl pyridinoline and lysyl pyridinoline cross-links form between (hydroxy)lysine and hydroxylysine residues in bone and cartilage (Eyre et al. 1984). Arginoline cross-links can form in cartilage (Eyre et al. 2010); mature bovine articular cartilage contains roughly equimolar amounts of arginoline and hydroxylysyl pyridinoline based on peptide yields. Mature collagen fibrils in skin are stabilized by the lysyl oxidase-mediated cross-link histidinohydroxylysino norleucine (Yamauch et al. 1987). Due to the quarter-staggered arrangement of collagen molecules in a fibril, telopeptides most often interact with the triple helix of a neighbouring collagen molecule in the fibril, except for collagen molecules in register staggered by 4D from another collagen molecule. Fibril aggregation in vitro can be unipolar or bipolar, influenced by temperature and levels of C-proteinase, suggesting a role for the N- and C-propeptides in regulation of the aggregation process (Kadler et al. 1996). In vivo, collagen molecules at the fibril surface may retain their N-propeptides, suggesting that this may limit further accretion, or alternatively represents a transient stage in a model whereby fibrils grow in diameter through a cycle of deposition, cleavage and further deposition (Chapman 1989).

In vivo, fibrils are often composed from more than one type of collagen. Type III collagen is found associated with type I collagen in dermal fibrils, with the collagen III on the periphery, suggesting a regulatory role (Fleischmajer et al. 1990). Type V collagen associates with type I collagen fibrils, where it may limit fibril diameter (Birk et al. 1990, White et al. 1997). Type IX associates with the surface of narrow diameter collagen II fibrils in cartilage and the cornea (Wu et al. 1992, Eyre et al. 2004). Highly specific patterns of crosslinking sites suggest that collagen IX functions in interfibrillar networking (Wess 2005). Type XII and XIV collagens are localized near the surface of banded collagen I fibrils (Nishiyama et al. 1994). Certain fibril-associated collagens with interrupted triple helices (FACITs) associate with the surface of collagen fibrils, where they may serve to limit fibril fusion and thereby regulate fibril diameter (Gordon & Hahn 2010). Collagen XV, a member of the multiplexin family, is almost exclusively associated with the fibrillar collagen network, in very close proximity to the basement membrane. In human tissues collagen XV is seen linking banded collagen fibers subjacent to the basement membrane (Amenta et al. 2005). Type XIV collagen, SLRPs and discoidin domain receptors also regulate fibrillogenesis (Ansorge et al. 2009, Kalamajski et al. 2010, Flynn et al. 2010).

Collagen IX is cross-linked to the surface of collagen type II fibrils (Eyre et al. 1987). Type XII and XIV collagens are found in association with type I (Walchli et al. 1994) and type II (Watt et al. 1992, Eyre 2002) fibrils in cartilage. They are thought to associate non-covalently via their COL1/NC1 domains (Watt et al. 1992, Eyre 2002).

Some non-fibrillar collagens form supramolecular assemblies that are distinct from typical fibrils. Collagen VII forms anchoring fibrils, composed of antiparallel dimers that connect the dermis to the epidermis (Bruckner-Tuderman 2009). During fibrillogenesis, the nascent type VII procollagen molecules dimer-
ize in an antiparallel manner. The C-propeptides are then removed by Bone morphogenetic protein 1 (Rattenholl et al. 2002) and the processed antiparallel dimers aggregate laterally. Collagens VIII and X form hexagonal networks and collagen VI forms beaded filament (Gordon & Hahn 2010, Ricard-Blum et al. 2011).

**Literature references**


**Editions**

<table>
<thead>
<tr>
<th>Date</th>
<th>Action</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-08-05</td>
<td>Authored</td>
<td>Jupe, S.</td>
</tr>
<tr>
<td>2012-10-08</td>
<td>Reviewed</td>
<td>Kalamajski, S., Raleigh, S.</td>
</tr>
<tr>
<td>2012-11-12</td>
<td>Edited</td>
<td>Jupe, S.</td>
</tr>
<tr>
<td>2012-11-19</td>
<td>Reviewed</td>
<td>Ricard-Blum, S.</td>
</tr>
</tbody>
</table>
# Table of Contents

- Introduction ........................................ 1
- Collagen formation .................................. 2
  - Collagen biosynthesis and modifying enzymes ...... 4
  - Assembly of collagen fibrils and other multimeric structures ........ 6

Table of Contents ........................................ 9