Molecular Recognition in the Assembly of Collagens: Terminal Noncollagenous Domains Are Key Recognition Modules in the Formation of Triple Helical Protomers*

Published, JBC Papers in Press, November 2, 2006, DOI 10.1074/jbc.R600025200

Jamshid Khoshnoodi‡1, Jean-Philippe Cartailer‡1, Keith Alvarese1, Arthur Veis1, and Billy G. Hudson1,2

From the 1Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37232-2372, 2Symmation LLC, Spring Hill, Tennessee 37174, and 3Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611

The α-chains of the collagen superfamily are encoded with information that specifies self-assembly into fibrils, microfibrils, and networks that have diverse functions in the extracellular matrix. A key self-organizing step, common to all collagen types, is trimerization that selects, binds, and registers cognate α-chains for assembly of triple helical protomers that subsequently oligomerize into specific suprastructures. In this article, we review recent findings on the mechanism of chain selection and infer that terminal noncollagenous domains function as recognition modules in trimerization and are therefore key determinants of specificity in the assembly of suprastructures. This mechanism is also illustrated with computer-generated animations.

Collagens are modular triple helical proteins that constitute the major structural components of the extracellular matrix of all animals. They occur as diverse suprastructures such as fibrils, microfibrils, and networks, which serve as self-organizing scaffolds for the attachment of other macromolecular complexes including laminin networks, proteoglycans, and cell surface receptors. The suprastructures play functional roles in cell adhesion, cell differentiation, tissue development, and the structural integrity of organs.

Collagen suprastructures are assembled from a large family of gene products called α-chains. To date, 43 unique α-chains that belong to 28 types of collagens (types I–XXVIII) have been discovered in vertebrates. Based on their supramolecular architectures, they are further classified as fibril-forming, fibril-associated containing interrupted triple helices (FACIT),3 beaded filament, anchoring fibril, network-forming, and transmembrane collagens (1–4).

The assembly of all collagen suprastructures begins with the association of three type-specific α-chains (trimerization) that subsequently intertwine to form triple helical protomers, the building blocks of larger assemblies (1–5). Protomers are homotrimmers or heterotrimers, composed of up to three different α-chains. They have in common at least one triple helical collagenous domain of varying length and two noncollagenous domains (NC) of variable sequence, size, and shape that are positioned at the N and C termini, designated herein as N-NC and C-NC domains, respectively. The terminal NC domains are excised, modified, or incorporated directly into the final suprastructure, depending on protomer type and function. Subsequently, specific protomers oligomerize into distinct suprastructures involving interactions that form end-to-end connections, lateral associations, and supercoiling of helices. Thus, protomer formation and oligomerization involve pivotal recognitions steps that target specific α-chains to assemble into a particular type of suprastructure.

How the α-chains selectively recognize each other is a fundamental question in matrix biology that remains largely unanswered. Early studies on collagens I and III suggested that the C-NC domains play a critical role in the trimerization step that involves selection, binding, and registration of three α-chains (6–9). In this article, we review recent findings on the trimerization of collagens with an emphasis on collagens I and IV, which are the best understood.

Fibril-forming Collagens

Fibril-forming collagens include the widely distributed types I, III, and V, types II and XI, prominent in cartilage and eye, and the newly discovered types XXIV and XXVII (10–12). All of the 12 fibril-forming α-chains share a long uninterrupted collagenous domain flanked by N- and C-terminal noncollagenous propeptides. The α-chains assemble into at least 12 type-specific protomers, characterized as homo- and heterotrimers. The terminal propeptides are further cleaved by specific proteases, which promote oligomerization and fibril formation.

Type I collagen is composed of α1- and α2-chains, forming preferably an α1α1α2 heterotrimer (Fig. 1). However, if only the triple helical domains are considered, the (α1)3 homotrimer would be the preferred form (13, 14). The pro-α1 homotrimer is formed as the default structure in the absence of pro-α2-chain, whereas the opposite is not true (15). However, if the α2 C-NC domain is replaced with an artificial trimeric NC domain, an α2 homotrimer can form (16, 17). Thus the α2 C-NC domain is the key domain required for heterotrimer assembly, although the α1 C-NC domain contains sufficient information for directing homotrimer formation. Disulfide bonding in the α2 C-NC domain is crucial for heterotrimer formation (18), particularly

ruptured triple helices; NC domain, noncollagenous domain; N-NC domain, amino-terminal noncollagenous domain, which is the same as N-propeptide; C-NC domain, carboxyl-terminal noncollagenous domain, which is the same as C-propeptide.
Within the most C-terminal sequence of the α2 C-NC domain (19, 20).

The putative conformation and energy-minimized structures of the human α1 C-NC and α2 C-NC reveal (21, 22) that each contains five subdomains (Fig. 1A) (8, 23). Domain I folds into subdomains Ia and Ib without α-helical or β-sheet conformations, inconsistent with suggestions that subdomain Ia participates in trimerization by forming α-helical coiled-coils (23–25). Subdomain Ib contains the interchain disulfide bonds in the assembled C-NC trimer. Domains II and IV fold into globular regions G1 and G2, respectively. These are linked by an antiparallel β-sheet assembled from domains III and V (Fig. 1). Energy minimization and molecular dynamics simulations have revealed key steps in the folding, docking, and assembly of two α1 C-NC (α11 and α12) and α2 C-NC domains into a heterotrimer (26). Folding is initiated at the C terminus by bimolecular association of the α2 and α12, near the junction of domain III and domain IV-G2, and proceeds toward the N terminus (Fig. 1B). The α2 domain V appears to have a dominant effect on heterotrimer formation (19, 20). Trimerization proceeds via a second interaction between the α2-α12 dimer and the α11 at domain II-G1, after which the interchain disulfide bonds become established in domain Ib. Folding of domain Ia is the last and slowest folding step but eventually drives the folding through the C-telo-Ia junction (26, 27). Overall, the α2 C-NC domain appears to provide the driving force for heterotrimer formation. Computer modeling of collagen I C-propeptide (26) also suggests that the homotrimer could resemble a cruciform as seen in the collagen III C-propeptide (28).

Putative trimerization control sequences have been located within the C-NC domains (17, 29). Replacing the α2(I) C-NC domain of the α2(I)-chain with the homotrimer-directing α1(III) C-NC permitted α2-α2-α2 homo-trimerization (17, 29), demonstrating that the α1 C-NC domain is sufficient to direct homotrimer assembly of α2-chains with aligned triple helices. Exchanging specific sequences from the α1 C-NC domain of collagen III with the corresponding sequences of the α2 C-NC of collagen I revealed the location of a putative recognition site, a discontinuous sequence of 15 amino acids, located at the C-terminal of subdomain III (Fig. 1). This α1 sequence, when transferred to the sequence-equivalent region of the α2 C-NC domain, enabled the α2-chains to homotrimerize (17, 29). Thus, this recognition site is both necessary and sufficient to ensure that collagen chains discriminate between each other to form type-specific protomers (Fig. 1, A and C).

**Network-forming Collagens**

The network-forming collagens include types IV, VIII and X. In contrast to the fibril-forming collagens, their C-NC domains are retained in their suprastructures. Importantly, crystal structures of their C-NC domains are known, which provide insight into the mechanisms of chain selection.

**Collagen IV**—Type IV collagen is the major constituent of basement membranes. It is composed of a family of six homologous α-chains (α1–α6), each characterized by a long collagenous domain of ~1400 residues, interrupted by about 20 short noncollagenous sequences, which is flanked by a short N-NC sequence of 25 residues and a globular C-NC domain of 230 residues. The six α-chains assemble into three heterotrimeric protomers (α1-α1-α2, α3-α4-α5, and α5-α5-α6), which further assemble into three distinct networks (30).

Collagen IV has provided a unique opportunity to gain insight into chain recognition mechanisms. First, the six α-chains assemble into 3 specific protomers out of 76 possible combinations, reflecting a remarkable specificity for chain selection in vivo (30, 31). Second, the extraordinary capacity of monomeric C-NC domains to recognize each other and assemble in vitro into their original trimeric and hexameric compositions (32), together with their critical role in trimerization in vivo (33), indicates a recognition role in assembly and provides the means to characterize specificity and binding parameters (34). Third, the crystal structure of the C-NC hexamer is known (31, 35), which reveals the nature of intermolecular interactions and their role in chain selectivity. Fourth, the known primary structure for all six C-NC domains from many species provides information needed to deduce candidate recognition sites from sequence variation (34). Finally, a study of sequence variation, coupled with structural and kinetic analysis, provides the requisite information needed to decipher the location and three-dimensional features of putative recognition sites (34).

Two putative recognition sites were identified in the C-NC domains as prominent mediators in the mechanism of chain recognition (34). These sites (Fig. 2D) constitute a 13-residue-long β-hairpin motif (site 1) involved in the domain swapping mechanism and a 15-residue-long variable region (site 2, VR3, docking site) with sequence hypervariability across all the chains of collagen IV (34). Accordingly, we have proposed that the combination of sequence variations at these two sites and at the neighboring C-NC domains to which they interact consti-
Collagen X—Type X collagen is expressed in the endochondral growth plate. The single α1-chain contains a short collagenous domain of 154 residues, flanked by a N-NC domain of 37 residues and a C-NC domain of 161 residues (41). Its protomer is a homotrimer. The C-NC domains function as nucleation sites for trimer and multimer formation, based on experiments using recombinant C-NC domains (42). A comparison of the x-ray crystal structures of the C-NC trimer of collagens X and VIII revealed them to be highly homologous, but they differ in intermolecular contact residues that govern trimer stability. Such differences may also confer specificity (40, 43).

FACIT Collagens

The FACIT collagens include types IX, XII, XIV, XVI, XIX, XX, XXI, and XXII. Collagen IX is composed of three α-chains and all others of one α-chain, each of which is characterized by short collagenous domains interrupted by several NC domains (1, 2). The protomer of collagen IX is a heterotrimer, and all others are homotrimers, accounting for at least 9 distinct protomers. Unlike the fibril-forming collagens, the FACITs have significantly shorter C-NC domains: 75 residues for collagen XII and fewer than 30 residues for collagen IX, whereas those of fibrillar collagen are about 260 residues. The FACITs share a remarkable sequence homology at their COL1/C-NC junctions by having two strictly conserved cysteine residues separated by four residues at the NC domain. Several lines of evidence suggest that the COL1 domain at the COL1/C-NC junction (which contains the first cysteine residue) and the C-NC domain (containing the second conserved cysteine) are involved in the mechanism of chain selection in the assembly of collagens XII and XIV (44–46). Independent studies by Mazorana et al. (44, 45) and Lesage et al. (46) strongly suggest that both the collagenous portion and the NC domain are involved in formation and stabilization of protomers in collagens XII and XIV (47).

Beaded Filament and Anchoring Fibril Collagens

Collagen VI—Type VI collagen is ubiquitously expressed in connective tissues. It is composed of three genetically distinct α-chains, α1, α2, and α3, that form extensive beaded filaments (48–50). Each chain contains a relatively short collagenous domain of ∼335 residues flanked by N-NC and C-NC globular domains. The α1- and α2-chains are similar in size and contain one N-NC domain and a C-NC domain with two subdomains, C1 and C2. In contrast, the α3-chain is much longer, and the N-NC domain contains ten subdomains, N1–N10, and five C-NC subdomains, C1–C5 (48). The protomer is an α1α2α3 heterotrimer. Studies conducted on cell lines lacking or expressing different combinations of recombinant α-chains have shown that the α3-chain contains sequences necessary for chain association and protomer formation (48, 50). Deletion studies have demonstrated that whereas the C5 subdomain of the α3-chain is required for the extracellular microfibril formation, the C1 subdomain, illustrated in Fig. 3, in all chains is

Minireview: Collagen Assembly

Collagen Assembly Proteins

The crystal structure of the C-NC domain of 173 residues (37) indicates that the chains assemble into NC hexamers that are stabilized by local and extensive hydrophobic and hydrophilic forces. Within each NC trimer, the C-NC domain with two subdomains, C1 and C2. In contrast, the α3-chain is much longer, and the N-NC domain contains ten subdomains, N1–N10, and five C-NC subdomains, C1–C5 (48). The protomer is an α1α2α3 heterotrimer. Studies conducted on cell lines lacking or expressing different combinations of recombinant α-chains have shown that the α3-chain contains sequences necessary for chain association and protomer formation (48, 50). Deletion studies have demonstrated that whereas the C5 subdomain of the α3-chain is required for the extracellular microfibril formation, the C1 subdomain, illustrated in Fig. 3, in all chains is

Minireview: Collagen Assembly

Collagen Assembly Proteins

The crystal structure of the C-NC domain of 173 residues (37) indicates that the chains assemble into NC hexamers that are stabilized by local and extensive hydrophobic and hydrophilic forces. Within each NC trimer, the C-NC domain with two subdomains, C1 and C2. In contrast, the α3-chain is much longer, and the N-NC domain contains ten subdomains, N1–N10, and five C-NC subdomains, C1–C5 (48). The protomer is an α1α2α3 heterotrimer. Studies conducted on cell lines lacking or expressing different combinations of recombinant α-chains have shown that the α3-chain contains sequences necessary for chain association and protomer formation (48, 50). Deletion studies have demonstrated that whereas the C5 subdomain of the α3-chain is required for the extracellular microfibril formation, the C1 subdomain, illustrated in Fig. 3, in all chains is
Collagen Assembly

MINIREVIEW: Collagen Assembly

Collagen VII—Type VII collagen is a major component of anchoring fibrils that maintain the epidermal-dermal adherence of skin. It is composed of three identical α-chains, each consisting of a collagenous domain of 1530 residues with 19 interruptions, which are flanked by N-NC and C-NC domains (51). Protomers form anti-parallel dimers that are stabilized by disulfide bonding via the C-NC domains of adjacent protomers. Following proteolytic modification of the C-NC domain, interruptions, flanked by N-NC and C-NC domains. The protomer of each α-chain that contains a collagenous domain, with frequent interruptions, flanked by N-NC and C-NC domains. The C-NC domain of XVIII forms trimers in vitro suggesting a role in protomer assembly (60).

Other Collagens

Little is known about the assembly of collagen XV (57) and XVIII (58) and the newly discovered XXVI (59) and XXVIII (4). Each type is composed of a single α-chain that contains a collagenous domain, with frequent interruptions, flanked by N-NC and C-NC domains. The C-NC domain of XVIII forms trimers in vitro suggesting a role in protomer assembly (60). The crystal structures of fragments of C-NC domains of XV and XVIII are highly similar but distinct (61, 62), and they differ greatly with the three-dimensional structures of the C-NC domains of collagens VIII, X, and IV (see above); such distinctions are consistent with a special role of C-NC domains in chain selection.

Conclusions and Future Perspectives

Over the last decade, findings have emerged on several collagen types that provide new insights into the chain selection mechanism. Based on the body of evidence reviewed here, we infer that the terminal NC domains function as recognition modules that select, bind, and register three cognate α-chains for self-assembly of triple helical protomers. The C-NC domains govern chain selection for all classes except for the transmembrane collagens, which are governed by the N-NC domains (Fig. 3). As recognition modules, they contain sites that are characterized by positive determinants of shape complementarity, electrostatic charge distribution and magnitude, and hydrophobicity for selection of cognate α-chains and negative determinants for the exclusion of all other unrelated α-chains.

Following trimerization, coupled with triple helix formation, new recognition sites are encoded, a consequence of forming quaternary structures that commit protomers to oligomerize with cognate protomers into a particular type of suprastructure such as a fibril or network. The quaternary recognition sites specify such features as enzymatic cleavage sites, end-to-end connections, lateral associations, supercoiling of helices, and cross-link formation, as well as interactions with other macromolecular complexes such as proteoglycans (63) and integrin receptors (64). Therefore, by directing formation of triple helical protomers with distinct chain compositions and quaternary structures, the NC recognition modules are key determinants of specificity that target a type of α-chain to assemble into distinct suprastructures.

Many important features remain to be elucidated about the NC recognition mechanisms. The C-NC modules of heterotrimeric collagen XIII, a short sequence of 21 residues located at the juxtapasmembrane linker subdomain, has been identified as a putative recognition site for trimerization (Fig. 3).
mers composed of three different chains (collagens IV, V, VI, and IX) are the most promising models for study because their trimeric forms display quaternary structural features that can be differentially analyzed with regard to the identity of recognition motifs and the cyclic order of chains in the triple helix. Such mechanistic information provides opportunities for engineering new suprastructures for biomaterials and development of therapeutic strategies that facilitate or interfere with collagen assembly in various disorders.

REFERENCES