Collagen Structure and Stability

Matthew D. Shoulders\textsuperscript{1} and Ronald T. Raines\textsuperscript{1,2}

\textsuperscript{1}Department of Chemistry and \textsuperscript{2}Department of Biochemistry, University of Wisconsin, Madison, Wisconsin 53706; email: raines@biochem.wisc.edu

Key Words
biomaterials, extracellular matrix, fibrillogenesis, proline, stereoelectronic effects, triple helix

Abstract
Collagen is the most abundant protein in animals. This fibrous, structural protein comprises a right-handed bundle of three parallel, left-handed polyproline II-type helices. Much progress has been made in elucidating the structure of collagen triple helices and the physicochemical basis for their stability. New evidence demonstrates that stereoelectronic effects and preorganization play a key role in that stability. The fibrillar structure of type I collagen—the prototypical collagen fibril—has been revealed in detail. Artificial collagen fibrils that display some properties of natural collagen fibrils are now accessible using chemical synthesis and self-assembly. A rapidly emerging understanding of the mechanical and structural properties of native collagen fibrils will guide further development of artificial collagenous materials for biomedicine and nanotechnology.
INTRODUCTION
Collagen is an abundant structural protein in all animals. In humans, collagen comprises one-third of the total protein, accounts for three-quarters of the dry weight of skin, and is the most prevalent component of the extracellular matrix (ECM). Twenty-eight different types of collagen, composed of at least 46 distinct polypeptide chains, have been identified in vertebrates, and many other proteins contain collagenous domains (1, 2). Remarkably, intact collagen has been discovered in soft tissue of the fossilized bones of a 68 million-year-old Tyrannosaurus rex fossil (3, 4), by far the oldest protein detected to date. That discovery is, however, under challenge (5, 6).

The defining feature of collagen is an elegant structural motif in which three parallel polypeptide strands in a left-handed, polyproline II-type (PPII) helical conformation coil about each other with a one-residue stagger to form a right-handed triple helix (Figure 1). The tight packing of PPII helices within the triple helix mandates that every third residue be Gly, resulting in a repeating XaaYaaGly sequence, where Xaa and Yaa can be any amino acid. This repeat occurs in all types of collagen, although it is disrupted at certain locations within the triple-helical domain of nonfibrillar collagens (8). The amino acids in the Xaa and Yaa positions of collagen are often (2S)-proline (Pro, 28%) and (2S,4R)-4-hydroxyproline (Hyp, 38%), respectively. ProHypGly is the most common triplet (10.5%) in collagen (9). In animals, individual collagen triple helices, known as tropocollagen (TC), assemble in a complex, hierarchical manner that ultimately leads to the macroscopic fibers and networks observed in tissue, bone, and basement membranes (Figure 2).

The categories of collagen include the classical fibrillar and network-forming collagens, the FACTIs (fibril-associated collagens with interrupted triple helices), MACTIs (membrane-associated collagens with interrupted triple helices), and MULTIPLEXINs (multiple triple-helix domains and interruptions). Collagen types, their distribution, composition, and pathology are listed in Table 1. It is noteworthy that, although the three polypeptide chains in the triple helix of each collagen type can be identical, heterotrimeric triple helices are more prevalent than are homotrimeric triple helices.

STRUCTURE OF THE COLLAGEN TRIPLE HELIX
In 1940, Astbury & Bell (11) proposed that the collagen molecule consists of a single
extended polypeptide chain with all amide bonds in the cis conformation. A significant advance was achieved when, in the same 1951 issue of the Proceedings of the National Academy of Sciences in which they put forth the correct structures for the α-helix and β-sheet, Pauling & Corey (12) proposed a structure for collagen. In that structure, three polypeptide strands were held together in a helical conformation by hydrogen bonds. Within each amino acid triplet, those hydrogen bonds engaged four of the six main chain heteroatoms, and their formation required two of the three peptide bonds to be in the cis conformation. In 1954, Ramachandran & Kartha (13, 14) advanced a structure for the collagen triple helix on the basis of fiber diffraction data. Their structure was a right-handed triple helix of three staggered, left-handed PPII helices with all peptide bonds in the trans conformation and two hydrogen bonds within each triplet. In 1955, this structure was refined by Rich & Crick (15–16) and by North and coworkers (17) to the triple-helical structure accepted today, which has a single interstrand N–H(Gly)···O==C(Xaa) hydrogen bond per triplet and a tenfold helical symmetry with a 28.6-Å axial repeat (10/3 helical pitch) (Figure 1).

Fiber diffraction studies cannot reveal the structure of collagen at atomic resolution. Exacerbating this difficulty, the large size, insolubility, repetitive sequence, and complex hierarchical structure of native collagen thwart most biochemical and biophysical analyses. Hence, a reductionist approach using triple-helical, collagen-related peptides (CRPs) has been employed extensively since the late 1960s (18).

In 1994, Berman and coworkers (19) reported the first high-resolution crystal structure of a triple-helical CRP (Figure 1a). This structure confirmed the existence of interstrand N–H(Gly)···O= C(Xaa) hydrogen bonds (Figure 1c,d) and provided additional insights, including that C=O(Hyp/Gly)··· O= C(Xaa/Gly) hydrogen bonds could likewise stabilize the triple helix (20). Using CRPs and X-ray crystallography, the structural impact of a single Gly → Ala substitution was observed (19), the effects of neighboring charged residues in a triple helix were analyzed (21), and a snapshot of the interaction of a triple-helical CRP with the I domain of integrin α2β1 was obtained (Figure 3) (22).

Most X-ray crystallographic studies on CRPs have been performed on proline-rich collagenous sequences. All of the resulting structures have a 7/2 helical pitch (20.0-Å axial repeat), in contrast to the 10/3 helical pitch (28.6-Å axial repeat) predicted for natural collagen by fiber diffraction (17). On the basis of X-ray crystal structures of proline-rich CRPs: collagen-related peptide

\[
\text{CRP: collagen-related peptide}
\]
Figure 2
Biosynthetic route to collagen fibers (110), which are the major component of skin. Size and complexity are increased by posttranslational modifications and self-assembly. Oxidation of lysine side chains leads to the spontaneous formation of desmosine and isodesmosine cross-links.

CRPs, and in accordance with an early proposal regarding the helical pitch of natural triple helices (23), Okuyama and coworkers (24) postulated that the correct average helical pitch for natural collagen is 7/2. The generality of this hypothesis is unclear, as few regions of natural collagen are as proline rich as the CRPs analyzed by X-ray crystallography. The actual helical pitch of collagen likely varies across the domains and types of natural collagen.
Table 1  Vertebrate collagens

<table>
<thead>
<tr>
<th>Type</th>
<th>Class</th>
<th>Composition</th>
<th>Distribution</th>
<th>Pathology</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Fibrillar</td>
<td>$\alpha_1[1]:\alpha_2[1]$</td>
<td>Abundant and widespread: dermis, bone, tendon, ligament</td>
<td>OI, Ehlers–Danlos syndrome, osteoporosis</td>
</tr>
<tr>
<td>II</td>
<td>Fibrillar</td>
<td>$\alpha_1[II]$</td>
<td>Cartilage, vitreous</td>
<td>Osteoarthrosis, chondrodysplasias</td>
</tr>
<tr>
<td>III</td>
<td>Fibrillar</td>
<td>$\alpha_1[III]$</td>
<td>Skin, blood vessels, intestine</td>
<td>Ehlers-Danlos syndrome, arterial aneurysms</td>
</tr>
<tr>
<td>IV</td>
<td>Network</td>
<td>$\alpha_1[IV]:\alpha_2[IV]$</td>
<td>Basement membranes</td>
<td>Alport syndrome</td>
</tr>
<tr>
<td>V</td>
<td>Fibrillar</td>
<td>$\alpha_1[V]:\alpha_2[V]$</td>
<td>Widespread: bone, dermis, cornea, placenta</td>
<td>Ehlers-Danlos syndrome</td>
</tr>
<tr>
<td>VI</td>
<td>Network</td>
<td>$\alpha_1[VI]:\alpha_2[VI]$</td>
<td>Widespread: bone, cartilage, cornea, dermis</td>
<td>Bethlem myopathy</td>
</tr>
<tr>
<td>VII</td>
<td>Anchoring fibrils</td>
<td>$\alpha_1[VII]:\alpha_2[VII]$</td>
<td>Dermis, bladder</td>
<td>Epidermolysis bullosa acquisita</td>
</tr>
<tr>
<td>VIII</td>
<td>Network</td>
<td>$\alpha_1[VIII]:\alpha_2[VIII]$</td>
<td>Widespread: dermis, brain, heart, kidney</td>
<td>Fuchs endothelia corneal dystrophy</td>
</tr>
<tr>
<td>IX</td>
<td>FACIT</td>
<td>$\alpha_1[X]:\alpha_2[X]$</td>
<td>Cartilage, cornea, vitreous</td>
<td>Osteoarthrosis, multiple epiphyseal dysplasia</td>
</tr>
<tr>
<td>X</td>
<td>Network</td>
<td>$\alpha_1[X]$</td>
<td>Cartilage</td>
<td>Chondrodysplasia</td>
</tr>
<tr>
<td>XI</td>
<td>Fibrillar</td>
<td>$\alpha_1[XI]:\alpha_2[X]$</td>
<td>Cartilage, intervertebral disc</td>
<td>Chondrodysplasia, osteoarthrosis</td>
</tr>
<tr>
<td>XII</td>
<td>FACIT</td>
<td>$\alpha_1[XII]$</td>
<td>Dermis, tendon</td>
<td>—</td>
</tr>
<tr>
<td>XIII</td>
<td>MACIT</td>
<td>—</td>
<td>Endothelial cells, dermis, eye, heart</td>
<td>—</td>
</tr>
<tr>
<td>XIV</td>
<td>FACIT</td>
<td>$\alpha_1[XIV]$</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>XV</td>
<td>MULTIPLEXIN</td>
<td>—</td>
<td>Capillaries, testis, kidney, heart</td>
<td>—</td>
</tr>
<tr>
<td>XVI</td>
<td>FACIT</td>
<td>—</td>
<td>Dermis, kidney</td>
<td>—</td>
</tr>
<tr>
<td>XVII</td>
<td>MACIT</td>
<td>$\alpha_1[XVII]$</td>
<td>Hemidesmosomes in epithelia</td>
<td>Generalized atrophic epidermolysis bullosa</td>
</tr>
<tr>
<td>XVIII</td>
<td>MULTIPLEXIN</td>
<td>—</td>
<td>Basement membrane, liver</td>
<td>Knobloch syndrome</td>
</tr>
<tr>
<td>XIX</td>
<td>FACIT</td>
<td>—</td>
<td>Basement membrane</td>
<td>—</td>
</tr>
<tr>
<td>XX</td>
<td>FACIT</td>
<td>—</td>
<td>Cornea (chick)</td>
<td>—</td>
</tr>
<tr>
<td>XXI</td>
<td>FACIT</td>
<td>—</td>
<td>Stomach, kidney</td>
<td>—</td>
</tr>
<tr>
<td>XXII</td>
<td>FACIT</td>
<td>—</td>
<td>Tissue junctions</td>
<td>—</td>
</tr>
<tr>
<td>XXIII</td>
<td>MACIT</td>
<td>—</td>
<td>Heart, retina</td>
<td>—</td>
</tr>
<tr>
<td>XXIV</td>
<td>Fibrillar</td>
<td>—</td>
<td>Bone, cornea</td>
<td>—</td>
</tr>
<tr>
<td>XXV</td>
<td>MACIT</td>
<td>—</td>
<td>Brain, heart, testis</td>
<td>Amyloid formation?</td>
</tr>
<tr>
<td>XXVI</td>
<td>FACIT</td>
<td>—</td>
<td>Testis, ovary</td>
<td>—</td>
</tr>
<tr>
<td>XXVII</td>
<td>Fibrillar</td>
<td>—</td>
<td>Cartilage</td>
<td>—</td>
</tr>
<tr>
<td>XXVIII</td>
<td></td>
<td>—</td>
<td>Dermis, sciatic nerve</td>
<td>Neurodegenerative disease?</td>
</tr>
</tbody>
</table>

*Information is updated from References 1 and 32.
*Partial listing of tissues in which the relevant collagen type occurs.
*For a discussion of the role of specific collagen types in human disease, see Reference 32.
*Abbreviations: FACIT, fibril-associated collagen with interrupted triple helices; MACIT, membrane-associated collagen with interrupted triple helices; MULTIPLEXIN, multiple triple-helix domains and interruptions.
*Collagen XXVIII was reported in 2006 (2).
Specifically, the helical pitch could be 10/3 in proline-poor regions and 7/2 in proline-rich regions. This proposal is supported by the observation that proline-poor regions within crystalline CRPs occasionally display a 10/3 helical pitch (25, 26). Variability in the triple-helical pitch of native collagen could play a role in the interaction of collagenous domains with other biomolecules (22, 27–29).

UNDERSTANDING TRIPLE-HELIX STRUCTURE AND STABILITY

The vital importance of collagen as a scaffold for animals demands a manifold of essential characteristics. These characteristics include thermal stability, mechanical strength, and the ability to engage in specific interactions with...
other biomolecules. Understanding how such properties are derived from the fundamental structural unit of collagen—the triple helix—necessitates a comprehensive knowledge of the mechanisms underlying triple-helix structure and stability.

Interstrand Hydrogen Bonds

The ubiquity of collagen makes the ladder of recurrent $\text{N}^\alpha\text{H}^{\text{Gly}} - \text{O}^\beta\text{C}^\gamma\text{(Xaa)}$ hydrogen bonds that form within the triple helix (Figure 1c,d) the most abundant amide–amide hydrogen bond in kingdom Animalia. Replacing the Yaa–Gly amide bond with an ester in a host-guest CRP (Figure 4a,b) enabled estimation of the strength of each amide–amide hydrogen bond as $\Delta G^\text{c} = -2.0$ kcal/mol (30). Boryskina and coworkers (31) used a variety of other experimental techniques to assess this same parameter, estimating the strength of each amide–amide hydrogen bond within a poly(GlyProPro) CRP as $\Delta G^\text{c} = -1.8$ kcal/mol and within native collagen as $\Delta G^\text{c} = -1.4$ kcal/mol.

Glycine Substitutions

Numerous collagen-related diseases are associated with mutations in both triple-helical and nontriple-helical domains of various collagens (Table 1). These diseases have been reviewed in detail elsewhere (32) and are not discussed extensively herein.

The Gly residue in the XaaYaaGly repeat is invariant in natural collagen, and favorable substitutions are unknown in CRPs (33). A computational study suggested that replacing the obligate Gly residues of collagen with $\delta$-alanine or $\delta$-serine would stabilize the triple helix (34) and thus that the Gly residues in collagen are surrogates for nonnatural $\delta$-amino acids. Subsequent experimental data demonstrated, however, that this notion was erroneous (35).

Many of the most damaging mutations to collagen genes result in the replacement of a Gly residue within the triple helix (Figure 1c,d). Both the identity of the amino acid replacing Gly and the location of that substitution can impact the pathology of, for example, osteogenesis imperfecta (OI) (33, 36). substitutions for Gly in proline-rich portions of the collagen sequence (Figure 3a) are far less disruptive than those in proline-poor regions, a testament to the importance of Pro derivatives for triple-helix nucleation (37). In vivo, triple helices fold in a C-terminal→N-terminal manner (38). The time delay between disruption of triple-helix folding by a Gly substitution and renucleation of the folding process N-terminal to the substitution site is much shorter when
Protocollagen: the nonhydroxylated, nontriple-helical form of collagen prior to the action of P4H, P3H, lysyl hydroxylase, and protein disulfide isomerase

Preorganization: the extent to which hosts and guests are organized for binding prior to their complexation, thereby increasing complex stability

P4H: prolyl 4-hydroxylase

Prolines in the Xaa and Yaa Positions

In the strands of human collagen, ~22% of all residues are either Pro or Hyp (9). The abundance of these residues preorganizes the individual strands in a PPII conformation, thereby decreasing the entropic cost for collagen folding (40). Despite their stabilizing properties, Pro derivatives also have certain deleterious consequences for triple-helix folding and stability that partially offset their favorable effects. For example, Pro has a secondary amino group and forms tertiary amides within a peptide or protein. Tertiary amides have a significant population of both the trans and the cis isomers (Figure 5), whereas all peptide bonds in collagen are trans. Thus, before a (ProHypGly)$_n$ strand can fold into a triple helix, all the cis peptide bonds must isomerize to trans. N-Methylalanine (an acyclic, tertiary amide missing only $\gamma$ of Pro) decreases triple-helix stability when used to replace Pro or Hyp in CRPs, presumably because it lacks the preorganization imposed by the pyrrolidine ring of Pro derivatives (41). In contrast, avoiding the issue of cis-trans isomerization altogether by replacing a Gly–Pro amide bond with a trans-locked alkene isostere also results in a destabilized triple helix, despite leaving all interchain hydrogen bonds intact (42). Clearly, the factors dictating triple-helix structure and stability are intertwined in a complex manner (vide infra).

Pro residues in the Yaa position of protocollagen triplets are modified by prolyl 4-hydroxylase (P4H), a nonheme iron enzyme that catalyzes the posttranslational and stereoselective hydroxylation of the unactivated $\gamma$-carbon of Pro residues in the Yaa position of collagen sequences to form Hyp (Figure 6). P4H activity is required for the viability of
both the nematode *Caenorhabditis elegans* and the mouse *Mus musculus* (43, 44). Thus, Hyp is essential for the formation of sound collagen in vivo.

**Role of Hyp**

The hydroxylation of Pro residues in the Yaa position of collagen increases dramatically the thermal stability of triple helices (Table 2). This stabilization occurs when the resultant Hyp is in the Yaa position (45, 46) but not in the Xaa position, nor when the hydroxyl group is installed in the 4S configuration as in (2S,4S)-4-hydroxyproline (hyp) (Table 2) (47, 48). These findings led to the proposal that the 4R configuration of a prolyl hydroxyl group is privileged in alone enabling the formation of water-mediated hydrogen bonds that stitch together the folded triple helix (49). Indeed, such water bridges between Hyp and main chain heteroatoms were observed by Berman and coworkers (19, 50) in their seminal X-ray crystallographic studies of CRPs. The frequency of Hyp in most natural collagen is, however, too low to support an extensive network of water bridges. For example, four or more repeating triads of Xaa–Hyp–Gly occur only twice in the amino acid sequence of human type I collagen.

The hypothesis that the water bridges observed in crystalline (ProHypGly), triple helices are meaningful was tested by replacing Hyp residues in CRPs with (2S,4R)-4-fluoroproline (Flp). As fluoro groups do not form strong hydrogen bonds (51), water bridges cannot play a major role in stabilizing a (ProFlpGly)₁₀ triple helix. Nonetheless, (ProFlpGly)₁₀ triple helices are hyperstable (Table 2) (52, 53). Accordingly, water bridges cannot be of fundamental importance for triple-helix stability. How, then, does 4R-hydroxylation of Yaa-position Pro residues stabilize the triple helix?

**A gauche effect.** Replacing Hyp in the Yaa position with (2S,4S)-4-fluoroproline (flp), a diastereomer of Flp, prevents triple-helix formation (Table 2) (54). This discovery that the stereochemistry of electronegative substituents at the 4-position of the Pro ring is important for the formation of stable triple helices suggests that Flp and Hyp in the Yaa position stabilize collagen via a stereoelectronic effect, rather than a simple inductive effect (54). Pro and its derivatives prefer one of two major pyrrolidine ring puckers, which are termed C'–exo and C'–endo (Figure 7). The ring actually prefers two distinct twist, rather than envelope, conformations (55). As C’ experiences a large out-of-plane displacement in the twisted rings, we refer to pyrrolidine ring

<table>
<thead>
<tr>
<th>(XaaYaaGly)₁₀</th>
<th>T’ₘ (°C)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ProFlpGly)₁₀</td>
<td>45</td>
<td>(54)</td>
</tr>
<tr>
<td>(ProHypGly)₁₀</td>
<td>36</td>
<td>(54)</td>
</tr>
<tr>
<td>(mepMepGly)₁₀</td>
<td>36</td>
<td>(65)</td>
</tr>
<tr>
<td>(flpProGly)₁₀</td>
<td>33</td>
<td>(68)</td>
</tr>
<tr>
<td>(ProMepGly)₁₀</td>
<td>29</td>
<td>(65)</td>
</tr>
<tr>
<td>(ProClpGly)₁₀</td>
<td>23</td>
<td>(61)</td>
</tr>
<tr>
<td>(mepProGly)₁₀</td>
<td>13</td>
<td>(65)</td>
</tr>
<tr>
<td>(clpProGly)₁₀</td>
<td>No helix</td>
<td>(61)</td>
</tr>
<tr>
<td>(ProProGly)₁₀</td>
<td>No helix</td>
<td>(79)</td>
</tr>
<tr>
<td>(flpFlpGly)₁₀</td>
<td>No helix</td>
<td>(79)</td>
</tr>
<tr>
<td>(ProFlpGly)₁₀</td>
<td>No helix</td>
<td>(79)</td>
</tr>
<tr>
<td>(ProFlpGly)₁₀</td>
<td>No helix</td>
<td>(54)</td>
</tr>
<tr>
<td>(flpFlpGly)₁₀</td>
<td>No helix</td>
<td>(68)</td>
</tr>
<tr>
<td>(ProFlpGly)₁₀</td>
<td>91</td>
<td>(53)</td>
</tr>
<tr>
<td>(ProMepGly)₁₀</td>
<td>70</td>
<td>(59)</td>
</tr>
<tr>
<td>(HypHypGly)₁₀</td>
<td>65</td>
<td>(85)</td>
</tr>
<tr>
<td>(ProHypGly)₁₀</td>
<td>61–69</td>
<td>(53, 85)</td>
</tr>
<tr>
<td>(flpProGly)₁₀</td>
<td>58</td>
<td>(69)</td>
</tr>
<tr>
<td>(ProClpGly)₁₀</td>
<td>52</td>
<td>(61)</td>
</tr>
<tr>
<td>(clpProGly)₁₀</td>
<td>33</td>
<td>(61)</td>
</tr>
<tr>
<td>(ProProGly)₁₀</td>
<td>31–41</td>
<td>(53, 64)</td>
</tr>
<tr>
<td>(flpFlpGly)₁₀</td>
<td>30</td>
<td>(94)</td>
</tr>
<tr>
<td>(clpClpGly)₁₀</td>
<td>No helix</td>
<td>(61)</td>
</tr>
<tr>
<td>(HypProGly)₁₀</td>
<td>No helix</td>
<td>(84)</td>
</tr>
<tr>
<td>(ProHypGly)₁₀</td>
<td>No helix</td>
<td>(47)</td>
</tr>
<tr>
<td>(FlpProGly)₁₀</td>
<td>No helix</td>
<td>(69)</td>
</tr>
<tr>
<td>(ClpProGly)₁₀</td>
<td>No helix</td>
<td>(61)</td>
</tr>
<tr>
<td>(hypHypGly)₁₀</td>
<td>No helix</td>
<td>(47)</td>
</tr>
</tbody>
</table>

*Values of T’ₘ depend on both CRP concentration and heating rate. Hence, detailed comparisons require knowledge of experimental procedures.*
Figure 7

Ring conformations of Pro and Pro derivatives. The \( C^\gamma\)-endo conformation is favored strongly by stereo-eelectronic effects when \( R^1 = \text{H}, R^2 = \text{F} \) (flp) or \( \text{Cl} \) (clp), and by steric effects when \( R^1 = \text{Me} \) (mep) or \( \text{SH} \) (mcp), \( R^2 = \text{H} \). The \( C^\gamma\)-exo conformation is favored strongly by stereo-electronic effects when \( R^1 = \text{OH} \) (Hyp), \( \text{F} \) (Flp), \( \text{OMe} \) (Mop), or \( \text{Cl} \) (Clp), \( R^2 = \text{H} \), and by steric effects when \( R^1 = \text{H}, R^2 = \text{Me} \) (Mep) or \( \text{SH} \) (Mcp). The \( C^\gamma\)-endo:\( C^\gamma\)-exo ratio is \( \sim 2 \) when \( R^1 = R^2 = \text{H} \) (56).

Pro itself has a slight preference for the \( C^\gamma\)-endo ring pucker (Table 3) (56). A key attribute of a 4\( R \) fluoro group on Pro (as well as the natural 4\( R \) hydroxyl group) is its imposition of a \( C^\gamma\)-exo pucker on the pyrrolidine ring via the gauche effect (Figure 8a,b) (56–58).

The \( C^\gamma\)-exo ring pucker preorganizes the main chain torsion angles (\( \phi, C_{i-1}^\gamma-N_i^\gamma-C_i^\gamma-C_{i+1}^\gamma; \psi, N_i-C_{i}^\gamma-C_i^\gamma-N_{i+1}^\gamma; \) and \( \omega, C_i^\gamma-C_{i}^\gamma-N_{i+1}^\gamma-C_{i+1}^\gamma \)) to those in the Yaa position of a triple helix (Table 4). Thus, 4\( R \)-hydroxylation of Pro residues in the Yaa position of collagen stabilizes the triple helix via a stereo-electronic effect. Flp is more stabilizing than Hyp because fluorine (\( \chi_F = 4.0 \)) is more electronegative than oxygen (\( \chi_O = 3.5 \)), and a fluoro group

Table 3 Conformation of Pro and its 4-substituted derivatives that prefer the Xaa position \[ \phi = -75^\circ, \ psi = 164^\circ (7) \] in a collagen triple helix

<table>
<thead>
<tr>
<th>Residue (References)</th>
<th>Crystal structure</th>
<th>Ring pucker</th>
<th>( (E_{\text{endo}}-E_{\text{exo}}) ) (kcal/mol)</th>
<th>( \phi ) (degrees)</th>
<th>( \psi ) (degrees)</th>
<th>( K_{\text{trans/cis}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro (54, 56, 57)</td>
<td><img src="https://example.com/pro_structure.png" alt="Pro structure" /></td>
<td>( C^\gamma)-endo</td>
<td>-0.41</td>
<td>-79\textsuperscript{b}</td>
<td>177\textsuperscript{b}</td>
<td>4.6</td>
</tr>
<tr>
<td>mcp (66)</td>
<td><img src="https://example.com/mcp_structure.png" alt="mcp structure" /></td>
<td>( C^\gamma)-endo</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>4.7</td>
</tr>
<tr>
<td>mep (65)</td>
<td><img src="https://example.com/mep_structure.png" alt="mep structure" /></td>
<td>( C^\gamma)-endo</td>
<td>-1.4</td>
<td>—</td>
<td>—</td>
<td>3.7</td>
</tr>
<tr>
<td>flp (54, 56, 68)</td>
<td><img src="https://example.com/flp_structure.png" alt="flp structure" /></td>
<td>( C^\gamma)-endo</td>
<td>-0.61</td>
<td>-76\textsuperscript{b}</td>
<td>172\textsuperscript{a}</td>
<td>2.5</td>
</tr>
<tr>
<td>clp (61)</td>
<td><img src="https://example.com/clp_structure.png" alt="clp structure" /></td>
<td>( C^\gamma)-endo</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2.2</td>
</tr>
</tbody>
</table>

\textsuperscript{a}From DFT calculations.

\textsuperscript{b}Values of \( \phi \) and \( \psi \) (here, \( N_i-C_{i-1}^\gamma-C_i^\gamma-O_{i+1} \)) are from the crystal structure of Ac-Pro-OMe, which has a cis peptide bond.

\textsuperscript{c}Values of \( K_{\text{trans/cis}} \) (Figure 5) were determined in solution by NMR spectroscopy.
Thus, a 4GlyNH₂ (62). In contrast, a host-guest CRP than does a 4 ring pucker of a Pro derivative more strongly stability, a (2

S

,4R)-4-methoxyproline residue (Mop) was incorporated into the Yaa position of a (ProYaaGly)₁₀ CRP (59). O-Methylation is perhaps the simplest possible covalent modification of a Hyp residue and reduces the extent of hydration without altering significantly the electron-withdrawing ability of the 4R substituent. Accordingly, Mop and Hyp residues have similar conformations (Table 4). Interestingly, reducing the hydration of (ProHypGly)₁₀ by methylation of Hyp residues enhances triple-helix stability significantly (Table 2). Moreover, alkylation with functional groups larger than a methyl group does not necessarily perturb triple-helix stability (60). Notably, (2S,4R)-4-chloroproline (Clp) residues also stabilize triple helices in the Yaa position (Table 2) (61). Like Flp, Clp has a strong preference for the C₁-exo ring pucker, and a (ProClpGly)₁₀ triple helix is therefore more stable than a (ProProGly)₁₀ triple helix. Thus, a plethora of data indicate that the hydroxyl group of Hyp stabilizes collagen through a stereoelectronic effect. Water bridges provide little (if any) net thermodynamic advantage to natural collagen (59).

Surprisingly, a host-guest CRP of the form AcGly–(ProHypGly)₄–ProFlpGly–(ProHypGly)₄–GlyNH₂ actually forms a less stable triple helix than does AcGly–(ProHypGly)₄–GlyNH₂ (62). In contrast, a host-guest CRP of the form (GlyProHyp)₄–GlyProFlp–GlyValCys–GlyAspLys–GlyAsnPro–GlyTrpPro–GlyAlaPro–(GlyProHyp)₄–NH₂ forms a more stable triple helix than one containing Hyp rather than Flp (63). These results suggest that a fluoro group might disrupt the hydration induced by a long string of Hyp residues. Kobayashi and coworkers (64) used differential scanning calorimetry to demonstrate that (ProHypGly)₁₀ triple helices are stabilized by enthalpy, whereas (ProFlpGly)₁₀ triple helices are stabilized by entropy. These findings are consistent with Hyp decreasing the entropic cost for folding via main chain preorganization but increasing that cost by specific hydration. This interpretation is in accord with the stability of (ProMopGly)₁₀ triple helices arising from a nearly equal contribution of enthalpy and entropy (59).

**Figure 8**

Stereoelectronic effects that stabilize the collagen triple helix. (a) A gauche effect and an n→π* interaction preorganize main chain torsion angles and enhance triple-helix stability. (b) A gauche effect, elicited by an electron-withdrawing group (EWG) in the 4R position, stabilizes the C₁-exo ring pucker. (c) An n→π* interaction stabilizes the trans isomer of the peptide bond but is substantial only when Pro derivatives are in the C₁-exo ring pucker (e.g., R₁ = OH or F, R₂ = H). (d) Depiction of overlap between n and π* natural bond orbitals (NBO[π*]) in a Pro residue with C₁-exo pucker.

A steric effect. Electronegative substituents on Pro rings are not the only means of enforcing an advantageous ring pucker. Pro ring pucker can also be dictated by steric effects, as in (2S,4S)-4-methylproline (Mep) (65) and (2S,4R)-4-mercaptoproline (Mpc) (Figure 7) (66). The 4-methyl substituent of Mep prefers the pseudoequatorial orientation and thus enforces the C₁-exo ring pucker of Pro (analogous results are observed for Mpc) (Table 4).
Indeed, triple helices formed from (ProMepGly)₇ have stability similar to those formed from (ProHypGly)₇ (Table 2) (65).

### Proline Derivatives in the Xaa Position

The C⁴-endo ring pucker of Pro residues in the Yaa position enhances triple-helix stability. Likewise, the ring pucker of Pro in the Xaa position is important for triple-helix stability. Typically, Pro residues in the Xaa position of biological collagen are not hydroxylated and usually display the C⁴-endo ring pucker (67). By employing C⁴-substituents, both the gauche effect and steric effects can be availed to preorganize the C⁴-endo ring pucker (Figures 7 and 8). Installation of flp, (2S,4S)-4-chloroproline (clp), or (2S,4R)-4-methylproline (mep) residues (all of which prefer the C⁴-endo ring pucker) (Table 3) in the Xaa position of collagen is stabilizing relative to Pro, but installation of Flp, Clp, or Hyp (which prefer the C⁴-endo ring pucker) is destabilizing (Table 2) (61, 65, 68–70). These results suggest that preorganizing the C⁴-endo ring pucker in the Xaa position of CRPs stabilizes triple helices. This conclusion is reasonable because Pro derivatives with a C⁴-endo ring pucker have φ and ψ main chain torsion angles similar to those observed in the Xaa position of triple helices (Table 3).

#### Table 4 Conformation of 4-substituted derivatives of Pro that prefer the Yaa position [φ = −60°, ψ = 152° (7)] in a collagen triple helix

<table>
<thead>
<tr>
<th>Residue</th>
<th>Crystal structure</th>
<th>Ring pucker</th>
<th>(E&lt;sub&gt;endo&lt;/sub&gt; − E&lt;sub&gt;exo&lt;/sub&gt;) &lt;sup&gt;a&lt;/sup&gt; (kcal/mol)</th>
<th>φ&lt;sup&gt;b&lt;/sup&gt; (degrees)</th>
<th>ψ&lt;sup&gt;b&lt;/sup&gt; (degrees)</th>
<th>K&lt;sub&gt;trans/cis&lt;/sub&gt; &lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mep (65)</td>
<td><img src="image" alt="Mep" /></td>
<td>C¹-exo</td>
<td>1.7</td>
<td>−62&lt;sup&gt;d&lt;/sup&gt;</td>
<td>153&lt;sup&gt;d&lt;/sup&gt;</td>
<td>7.4</td>
</tr>
<tr>
<td>Mop (59)</td>
<td><img src="image" alt="Mop" /></td>
<td>C¹-exo</td>
<td>—</td>
<td>−58</td>
<td>148</td>
<td>6.7</td>
</tr>
<tr>
<td>Flp (52, 54, 56, 57)</td>
<td><img src="image" alt="Flp" /></td>
<td>C¹-exo</td>
<td>0.85</td>
<td>−55</td>
<td>141</td>
<td>6.7</td>
</tr>
<tr>
<td>Hyp (46, 54, 57, 58)</td>
<td><img src="image" alt="Hyp" /></td>
<td>C¹-exo</td>
<td>—</td>
<td>−57</td>
<td>151</td>
<td>6.1</td>
</tr>
<tr>
<td>Clp (61)</td>
<td><img src="image" alt="Clp" /></td>
<td>C¹-exo</td>
<td>—</td>
<td>−56</td>
<td>148</td>
<td>5.4</td>
</tr>
<tr>
<td>Mcp (66)</td>
<td><img src="image" alt="Mcp" /></td>
<td>C¹-exo</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>5.4</td>
</tr>
</tbody>
</table>

<sup>a</sup>From DFT calculations.

<sup>b</sup>Values of φ and ψ (here, Nᵢ-Cᵢ-Cᵢ-Cᵢ₊₁) are from the crystal structure of Ac-Yaa-OMe.

<sup>c</sup>Values of K<sub>trans/cis</sub> (Figure 5) were determined in solution by NMR spectroscopy.

<sup>d</sup>M.D. Shoulders, I.A. Guzei & R.T. Raines, unpublished data.
Notably, replacing Pro in the Xaa position of (ProProGly)$_7$ with hyp, a Pro derivative that, like flp and clp, should prefer the $C^i$-endo ring pucker owing to the gauche effect, yields CRPs that do not form triple helices (Table 2) (47). This anomalous result for hyp in the Xaa position could be attributable to deleterious hydration, idiosyncratic conformational preferences of hyp residues, or both (71).

Type IV collagen, which is the primary component of basement membranes, has a high incidence of (2S,3S)-3-hydroxyproline (3S-Hyp) in the Xaa position (72). This modification is present in some other collagen types and in invertebrate collagens. 3S-Hyp, which prefers a $C^i$-endo ring pucker (73), is introduced almost exclusively within ProHypGly triplets via post-translational modification of individual collagen strands by prolyl 3-hydroxylase (P3H), which is distinct from P4H (74). A recessive form of OI is associated with a P3H deficiency (75, 76). Certain mutations to the gene encoding cartilage-associated protein, a P3H-helper protein, prevent 3S-hydroxylation of $\alpha_1$(I)Pro986 as well as 3S-hydroxylation of some other Xaa-position Pro residues, resulting in a phenotype nearly identical to classical OI. The underlying basis for the importance of 3S-hydroxylation of $\alpha_1$(I)Pro986 is unclear but could involve lower rates of triple-helix secretion (76). Replacing Pro with 3S-Hyp in the Xaa position of CRPs can enhance triple-helix stability slightly (73, 77). A crystal structure of a triple helix containing 3S-Hyp substitutions reveals the maintenance of the prototypical triple-helix structure and the absence of unfavorable steric interactions (Figure 4c) (78). In contrast, replacing 3S-Hyp with (2S,3S)-3-fluoroproline destabilizes a triple helix markedly, possibly owing to a through-bond inductive effect that diminishes the ability of its main chain oxygen to accept a hydrogen bond (Figure 4d) (79).

An n→π* Interaction

A general principle in the design of CRPs is that Pro residues with either a $C^i$-endo or $C^i$-exo ring pucker will stabilize triple helices in the Xaa and Yaa positions, respectively (Tables 2–4). Appropriate ring pucker, enforced by a stereoelectronic or steric effect, preorganizes the $\phi$ and $\psi$ torsion angles to those required for triple-helix formation.

Intriguingly, the stability of a (flpProGly)$_7$ or (clpProGly)$_7$ triple helix is significantly less than that of a (ProFlpGly)$_7$ or (ProClpGly)$_7$ triple helix, respectively (Table 2) (61, 68). Likewise, a (mepProGly)$_7$ triple helix is less stable than a (ProMepGly)$_7$ triple helix (Table 2) (65). Two factors contribute to the lower stability of triple helices formed from CRPs with stabilizing Pro derivatives substituted in the Xaa position rather than the Yaa position. First, a $C^i$-endo ring pucker is already favored in Pro (56); flp, clp, and mep merely enhance that preference (Table 3). In contrast, Flp, Clp, Hyp, and Mep have the more dramatic effect of reversing the preferred ring pucker of Pro, thereby alleviating the entropic penalty for triple-helix formation to a greater extent (Table 4). Second, Flp, Clp, and Mep in the Yaa position cause favorable preorganization of all three main chain torsion angles ($\phi$, $\psi$, and $\omega$) (Table 4). In contrast, flp, clp, and mep have a low probability of adopting a trans peptide bond ($\omega = 180^\circ$) (54, 61, 65) relative to Pro (Table 3), thereby mitigating the benefit accrued from proper preorganization of $\phi$ and $\psi$.

Notably, $^{13}$C-NMR studies on collagen in vitro show that 16% of Gly–Pro bonds in unfolded collagen are in the cis conformation, whereas only 8% of Xaa–Hyp bonds in unfolded collagen are cis, an observation that confirms the effect of $C^i$-substitution on the conformation of the preceding peptide bond (80).

How does the effect of a 4-X substituent on Pro ring pucker influence the peptide bond isomerization equilibrium constant ($K_{trans/cis}$) (Figure 5 and Tables 3 and 4)? The explanation stems from another stereoelectronic effect: an n→π* interaction (56, 81). In an n→π* interaction, the oxygen of a peptide bond ($O_{i-1}$) donates electron density from its lone pairs into the antibonding orbital of the carbonyl in the subsequent peptide bond ($C^i$═O).
The $\gamma$-exo ring pucker of a Pro residue provides a more favorable $O_{\cdots} \cdots C_{\gamma}=\cdots O_{\cdots}$ distance and angle for an $n\to\pi^*$ interaction than does the $\gamma$-endo pucker (56). Importantly, $K_{\text{trans/\alpha}}$ for peptidyl prolyl amide bonds is determined by the pyrrolidine ring pucker and is not generally affected by the identity of substituents in the 4-position of the pyrrolidine ring (82). Because an $n\to\pi^*$ interaction can occur only if the peptide bond containing $O_{\cdots}{\gamma}$ is trans, the $n\to\pi^*$ interaction has an impact on the value of $K_{\text{trans/\alpha}}$ for main chains with appropriate torsion angles (Table 4). Thus, imposing a $\gamma$-exo pucker on a pyrrolidine ring in the Xaa position of a CRP preorganizes not only the $\phi$ and $\psi$ angles for triple-helix formation, but also the $\omega$ angle. Indeed, a single $n\to\pi^*$ interaction can stabilize the trans conformation by $\Delta G^* = -0.7$ kcal/mol (81, 83).

### Hyp in the Xaa Position

In the Xaa position, a Pro residue with a $\gamma$-endo pucker generally stabilizes a triple helix, whereas one with a $\gamma$-exo pucker destabilizes a triple helix. For example, $(\text{HypProGly})_n$ triple helices are far less stable than $(\text{ProProGly})_n$ triple helices (Table 2) (84) because Hyp prefers the $\gamma$-exo ring pucker and thus preorganizes the $\phi$ and $\psi$ torsion angles improperly for the Xaa position of a collagen triple helix (Table 4). Surprisingly, $(\text{HypHypGly})_{10}$ triple helices are actually slightly more stable than $(\text{ProHypGly})_{10}$ triple helices (Table 2) (85, 86) despite the Hyp residues in the Xaa position of $(\text{HypHypGly})_{10}$ displaying the $\gamma$-exo ring pucker in the triple helix (87, 88). It is noteworthy that crystal structures of $(\text{HypHypGly})_{10}$ show that the main chain torsion angles in the Xaa position of a $(\text{HypHypGly})_{10}$ triple helix adjust to accommodate a $\gamma$-exo ring pucker in that position (87, 88).

The finding that Hyp can stabilize triple helices in the Xaa position in a context-dependent manner was presaged in a study by Gruskin and coworkers (89) on the global substitution of Hyp for Pro in recombinant type I collagen polypeptides that formed stable triple helices. Notably, Hyp is found in the Xaa position of some invertebrate collagens (90) and can be acceptable in CRPs in which the Yaa position residue is not Pro (86, 91, 92). Berisio and coworkers (93) have suggested that $(\text{HypHypGly})_{10}$ triple helices might be hyperstable owing to intrastrand dipole-dipole interactions between proximal $\gamma$-OH bonds of adjacent Hyp residues. Kobayashi and coworkers (87) have proposed that the stability of $(\text{HypHypGly})_{10}$ triple helices is attributable to the high hydration level of the peptide chains in the single-coil state prior to triple-helix formation, which could reduce the entropic cost of water bridge formation. A combination of these factors is likely to be responsible for this anomaly.

### Heterotrimeric Synthetic Triple Helices

Both flp and Flp greatly enhance triple-helix stability when in the Xaa and Yaa position, respectively. Nonetheless, $(\text{flpFlpGly})_n$ forms much less stable triple helices than does $(\text{ProProGly})_n$ (Table 2) (79, 94). In such a helix, the fluorine atoms of flp and Flp residues in alternating strands would be proximal, and the C-F dipoles would interact unfavorably (Figure 9a) (79). These negative steric and electronic interactions presumably compromise triple-helix stability despite appropriate preorganization of main chain torsion angles. This hypothesis was confirmed by two other findings. First, a $(\text{clpClpGly})_{10}$ triple helix does not even form at 40°C, whereas a $(\text{flpFlpGly})_{10}$ triple helix has $T_m = 30°C$ (Table 2) (61, 94). The steric clash between chlorine atoms of opposing clp and Clp residues is exacerbated by the large size of chlorine relative to fluorine (Figure 9b). Second, $(\text{mepMepGly})_n$ forms more stable triple helices than do either of the corresponding mono-substituted CRPs, $(\text{mepProGly})_n$ and $(\text{ProMepGly})_n$ (Table 2). The 4-methyl groups protrude radially from the triple helix (Figure 9c) and thus cannot interact detrimentally with each other (65).
The steric and stereoelectronic effects on triple-helix stability manifested in the (flpFlpGly)7:CRP provided, for the first time, a means to generate noncovalently linked, heterotrimeric triple helices with defined stoichiometry. Analysis of triple-helix cross sections suggested a triple helix composed of (flpFlpGly)7:(ProProGly)7 in either a 1:2 or 2:1 ratio could be stable, as the presence of some Pro residues in the Xaa and Yaa positions would eliminate deleterious steric interactions between fluorine residues in opposing strands. A (flpFlpGly)7:(ProProGly)7 ratio of 2:1 yielded the most stable triple helices, thereby demonstrating the first instance of heterotrimeric assembly of triple helices with controlled stoichiometry (79) and suggesting the possibility of developing a “code” for triple-helix assembly along the lines of the Watson-Crick code for DNA assembly.

Gauba & Hartgerink (95) developed an alternative strategy that employs Coulombic interactions to guide the assembly of heterotrimeric triple helices. They observed that a 1:1:1 mixture of (ProArgGly)10:(GluHypGly)10:(ProHypGly)10 produces triple helices containing one negatively charged, one positively charged, and one neutral CRP. Intriguingly, a (ProLysGly)10:(AspHypGly)10:(ProHypGly)10 triple helix has...
Self-assembly: a process in which specific, local interactions between disordered components lead to an organized structure, without external direction.

Collagen fibrillogenesis: the process of tropocollagen monomers assembling into mature fibrils.

**AVOIDING AGGREGATION**

Long, unfolded polypeptides have an innate tendency to form aggregates (145), such as the amyloid fibrils implicated in neurodegenerative diseases. Interestingly, despite their long length and slow folding, protocollagen strands are not known to aggregate. Amyloid fibrils and other aggregates are composed largely of β-sheets (146). Pro and Gly are the two amino acid residues with the lowest propensity to form a β-sheet (147, 148), and Gly residues are known explicitly to reduce protein aggregation rates (149).

We propose that the prevalence of Pro and Gly residues in protocollagen is necessary to avert the formation of harmful aggregates. This proposal is supported by the remarkably high Pro/Gly content of other fibrous, structural proteins in plants and animals, such as elastin, extensin, glycine-rich proteins, and proline-rich proteins. Molecular dynamics simulations of elastin polypeptides likewise support this proposal, as a minimum threshold of Pro/Gly content must be attained to realize elastomers instead of amyloid fibrils (150). Apparently, the molecular evolution of collagen and other fibrous, structural proteins has availed Pro and Gly residues to avoid β-sheet formation and the consequent formation of harmful aggregates.

**Nonproline Substitutions in the Xaa and Yaa Positions**

Brodsky and coworkers (9) determined the frequency of occurrence of all possible tripeptides in a set of fibrillar and nonfibrillar collagen sequences. Only a few of the 400 possible triplets formed from the 20 natural amino acids are observed with any frequency in collagen. Additionally, they have examined exhaustively the incorporation of all 20 common amino acids in the Xaa and Yaa positions of CRPs using a host-guest model system wherein a single XaaYaaGly triplet is placed within a (ProProGly)\textsubscript{n} or (ProHypGly)\textsubscript{n} CRP (98). These host-guest studies revealed a correlation between the propensity of a particular residue to adopt a PPII conformation and its contribution to triple-helix stability (98). Notably, Arg in the Yaa position confers triple-helix stability similar to Hyp (99). The aromatic amino acid residues Trp, Phe, and Tyr are all strongly destabilizing to the triple helix (98), although the structural basis for this destabilization is unclear. Brodsky and coworkers (100) used their data on host-guest CRPs to develop an algorithm that enables a priori calculation of the effect of Xaa and Yaa substitutions on triple-helix stability.

**HIGHER-ORDER COLLAGEN STRUCTURE**

In vivo collagen has a hierarchical structure (Figure 2). Individual TC monomers self-assemble into the macromolecular fibers that are essential components of tissues and bones. The self-assembly processes involved in collagen fibrillogenesis are of enormous importance to ECM pathology and proper animal development (see the sidebar “Avoiding Aggregation” for a discussion of how collagen self-assembly might be directed away from deleterious protein aggregates).

**Fibril Structure**

There are many classes of collagenous structures in the ECM, including fibrils, networks,
and transmembrane collagenous domains. For brevity, we focus here on fibrils composed primarily of type I collagen.

TC monomers of type I collagen have the unique property of actually being unstable at body temperature (101); that is, the random coil conformation is the preferred one. How can stable tissue structures form from an unstable protein? The answer must be that collagen fibrillogenesis has a stabilizing effect on triple helices. Moreover, the assembly of strong macromolecular structures is essential to enable collagen to support stress in one, two, and three dimensions (102). The importance of collagen fibrillogenesis is underscored by the conclusion of Kadler and coworkers (103) that the fundamental principles underlying the formation of some types of modern collagen fibrils were established at least 500 Mya.

Collagen fibrillogenesis in situ occurs via assembly of intermediate-sized fibril segments, called microfibrils (Figure 2) (104). Thus, there are two important issues for understanding the molecular structure of the collagen fibril. First, what is the arrangement of individual TC monomers within the microfibril? Second, what is the arrangement of the individual microfibrils within the collagen fibril? These questions have proven difficult to answer, as individual natural microfibrils are not isolable and the large size and insolubility of mature collagen fibrils prevent the use of standard structure-determination techniques.

Collagen fibrils formed mainly from type I collagen (all fibrous tissues except cartilage) and fibrils formed largely from type II collagen (cartilage) have slightly different structures. Although we focus solely on type I collagen fibrils, recent data have enabled the determination of thin cartilage fibril structure to intermediate resolution (~4 nm). This structure suggests that cartilage collagen fibrils have a 10 + 4 heterotypic microfibril structure—meaning that the fibril surface presents ten equally spaced microfibrils and that there are four equally spaced microfibrils in the core of the fibril (105).

Fibrils of type I collagen in tendon are up to 1 cm in length (106) and up to ~500 nm in diameter. An individual triple helix in type I collagen is ~2 nm in diameter and ~500 nm long. Clearly, fibrillogenesis on an extraordinary scale is necessary to achieve the structural dimensions of natural collagen fibrils. The most characteristic feature of collagen fibrils is that they are D-periodic with $D_{\parallel} = 67$ nm. The banded structure observed in transmission electron microscopy (TEM) images of collagen fibrils occurs because the actual length of a TC monomer is not an exact multiple of $D$ but $L = 4.46D_\parallel$, resulting in gaps of 0.54$D_\parallel$ and overlaps of 0.46$D_\parallel$ (Figure 2). This regular array of gap and overlap regions must be accounted for in structural models of the collagen fibril and microfibril.

The initial proposal for the three-dimensional structure of fibrillar collagen was a simplified structural model for collagen microfibrils advanced by Hodge & Petruska (107) in 1963. Their model consists of a two-dimensional stack in which five TC monomers within a microfibril are offset by $D = 67$ nm between neighboring strands (Figure 2). This model accounts for the gap and overlap regions apparent in mature collagen fibrils by TEM and atomic force microscopy (AFM). Many research groups began efforts to determine the three-dimensional structure of type I collagen fibrils at higher resolution. Numerous models were proposed to account for the features of fiber diffraction and of TEM and AFM images of such fibrils (108–111). Researchers generally agreed on a quasi-hexagonal unit cell containing five TC monomers as the basis for an accurate model of the collagen fibril, but important details were in dispute. Recent findings indicate that the fibril structure controversy is approaching resolution.

In 2001, Orgel and coworkers (112, 113) reported the first electron-density map of a type I collagen fiber at molecular anisotropic resolution (axial: 5.16 Å; lateral: 11.1 Å) using synchrotron radiation. Their data confirm that collagen microfibrils have a quasi-hexagonal unit cell. The molecular packing of the TC monomers in this model results in TC neighbors arranged to form supertwisted,
Collagen telopeptides: N- and C-terminal 11- to 26-residue nontriple-helical domains of tropocollagen strands involved in fibrillogenesis and cross-linking

Procollagen: the hydroxylated form of collagen prior to collagen propeptide cleavage

Collagen propeptides: N- and C-terminal nontriple-helical domains of collagen strands that direct triple-helix folding prior to fibrillogenesis

Collagen fibrillogenesis requires completion of two stages of self-assembly: nucleation and fiber growth. Collagen fibrillogenesis begins only after procollagen N- and C-propeptidases cleave the collagen propeptides at each triple-helix terminus to generate TC monomers. The C-terminal propeptides are essential for proper triple-helix formation but prevent fibrillogenesis (116). After cleavage of the propeptides, TC monomers are composed of a lengthy triple-helical domain consisting of a repeating XaaYaaGly sequence flanked by short, nontriple-helical telopeptides (Figure 2).

The C-terminal telopeptides of TC are important for initiating proper fibrillogenesis. Prockop and Fertala (117) suggested that collagen self-assembly into fibrils is driven by the interaction of C-terminal telopeptides with specific binding sites on triple-helical monomers. The addition of synthetic telopeptide mimics can inhibit collagen fibrillogenesis, presumably by preventing the interaction between collagen telopeptides and TC monomers. Triple helices lacking the telopeptides can, however, assemble into fibrils with proper morphology (118). Thus, collagen telopeptides could accelerate fibril assembly and establish the proper register within microfibrils and fibrils but might not be essential for fibrillogenesis.

Collagen telopeptides have a second role in stabilizing mature collagen fibrils. Lys side chains in the telopeptides are cross-linked subsequent to fibril assembly, forming desmosine and isodesmosine cross-links between Lys and hydroxylysine residues with the aid of lysyl oxidase (Figure 2) (119). The cross-linking process endows mature collagen fibrils with strength and stability, but is not involved in fibrillogenesis. Thus, although collagen telopeptides might not be essential for nucleating collagen fibrillogenesis, their absence greatly weakens the mature fibril owing to the lack of cross-links within and between triple helices (119).

Nucleation and Modulation of Collagen Fibrillogenesis

Mechanical Properties of Collagen Fibrils

The hierarchical nature of collagen structure theoretically enables evaluation of the mechanical properties of collagen at varying levels of structural complexity, including the TC monomer, individual collagen fibrils, and collagen fibers. Perhaps the most direct measures of the mechanical properties of collagen have been obtained by studying TC monomers and fibrils formed from type I collagen. Researchers have employed various biophysical and theoretical techniques over the past 20 years, and
recent advances in AFM methodology have enabled more refined evaluations.

In 2006, Buehler estimated the fracture strength of a TC monomer to be 11 GPa, which is significantly greater than that of a collagen fibril (0.5 GPa) (102). This difference is reasonable, given that fracture of a TC monomer requires unraveling of the triple helix and ultimately breaking of covalent bonds, whereas fracture of a fibril does not necessarily require the disruption of covalent bonds. For comparison, the tensile strength of collagen in tendon is estimated to be 100 MPa (120).

The Young’s modulus of a TC monomer is $E = 6–7$ GPa (102, 121), whereas AFM measurements show that dehydrated fibrils of type I collagen from bovine Achilles tendon (122) and rat tail tendon (123) have $E \approx 5$ GPa and $E \leq 11$ GPa, respectively. Because collagen fibrils are anisotropic, the shear modulus (which is a measure of rigidity) is also an important measure of the strength of a collagen fibril. In 2008, AFM revealed that dehydrated fibrils of type I collagen from bovine Achilles tendon have $G = 33$ MPa (124). Hydration of these fibrils reduced their shear modulus significantly, whereas carbodiimide-mediated cross-linking increased their shear modulus. It is noteworthy that a certain level of cross-linking is favorable for the mechanical properties of collagen fibrils, but excessive cross-linking results in extremely brittle collagen fibrils (102), a common symptom of aging.

An analysis by Buehler (102) of the mechanical properties of collagen fibrils suggests that nature has selected a length for the TC monomer that maximizes the robustness of the assembled collagen fibril via efficient energy dissipation. Simulations indicate that TC monomers either longer or shorter than $\sim 300$ nm (which is the length of a type I collagen triple helix) would form collagen fibrils with less favorable mechanical properties.

### COLLAGENOUS BIOMATERIALS

Research on the structure and stability of collagen triple helices has focused on blunt-ended triple helices composed of $(XaaYaaGly)_{n \leq 10}$ CRPs. These short triple helices, although valuable for studies directed at understanding the physicochemical basis of triple-helix structure and stability, are not useful for many potential biomaterial applications because of their small size, which does not approach the scale of natural collagen fibers (Figure 2).

Bovine collagen is readily available and useful for some biomedical purposes, but it suffers from heterogeneity, potential immunogenicity, and loss of structural integrity during the isolation process. An efficient recombinant or synthetic source of collagen could avoid these complications. The heterologous production of collagen is made problematic by the difficulty of incorporating posttranslational modifications, such as that leading to the essential Hyp residues (Figure 6), and by the need to use complex expression systems (125). These challenges underscore the need for synthetic sources of collagen-like proteins and fibrils.

### Collagen via Chemical Synthesis

Early approaches to long synthetic collagen triple helices relied on the condensation (126, 127) or native chemical ligation of short CRPs (127). Interestingly, concentrated aqueous solutions of $(\text{ProHypGly})_{10}$ self-assemble into highly branched fibrils (128). Brodsky and coworkers (129) have shown that the rate of $(\text{ProHypGly})_{10}$ self-assembly and the morphology of the resultant fibrils are sequence dependent. CRPs containing a single $\text{Pro} \rightarrow \text{Ala}$ or $\text{Pro} \rightarrow \text{Leu}$ substitution display slower self-assembly; fibril morphology can be modified by a Gly $\rightarrow$ Ser substitution, or prevented by a single Gly $\rightarrow$ Ala substitution or global Hyp $\rightarrow$ Pro substitutions. Regardless, the higher-order structures formed by the self-assembly of $(\text{ProHypGly})_{10}$ and related CRPs do not resemble natural collagen fibrils.

Long collagen triple helices have been prepared by using a design that takes advantage of the intrinsic propensity of individual CRP strands to form triple helices. Specifically, a
cystine knot within short collagen fragments was utilized to set the register of individual collagen strands such that short, “sticky” ends preorganized for further triple-helix formation were displayed at the end of each triple-helical, monomeric segment (Figure 10a) (130, 131). Self-assembly of these short, triple-helix fragments was then mediated by association of the sticky ends, resulting in collagen assemblies as long as 400 nm—significantly longer than natural TC monomers (131). Koide and coworkers (132) used this system to prepare tunable collagen-like gels with potential biomaterial applications.

Maryanoff and coworkers (133) developed another approach to long triple helices, one that relied on the predilection of electron-rich phenyl rings of C-terminal phenylalanine residues installed in a short CRP to engage in π-stacking interactions preferentially with electron-poor pentafluorophenyl rings of N-terminal pentafluorophenylalanine residues (Figure 10b). Their strategy produced micrometer-scale triple-helical fibers. This π-stacking approach has been used to generate thrombogenic collagen-like fibrils for applications in biomedicine (134). In addition, attachment of gold nanoparticles to these fibrils and subsequent electroless silver plating yielded collagen-based nanowires that conduct electricity (135).

Przybyla & Chmielewski (136) used metal-triggered self-assembly to obtain collagen fibrils from a CRP. A single Hyp residue in Ac-(ProHypGly)9-NH2 was replaced with a bipyridyl-modified Lys residue. Addition of Fe(II) to a solution of this CRP triggered self-assembly into morphologically diverse fibrils of up to 5 μm in length with a mean radius of 0.5 μm.

A major advance in the development of synthetic CRP assemblies with improved similarity to collagen fibrils was reported by Chaikof and coworkers (137). They synthesized a CRP with the sequence (ProArgGly)4–(ProHypGly)4–(GluHypGly)4 and observed self-assembly in solution into fibrils 3–4 μm in length and 12–15 nm in diameter. Upon heating the peptide solution to 75°C for 40 min and then cooling to room temperature, they observed thicker fibrils (~70 nm in diameter). Importantly, these fibrils exhibited two key characteristics of natural collagen fibrils. First, the fibrils displayed tapered tips at their termini—a feature observed in type I collagen fibers and thought to be important for fiber growth (138). Second, Chaikof and coworkers observed D-periodic structure in synthetic collagen fibrils, with $D \approx 18$ nm. The self-assembly process presumably relies on Coulombic interactions and hydrogen bonds between charged Arg and Glu residues in individual, axially staggered triple helices (Figure 10c).

The methodologies described above enable the creation of long, triple-helical, collagen-like fibrils. Despite major advances, synthetic collagen-mimetic fibrils still lack many of the characteristics of higher-order collagen structures. In addition, the mechanical properties of synthetic collagenous materials have not been studied to date. Synthetic collagens that closely mimic the length, girth, patterns, mechanical properties, and complexity of natural collagen fibrils remain to be developed, but rapid progress in the past few years engenders great optimism.

**Biological and Biomedical Applications of Synthetic Collagen**

Relatively few CRPs have been tested as biomaterials. Goodman and coworkers (139) showed that peptoid-containing CRPs have a notable ability to bind to epithelial cells and fibroblasts, particularly when displayed on a surface. CRPs are also useful for inducing platelet aggregation, which can aid the wound-healing process (140, 141).

A key step toward utilizing collagenous biomaterials for therapeutic purposes is the development of CRPs that can either adhere to or bury themselves within biological collagen. Most efforts toward these objectives have relied on immobilization of CRPs on an unrelated substance. Yu and coworkers (142) prepared CRP-functionalized gold nanoparticles...
Figure 10
Strategies for the self-assembly of long, synthetic collagen triple helices and fibrils. (a) Disulfide bonds enforce a strand register with sticky ends that self-assemble (131). (b) Stacking interactions between electron-poor pentafluorophenyl rings and electron-rich phenyl rings lead to self-assembly (133, 134). (c) Coulombic forces between cationic and anionic blocks encourage self-assembly. TEM image of a resulting fiber shows $D$-periodicity with $D = 17.9$ nm (137). Natural type I collagen has $D = 67$ nm.
and demonstrated binding of the gold nanoparticles to the gap region of natural collagen. Maryanoff and coworkers found that CRPs displayed on latex nanoparticles can stimulate human platelet aggregation with a potency similar to that of type I collagen (140). In an important extension of this work, they demonstrated that triple-helical fibrils obtained via aromatic interactions had a similar level of thrombogenic activity to the CRPs immobilized on latex nanoparticles (134). Finally, single strands of CRPs and polyethylene glycol-conjugated CRPs bind to collagen films even without immobilization on nanoparticles (143) and are of potential use in collagen imaging (144) and wound-healing applications. The future of these approaches appears to be especially bright.

SUMMARY POINTS

1. High-resolution crystal structures and modern biophysical approaches have enabled detailed study of the structure and stability of collagen triple helices. The ladder of hydrogen bonds observed in these crystal structures is essential for holding the triple helix together, and its absence in natural collagen leads to a variety of pathological conditions.

2. Steroelectronic effects impart significant structural stability to collagen by preorganizing individual polypeptides for triple-helix formation. For example, Hyp in the Yaa position stabilizes the triple helix via a stereoelectronic effect. Steroelectronic effects are also important for the structure and stability of numerous other peptides and proteins.

3. Posttranslational modifications to protocollagen are of fundamental importance to the synthesis of a stable ECM. These modifications include hydroxylation and cross-linking reactions.

4. Collagen fibrillogenesis is an essential process for the formation of macromolecular biological scaffolds. Relatively high-resolution models of type I and type II collagen fibrils are now available and, for type I collagen, show that collagen fibrils can be described as nanoscale ropes.

5. Simple means to synthesize long collagen triple helices and fibrils have become apparent. The resultant materials are poised for use in biomedicine and nanotechnology.

FUTURE ISSUES

1. The factors that affect triple-helix stability for Pro derivatives in the Yaa position are now clear. In comparison, the Xaa position is understood only poorly. What, for example, is the physicochemical basis for the anomalous effects of hyp and Hyp on triple-helix stability in the Xaa position?

2. The current understanding of triple-helix structure and stability derives from analyses of triple-helical CRPs. Do these analyses provide insight on the stability and mechanical properties of natural collagen fibrils?

3. What functionalities in natural collagen are important for proper fibril formation in the ECM? How might diseases stemming from improper fibril formation be subject to therapeutic intervention?
4. Can improved methods be developed to synthesize long collagen triple helices and relevant mimics of complex, hierarchical collagen assemblies?

5. What are the molecular structures of nonfibrillar collagen assemblies? How are those assemblies formed in vivo?

6. Natural collagens appear to engage many other proteins and biomolecules. Which ones? How? Can those interactions be manipulated to treat disease?

7. How can synthetic collagen-based biomaterials lead to expeditious therapies?

DISCLOSURE STATEMENT
The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS
The authors acknowledge Dr. Jeet Kalia for critical reading of the manuscript and Amit Choudhary for creating Figure 8d. M.D.S. was supported by graduate fellowships from the Department of Homeland Security and the Division of Medicinal Chemistry, American Chemical Society. Collagen research in our laboratory is supported by Grant AR044276 (NIH).

LITERATURE CITED


43. Demonstration that 4R-hydroxylation of Pro residues in the Yaa position of collagen strands is required for animal life.
61. Shoulders MD, Guzei IA, Raines RT. 2008. 4-Chloroprolines: synthesis, conformational analysis, and effect on the collagen triple helix. *Biopolymers* 89:443–54


**RELATED RESOURCES**


**NOTE ADDED IN PROOF**
Contents

Preface ................................................................. v

Prefatory Articles

Frontispiece
E. Peter Geiduschek ........................................................ xii

Without a License, or Accidents Waiting to Happen
E. Peter Geiduschek ........................................................ 1

Frontispiece
James C. Wang ........................................................... 30

A Journey in the World of DNA Rings and Beyond
James C. Wang ........................................................... 31

Biochemistry and Disease Theme

The Biochemistry of Disease: Desperately Seeking Syzygy
John W. Kozarich .......................................................... 55

Biosynthesis of Phosphonic and Phosphinic Acid Natural Products
William W. Metcalf and Wilfred A. van der Donk ....................... 65

New Antivirals and Drug Resistance
Peter M. Colman .......................................................... 95

Multidrug Resistance in Bacteria
Hiroshi Nikaido ............................................................. 119

Conformational Pathology of the Serpins: Themes, Variations, and Therapeutic Strategies
Bibek Gooptu and David A. Lomas .................................... 147

Getting a Grip on Prions: Oligomers, Amyloids, and Pathological Membrane Interactions
Byron Caughey, Gerald S. Baron, Bruce Chesebro, and Martin Jeffrey ........... 177

Ubiquitin-Mediated Protein Regulation

RING Domain E3 Ubiquitin Ligases
Raymond J. Deshaies and Claudio A.P. Joazeiro ......................... 399

Regulation and Cellular Roles of Ubiquitin-Specific Deubiquitinating Enzymes
Recognition and Processing of Ubiquitin-Protein Conjugates by the Proteasome
Daniel Finley .................................................................................................................. 477

Degradation of Activated Protein Kinases by Ubiquitination
Zhimin Lu and Tony Hunter .......................................................................................... 435

The Role of Ubiquitin in NF-κB Regulatory Pathways
Brian Skau, Xiaomo Jiang, and Zhijian J. Chen ............................................................. 769

Biological and Chemical Approaches to Diseases of Proteostasis
Deficiency
Evan T. Powers, Richard I. Morimoto, Andrew Dillin, Jeffery W. Kelly, and William E. Balch ................................................................. 959

Gene Expression
RNA Polymerase Active Center: The Molecular Engine of Transcription
Evgeny Nudler ............................................................................................................... 335

Genome-Wide Views of Chromatin Structure
Oliver J. Rando and Howard Y. Chang ......................................................................... 245

The Biology of Chromatin Remodeling Complexes
Cedric R. Clapier and Bradley R. Cairns .................................................................... 273

The Structural and Functional Diversity of Metabolite-Binding Riboswitches
Adam Roth and Ronald R. Breaker .............................................................................. 305

Lipid and Membrane Biogenesis
Genetic and Biochemical Analysis of Non-Vesicular Lipid Traffic
Dennis R. Voelker ......................................................................................................... 827

Cholesterol 24-Hydroxylase: An Enzyme of Cholesterol Turnover in the Brain
David W. Russell, Rebekka W. Halford, Denise M.O. Ramirez, Rabul Shab, and Tiina Kotti ........................................................................................................... 1017

Lipid-Dependent Membrane Protein Topogenesis
William Dowhan and Mikhail Bogdanov ..................................................................... 515

Single-Molecule Studies of the Neuronal SNARE Fusion Machinery
Axel T. Brunger, Keith Weninger, Mark Bowen, and Steven Chu ............................... 903

Mechanisms of Endocytosis
Gary J. Doherty and Harvey T. McMahon .................................................................. 857
Recent Advances in Biochemistry

Motors, Switches, and Contacts in the Replisome
Samir M. Hamdan and Charles C. Richardson ........................................ 205

Large-Scale Structural Biology of the Human Proteome
Aled Edwards ................................................................. 541

Collagen Structure and Stability
Matthew D. Shoulders and Ronald T. Raines ...................................... 929

The Structural and Biochemical Foundations of Thiamin Biosynthesis
Christopher T. Jurgenson, Tadhg P. Begley, and Steven E. Ealick ............ 569

Proton-Coupled Electron Transfer in Biology: Results from
Synergistic Studies in Natural and Model Systems
Steven Y. Reece and Daniel G. Nocera ........................................ 673

Mechanism of Mo-Dependent Nitrogenase
Lance C. Seefeldt, Brian M. Hoffman, and Dennis R. Dean .................... 701

Inorganic Polyphosphate: Essential for Growth and Survival
Narayana N. Rao, María R. Gómez-García, and Arthur Kornberg ............. 605

Essentials for ATP Synthesis by F1F0 ATP Synthases
Christoph von Ballmoos, Alexander Wiedenmann, and Peter Dimroth .......... 649

The Chemical Biology of Protein Phosphorylation
Mary Katherine Tarrant and Philip A. Cole ..................................... 797

Sphingosine 1-Phosphate Receptor Signaling
Hugh Rosen, Pedro J. Gonzalez-Cabrera, M. Germana Sanna, and Steven Brown .... 743

The Advent of Near-Atomic Resolution in Single-Particle Electron
Microscopy
Yifan Cheng and Thomas Walz ........................................ 723

Super-Resolution Fluorescence Microscopy
Bo Huang, Mark Bates, and Xiaowei Zhuang .................................. 993

Indexes

Cumulative Index of Contributing Authors, Volumes 74–78 .................. 1041
Cumulative Index of Chapter Titles, Volumes 74–78 ............................ 1045

Errata

An online log of corrections to Annual Review of Biochemistry articles may be found at http://biochem.annualreviews.org/errata.shtml