PNAS COMMENTARY

A two-step dance commits collagen to folding

Barbara Brodsky[a] and Anton V. Persikov[b,1]

There is a story that Francis Crick referred to collagen as the "excelsior of the body." According to the dictionary, *excelsior* means "fine curled wood shavings used especially for packing fragile items." This reflects early views of the fibrous protein collagen as unremarkable and primarily a structural filler, cushioning the body's more dynamic biological processes and organs. Indeed, the amino acid sequence of collagen polypeptide chains is highly repetitive with glycine as every third residue and a high content of imino acids, features which dictate collagen's distinctive triple-helical structure. The triple helices then self-associate within the extracellular matrix to form fibers, which provide essential mechanical properties to bone, skin, tendons, and other tissues (1). However, the apparent simplicity of collagen's sequence and structure belies the complexity of its synthesis and folding processes, as well as the nuanced ways in which the sequence/structure variations influence cell signaling and interactions. In PNAS, Yammine et al. report important findings that prompt a reevaluation of folding and chain selection mechanisms for type I collagen, the most abundant protein in the human body (2).

Type I collagen is a heterotrimer composed of two homologous chains: two α1 chains and one α2 chain. These chains each contain a central $(Gly\text{-}Xaa\text{-}Yaa)_{338}$ triple-helical domain (THD) and have a similar charge distribution. However, compared to the α1 chain, the α2 THD has fewer imino acids (30% vs. 35% of Xaa and Yaa positions) and significantly more hydrophobic residues (16% vs. 9% of Xaa and Yaa positions), which must influence structural stability and functional interactions. The amino acid sequence alone has been shown to specify chain composition and register in short Gly-Xaa-Yaa repeat model peptides, where charged and cation–π interactions at the Xaa and Yaa positions promote formation of defined triple-helical heterotrimers (3). However, attempts to renature the long type I α1 and α2 THDs extracted from tissues led to poor nonspecific triple helix formation arising within and between chains (4). Type I collagen molecules require both a C-propeptide trimerization domain, as well as chaperones to achieve proper trimerization, chain selection, correct registration of the three chains, and triple helix folding (5, 6).

Both α1 and α2 chains of type I collagen are synthesized as procollagens, with a C-propeptide that drives trimerization and promotes registration of the chains, followed by zipper-like folding of the triple helix from the C-terminus toward the N terminus (4). The crystal structure of the α1 homotrimer C-propeptide revealed a coiled-coil stalk, a base, and three petal-like regions, and key residues were mutated to confirm their role in trimerization (7). In vertebrate α1 C-propeptides, the cysteine residues are highly conserved, and interchain disulfide bonds between Cys residues at two specific positions in the N-terminal region in the second and third positions were considered necessary for stable trimeric C-propeptides which promote triple helix formation (8).

In addition, it was demonstrated that coexpression of α1 and α2 C-propeptides in the presence of calcium leads to transient formation of all possible trimer combinations: α1α1α1, α1α1α2, α1α2α2, and α2α2α2 (8). Among these transient trimers, interchain stabilizing disulfide bonds can be formed only for the normal α1α1α2 collagen and for the α1α1α1 homotrimer. Because the α2 sequence has a Cys to Ser change in one of the critical positions, interchain disulfides linking all three chains could not be formed by α1α2α2 and α2α2α2 C-propeptide trimers, and collagens with these chain compositions have never been reported. It was assumed that formation of these interchain disulfides was the commitment step for proper chain selection and for triple helix formation.

Yammine et al., using an HT-1080 cell expression system capable of producing properly folded and posttranslationally modified procollagen molecules, introduced Cys to Ser mutations to show that interchain covalent disulfide bonds within the trimeric C-propeptide domain are not necessary for stable triple helix formation in procollagen (2). This finding was unexpected and indicates that a new mechanism is needed to explain why some transiently folded C-propeptide trimers can form stable triple-helical molecules (α1α1α2 and α1α1α1) while others (α1α2α2 and α2α2α2) cannot.

While the C-propeptide initiates assembly by facilitating transient trimer formation, a second step involving the sequence of the THD is also shown to be required to commit the protein to a stable, protease-resistant triple helix (2). Homotrimers of Pro-α1 chains can form stable triple-helical molecules resistant to enzyme digestion, while homotrimers of Pro-α2 chains cannot. Chimeric procollagen constructs containing α2 propeptides and telopeptides surrounding an α1 THD sequence formed triple helices, while constructs with α1 propeptides and telopeptides surrounding an α2 THD did not (2). This highlights the importance of a sufficiently stable $(Gly\text{-}Xaa\text{-}Yaa)_n$ sequence near the C-terminal initiation region and is consistent with the higher stability of the C-terminal tripeptides in α1 versus α2, as predicted and reported experimentally by Yammine et al. (2). The inability of α2 chains to form a stable homotrimeric triple helix aligns with the embryonic lethality observed in transgenic mice where expression of α1 chains has been eliminated (9).

Author affiliations: [a]Department of Biomedical Engineering, Tufts University, Medford, MA 02155; and [b]Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY 10010

Author contributions: B.B. and A.V.P. wrote the paper.

The authors declare no competing interest.

See companion article, "An outcome-defining role for the triple-helical domain in regulating collagen-I assembly," 10.1073/pnas.2412948121.

[1]To whom correspondence may be addressed. Email: apersikov@flatironinstitute.org.

Published December 16, 2024.

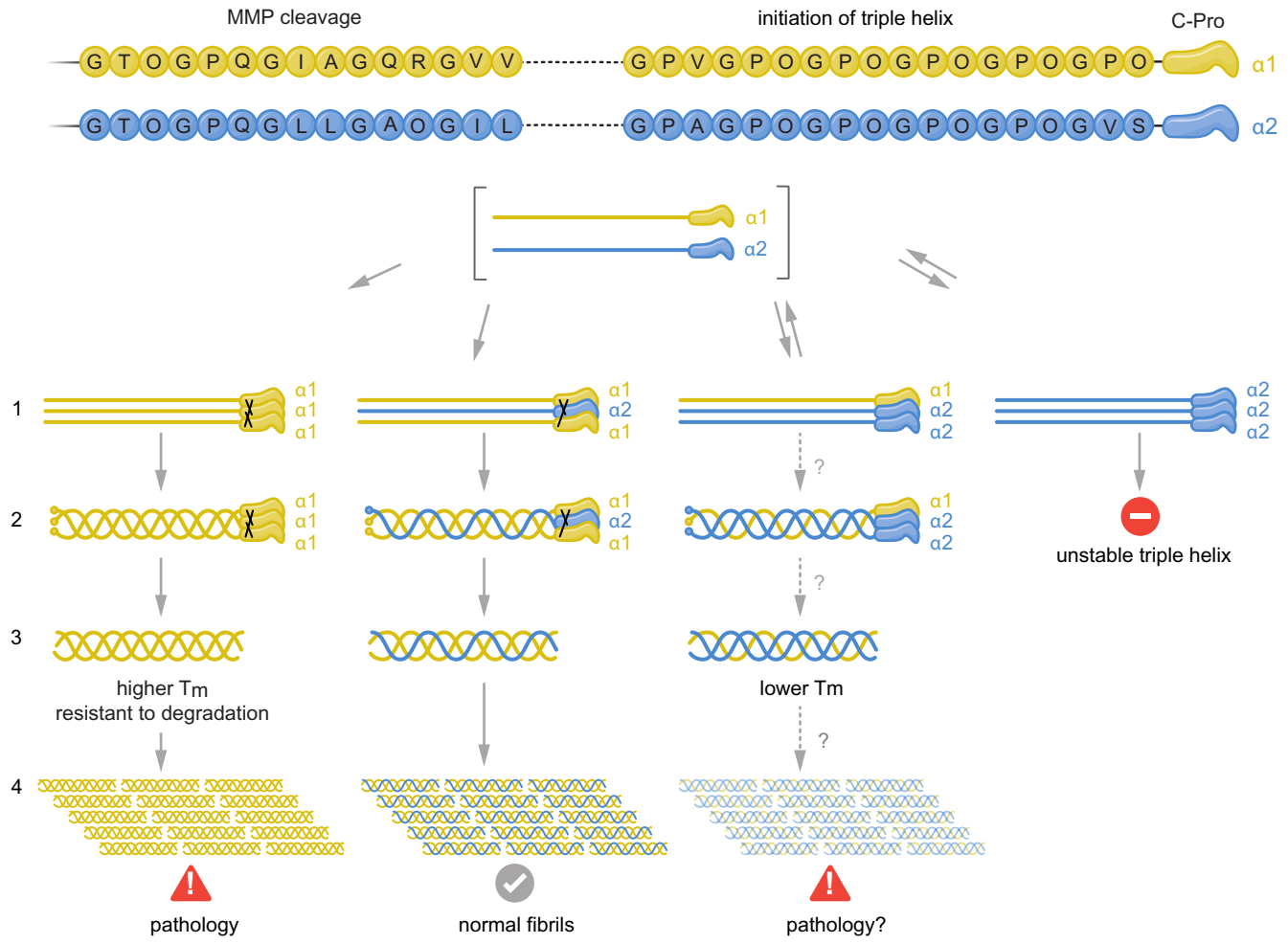

Fig. 1. Schematic view of potential type I collagen assembly pathways, demonstrating how α1 and α2 chain expression can lead to normal α1α1α2 heterotrimers, while alterations in chain expression may contribute to diseased states. On the top are sections of the amino acid sequence of the triple helix domains of the α1 and α2 chains, highlighting the imino acid–rich C-terminal sequence which initiates triple helix formation, the MMP cleavage site, and the C-propeptides. The folding pathways of the four possible trimer combinations are illustrated: *α1α1α1* homotrimers; the normal type I α1α1α2 heterotrimers; α1α2α2 heterotrimers, which have never been observed but are postulated to form when there is excess α2 chain expression; α2α2α2 homotrimers which cannot form stable triple helices. The steps described are (1) trimerization of the C-Propeptides, with interchain disulfide linkages stabilizing the *α1α1 α1* and α1α1α2 species; (2) triple helix folding; (3) cleavage of the propeptides in the pericellular space; and (4) formation of D-periodic fibrils in the extracellular matrix.

Under normal physiological circumstances, α1 and α2 chains are expressed in a 2:1 ratio, and characteristic α1α1α2 heterotrimers are predominant in tissues. Small amounts of α1 homotrimers have been observed in skin and embryonic tissues, but larger amounts are found in pathological states (10, 11). The absence of expression of intact α2 chains leads to α1 homotrimers in some cases of Osteogenesis Imperfecta and in mouse models 12, 13. Notably, α1 homotrimers form a triple helix approximately 2.5° more stable than normal α1α1α2 heterotrimers (14), and these homotrimers can form D-periodic fibrils with a normal axial structure, but lack ordered lateral packing and have reduced cross-linking (13, 15). Another striking feature is the resistance of α1 homotrimers to degradation by matrix metalloproteinases, the usual mechanism of collagen turnover, and a detailed analysis shows the basis of this resistance lies in differences between the α1 and α2 chains in the cleavage region (16) (Fig. 1). Such α1 homotrimers are also produced by some cancer cells and are found in tumors (17). In cancer cell lines, epigenetic silencing of the α2 chain has been shown to lead to α1 homotrimer formation 5, 18, 19 and it has been speculated that the resistance to degradation of fibrils formed from α1 homotrimers may facilitate cancer cell proliferation, migration, and altered tissue remodeling (16). The presence of α1 homotrimers have also been associated with other diseases, including liver fibrosis, intervertebral disc degeneration, and osteoarthritis (20).

While α2 chain deficiency can lead to α1 homotrimers and associated pathologies, higher than normal levels of α2 are also linked to disease states. In glioblastomas and colorectal cancers, elevated α2 chain expression correlates with poorer prognosis, likely due to effects on cell migration and tumor aggressiveness (19, 21). Research from the Shoulders laboratory suggests that an elevated α2 chain concentration might theoretically lead to transient C-propeptide trimers containing three α2 chains or α1α2α2 heterotrimers (8), although triple-helical molecules with these compositions have not been observed experimentally. Transient C-propeptide trimers containing three α2 chains cannot progress to triple helix formation and would likely dissociate and reform heterotrimers (2). The stability of a hypothetical α1α2α2 triple helix would be expected to be greater than the unstable α2α2α2 trimers but less than the normal α1α1α2 heterotrimers (Fig. 1) (22). If α1α2α2 triple helices are formed, such

molecules could be incorporated into the matrix, potentially forming abnormal fibrils that may underlie the negative outcomes reported.

> **"In PNAS, Yammine et al. report important findings that prompt a reevaluation of folding and chain selection mechanisms for type I collagen, the most abundant protein in the human body."**

The clarification of requirements for triple helix folding provides a mechanistic basis for considering disease states associated with altered expression levels of collagen chains, where either insufficient or excessive α2 expression can contribute to pathology. Understanding the molecular basis of these diseases will require a fuller characterization of each stage of the multistep process of collagen production, ranging from chain expression to trimer formation, stable triple helix formation, and fibril assembly (Fig. 1). The trimerization of the C-propeptide domains has long been recognized as a critical first step in collagen folding. The definition of a second step involving stable Gly-Xaa-Yaa triplets of the THD as necessary for commitment to irreversible triple helix folding paves the way for further research into both normal and abnormal collagen chain assembly mechanisms.

ACKNOWLEDGMENTS. We thank Barbara Smith Koff for assistance in reviewing the manuscript and the suggestion of relevant references. We thank Lucy Reading-Ikkanda, Simons Foundation for help with illustration. A.V.P. is grateful for the ongoing support through the Flatiron Institute, a division of the Simons Foundation.

1. B. Brodsky, A. V. Persikov, Molecular structure of the collagen triple helix. *Adv. Protein Chem.* **70**, 301–339 (2005).
2. K. M. Yammine *et al.*, An outcome-defining role for the triple-helical domain in regulating collagen-I assembly. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2412948121 (2024).
3. C. C. Cole *et al.*, Heterotrimeric collagen helix with high specificity of assembly results in a rapid rate of folding. *Nat. Chem.* **16**, 1698–1704 (2024).
4. S. P. Boudko, J. Engel, H. P. Bachinger, The crucial role of trimerization domains in collagen folding. *Int. J. Biochem. Cell Biol.* **44**, 21–32 (2012).
5. H. Su, M. Karin, Collagen architecture and signaling orchestrate cancer development. *Trends Cancer* **9**, 764–773 (2023).
6. N. J. Bulleid, Solving the mystery of procollagen chain selectivity. *Nat. Struct. Mol. Biol.* **19**, 977–978 (2012).
7. U. Sharma *et al.*, Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* **8**, 14671 (2017).
8. A. S. DiChiara *et al.*, A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* **9**, 4206 (2018).
9. A. Stacey *et al.*, Perinatal lethal osteogenesis imperfecta in transgenic mice bearing an engineered mutant pro-alpha 1(I) collagen gene. *Nature* **332**, 131–136 (1988).
10. J. Uitto, Collagen polymorphism: Isolation and partial characterization of alpha 1(I)-trimer molecules in normal human skin. *Arch. Biochem. Biophys.* **192**, 371–379 (1979).
11. S. A. Jimenez, R. I. Bashey, M. Benditt, R. Yankowski, Identification of collagen alpha1(I) trimer in embryonic chick tendons and calvaria. *Biochem. Biophys. Res. Commun.* **78**, 1354–1361 (1977).
12. A. C. Nicholls *et al.*, The clinical features of homozygous alpha 2(I) collagen deficient osteogenesis imperfecta. *J. Med. Genet.* **21**, 257–262 (1984).
13. D. J. McBride Jr., V. Choe, J. R. Shapiro, B. Brodsky, Altered collagen structure in mouse tail tendon lacking the alpha 2(I) chain. *J. Mol. Biol.* **270**, 275–284 (1997).
14. N. V. Kuznetsova, D. J. McBride, S. Leikin, Changes in thermal stability and microunfolding pattern of collagen helix resulting from the loss of alpha2(I) chain in osteogenesis imperfecta murine. *J. Mol. Biol.* **331**, 191–200 (2003).
15. T. J. Sims, C. A. Miles, A. J. Bailey, N. P. Camacho, Properties of collagen in OIM mouse tissues. *Connect Tissue Res.* **44**, 202–205 (2003).
16. S. Han *et al.*, Molecular mechanism of type I collagen homotrimer resistance to mammalian collagenases. *J. Biol. Chem.* **285**, 22276–22281 (2010).
17. E. Makareeva *et al.*, Carcinomas contain a matrix metalloproteinase-resistant isoform of type I collagen exerting selective support to invasion. *Cancer Res.* **70**, 4366–4374 (2010).
18. P. K. Sengupta, E. M. Smith, K. Kim, M. J. Murnane, B. D. Smith, DNA hypermethylation near the transcription start site of collagen alpha2(I) gene occurs in both cancer cell lines and primary colorectal cancers. *Cancer Res.* **63**, 1789–1797 (2003).
19. Y. Chen *et al.*, Oncogenic collagen I homotrimers from cancer cells bind to alpha3beta1 integrin and impact tumor microbiome and immunity to promote pancreatic cancer. *Cancer Cell* **40**, 818–834.e9 (2022).
20. K. J. Lee *et al.*, Collagen (I) homotrimer potentiates the osteogenesis imperfecta (oim) mutant allele and reduces survival in male mice. *Dis. Model Mech.* **15**, dmm049428 (2022).
21. X. Yuan, Y. He, W. Wang, ceRNA network-regulated COL1A2 high expression correlates with poor prognosis and immune infiltration in colon adenocarcinoma. *Sci. Rep.* **13**, 16932 (2023).
22. A. V. Persikov, J. A. M. Ramshaw, B. Brodsky, Prediction of collagen stability from amino acid sequence. *J. Biol. Chem.* **280**, 19343-19349 (2005). https://doi.org/10.1074/jbc.m501657200.