THE PROTEIN SOCIETY WILEY

The predominant roles of the sequence periodicity in the self-assembly of collagen-mimetic mini-fibrils

Fangfang Chen[1] | Rebecca Strawn[2] | Yujia Xu[3]

[1]Department of Biologics, Frontage Laboratories, Exton, Pennsylvania

[2]SGS, West Chester, Pennsylvania

[3]Department of Chemistry, Hunter College of the City University of New York, New York, New York

Correspondence
Yujia Xu, Department of Chemistry, Hunter College of the City University of New York, 695 Park Ave., New York, NY 10065.
Email: yujia.xu@hunter.cuny.edu

Abstract

Collagen fibrils represent a unique case of protein folding and self-association. We have recently successfully developed triple-helical peptides that can further self-assemble into collagen-mimetic mini-fibrils. The 35 nm axially repeating structure of the mini-fibrils, which is designated the *d*-period, is highly reminiscent of the well-known 67 nm *D*-period of native collagens when examined using TEM and atomic force spectroscopy. We postulate that it is the pseudo-identical repeating sequence units in the primary structure of the designed peptides that give rise to the *d*-period of the quaternary structure of the mini-fibrils. In this work, we characterize the self-assembly of two additional designed peptides: peptide Col877 and peptide Col108rr. The triple-helix domain of Col877 consists of three pseudo-identical amino acid sequence units arranged in tandem, whereas that of Col108rr consists of three sequence units identical in amino acid composition but different in sequence. Both peptides form stable collagen triple helices, but only triple helices Col877 self-associate laterally under fibril forming conditions to form mini-fibrils having the predicted *d*-period. The Co108rr triple helices, however, only form nonspecific aggregates having no identifiable structural features. These results further accentuate the critical involvement of the repeating sequence units in the self-assembly of collagen mini-fibrils; the actual amino acid sequence of each unit has only secondary effects. Collagen is essential for tissue development and function. This novel approach to creating collagen-mimetic fibrils can potentially impact fundamental research and have a wide range of biomedical and industrial applications.

KEYWORDS

axial-repeating structure of protein, collagen-mimetic fibrils, fibrous molecular assembly, protein design: fibrous protein, self-association of collagen triple helices, self-association of protein, sequence periodicity and protein structure, tailored functional collagen-mimetic biomaterial

1 | INTRODUCTION

Collagen is a highly versatile protein, acting as the stable molecular scaffold of tissues as varied as bones, skin, blood vessel walls, and the cornea of the eye. The ability to develop collagen-mimetic supramolecular structures using designed peptides tests the robustness of our understanding of proteins and also has tremendous impact on the advanced studies of tissues and cells and on biomedical and industrial applications. In this work, we focus on the design strategy of

mimetic collagen fibrils bearing the well-known axial structural motif of the D-period. The D-period is characterized by 67 nm striations observed on electron micrographs of collagen fibrils,[1,2] on images taken by atomic force spectroscopy (AFM),[3] or using X-diffraction.[4] This unique structure of collagen fibrils emerges from the lateral packing of collagen triple helices in a structurally specific way and is common to all fibrillar collagens including the three major types of collagens of the connective tissues: collagen types I, II, and III.[5,6] The fibrils will undergo further assembly in tissues to form fibers of varied diameter and to incorporate other collagens and/or other proteins.[6–8] It is not fully understood what molecular interactions lead to the formation of this specific structure, and how the D-period affects the functional and biomechanical properties of the tissues. The D-period is, nonetheless, considered the crucial structural element lending to the various functions of collagens and has been implicated in the load bearing properties of tissues, in the mineralization of bones, and in the regulation of cell differentiation and adhesion during the development of tissues.[9–12]

Development of collagen mimetic fibrils having D-period like structures through self-association of triple helices represents a unique case of protein design. The D-periodic fibrils embody the complexity of the structural hierarchy of proteins. Starting at the first-order structure, the long polypeptide chains of collagen have extensive, non-interrupted (Gly-Xaa-Yaa)$_n$ repeating sequences, where n is often greater than 300 in fibrillar collagens, and Xaa and Yaa can be any amino acid residues but are frequently Pro and hydroxyl proline (Hyp), respectively. Three such polypeptide chains, which can be identical or different, then wrap around each other about a common axis to form a collagen triple helix. The three polypeptide chains of a triple helix are arranged in parallel and with a mutual staggering of one-residue at the ends. The Gly residues are unequivocally buried at the center of the triple helix, whereas the side chains of residues in the X and Y positions are displayed on the surface of the helix in a *linear* N-to-C directionality. The backbone of the triple helix is rigid and has a uniform conformation characterized by an average axial rise of ~0.8-0.9 nm per Gly-Xaa-Yaa tripeptide.[13] The triple helix of fibrillar collagens consisting of ~1000 amino acid residues (per single peptide chain) is often characterized as a "rigid rod" about 300 nm in length. Furthermore, within this structural framework, the 67 nm D-period would correspond to a section of the triple helix formed by 234 amino acid residues (per single chain). Despite consisting of three polypeptide chains, the triple helix is often considered the "secondary" structure of collagen.[14] The triple helices in tissues inevitably further assemble to form different, biologically functional supramolecular structures. For fibrillar collagens, the triple helices further self-associate laterally with a mutual staggering of multiples of D-periods, or multiples of 234-residue units, between the neighboring helices. Because each triple helix

encompasses about 4.4 D by length, this D-staggered arrangement would thus, generate a gap (0.6D) and an overlap (0.4D) zone every 67 nm propagating throughout the fibril.

Collagen fibrils isolated from tissues can be dissolved into constituent triple helices under acidic conditions; the triple helices will reassemble into fibrils having the same D-period upon incubating in neutral pH and at a temperature close to the physiological temperature of the original host.[1,15,16] This spontaneous *in vitro* fibrillogenesis indicates the D-period is the most stable conformation formed by interactions between triple helices, akin to the spontaneous folding of proteins into globular structures. Unique to fibrillogenesis, however, is the rigid conformation of the triple helix that allows little flexibility for bending, coiling, or super twisting.[17–19] The rigidity of the triple helix effectively limits the available conformational space for the fibrils during the self-assembly. The conformational bias of the D-period would thus emerge from maximizing the number of interactions between helices in a specific mutual staggering of multiples of 234-residues during the self-assembly.[20]

Based on the understanding of the fibril structure of collagen, we devised a design strategy of incorporating *repeating sequence units* within collagen mimetic peptides and generated two triple helices that self-associate into mini-fibrils having a D-period-like axial repeating structure about 35 nm in size.[21,22] The two triple helices, referred to as peptide Col108 and peptide 2U108, consist of three and two pseudo-identical sequence units, respectively, placed in tandem in their first-order structure. Each sequence unit has about 123 amino acid residues. The 35 nm periodic structure, which is subsequently termed the d-period,[21] coincides with the size of a section of the triple helix formed by the 123 residues included in one sequence unit. The fully folded Col108 and 2U108 helices are about 3.3d and 2.3d in length, respectively. Thus, similar to collagen fibrils, by a mutual stagger of one (or two, in the case of Col108) sequence unit at the N-termini the self-assembled mini-fibrils will give rise to alternating 0.3d overlap (~10 nm) and 0.7d gap (~25 nm) regions every 35 nm. The alternating gap-overlap regions of the d-period were characterized by both electron microscopy[21,22] and atomic force microscopy.[21] We further demonstrated that a minimum of two identical sequence units is necessary and sufficient to form the d-period fibrils;[22] another peptide, peptide 1U108, consisting of only one sequence unit was found to form only nonspecific aggregates without the d-periodic axial structure.

The tandem repeats of the identical sequence units in the first-order structure of the peptides imply a long-range periodicity in the amino acid sequences: the same amino acid residues in a sequence unit are repeated every 123 residues in Col108 and 2U108. This 123-residue periodicity represents a higher-order, long-range sequence periodicity

existing on top of the local sequence periodicity inherited in the (Gly-Xaa-Yaa) repeating sequence of a triple helix. We argued that, because of this sequence periodicity, the in-register alignment of the interacting residues in the unit-staggered arrangement provides an effective way to optimize the interactions of the associating helices and selectively promotes the d-periodic conformation to emerge from the process of self-assembly. The same periodic placement of interacting residues was also postulated to be the determining factor of the fibrillogenesis of type I collagen.[20,23] Clusters of charged and of hydrophobic residues within the sequence unit were identified that may potentially be involved in the stabilization of the d-period mini-fibrils, although the specifics of these interactions are still lacking.[21]

This simple design rational was supported by the studies of Col108, 2U108, and 1U108. However, it should be noted that all three peptides share the same implicated set of amino acid sequence—that of the 123-residue sequence unit. To develop this approach into a general strategy for applications, we studied the self-assembly of two new peptides to delineate the crucial involvements of the sequence architecture of repeating sequence units verses that of the specific amino acid residues of a given sequence unit. Peptide Col877 was created to have tandem repeats of three pseudo-identical sequence units similar to Col108, except the sequence unit is comprised of a completely unrelated amino acid sequence to that of Col108. In contrast, another peptide, peptide Col108rr, was designed which kept the same amino acid *composition* (i.e., the same amino acid identities in the same quantities) of Col108, but the amino acid sequences of the three sequence units were randomized to potentially affect the in-register alignment of the like residues in the unit-staggered assembly. In other words, Col877 retains the same sequence architecture of a 123-residue

long-range sequence periodicity as Col108 but includes different amino acid residues, whereas peptide Col108rr consists of the same amino acid residues as that of Col108 but lacks the long-range sequence periodicity. The outcomes of this study further accentuate the importance of the overall long-range periodicity in the first order structure of the peptides; the actual amino acid residues conforming to the periodicity purportedly have only secondary effects.

2 | RESULTS

2.1 | Design of Col877 and Col108rr

The two new peptides were developed following the same amino acid sequence architecture of the original Col108 peptide (Figure 1). In peptide Col877, a new 108-residue domain, the C877 domain, replaces the Col domain in all three sequence-units. The 108 residues of the C877 domain copy the sequence of residues 877-985 of the α1 chain of human type I collagen and differ in both composition and sequence from that of the Col domain (Table 1). Thus, despite the same 123-residue periodicity of the first-order structure, a different set of amino acid residues from those in Col108 will be involved in the self-association of Col877. In peptide Col108rr, the amino acid sequences of the Col domain in the second and the third repeating sequence units of Col108 are replaced by, respectively, the rndCol and the rvCol domain. The sequence of the rndCol domain is generated by randomizing the sequence of the original Col domain, whereas that of the rvCol domain is generated by reversing the positions of the X- and Y-residues of the Col domain (Table 1). Therefore, differing from Col877, all the potential interacting residues of Col108 are present in Col108rr, except these residues do not
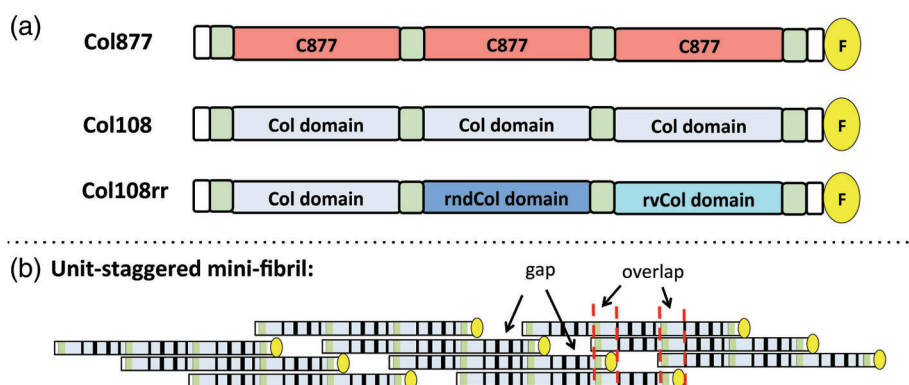

FIGURE 1 The sequence architecture of the peptides. (a) Schematic depictions of peptides Col877, Col108, and Col108rr. The four different, 108-residue sequence domains (see text) are shown in different colors: C877 (red), Col domain (light blue), rndCol (dark blue), and rvCol (cyan). The common features of all three peptides are shown as the following: foldon domain (yellow circle), GPP4 linkers (green blocks), and the sequence of the Cys-knot (white blocks). All three peptides have the same size: 417 amino acid residues including a 378-residue triple-helical domain consisting of non-interrupted Gly-Xaa-Yaa repeating sequences. (b) A schematic drawing of the unit-staggered Col108 mini-fibril based on previous studies[21,22] to highlight the alternating gaps and overlap zones and the in-register alignment of the interacting residues (black vertical bands). For visual clarity, we only selectively marked the four clusters of charged residues using the black vertical bars; other potential interacting residues are not marked

TABLE 1 Amino acid sequences of the 108-residue sequence domains

Domain	Amino acid sequence
C877[a]	GPVGPAGKSGDRGETGPAGPAGPVGPVGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQ GP*PGPPGSPGEQGPSGASGPAGPRGPPGSAGAPGKDGLNGLPGPI
Col domain	GERGPPGPQGARGLPGAPGQMGPRGLPGERGRPGAPGPAGARGEPGAPGSKGDTGAKGEPGPV GVQGPPGPAGEEGKRGARGEPGPTGPAPGPKGSPGEAGRPGEAGLP
rndCol domain[b]	GMQGLPGPR<u>GAPGR</u>AGPEGRAGRPGPAGLPGEQGRAGRPGPPGAEGAEGKPGAPG SPGPRGLPGPAGDTGPEGQVGARGPKGEPGTPGRAGEPGEPGVPGKAGPKGES
rvCol domain	GREGPPGQPGRAGPLGPAGMQGRPGPLGREGPRGPA<u>GAPGR</u>AGPEGPAGKSGTDG KAGPEGVPGQVGPPGAPGEEGRKGRAGPEGTPGAPGKPGPSGAEGPRGAEGPL

[a]The 108 residues of the C877 domain comprised of residues 877-985 of the α1 chain of human type I collagen. The six italicized proline residues in the Y-position of the (Gly-Xaa-Yaa)$_n$ repeating sequences are generally hydroxylated to form 4R-hydroxyproline in tissues.

[b]The nine-residue sequences identical in rndCol and rvCol domains are underlined.

TABLE 2 Sequence homology of the 108-residue sequence domains

Sequence domain	Hydrophobic[a]	Charged residues	Pro	Identical GXY triplets compared to Col domain[b]	Stabilizing tri-peptide			T_m^{cal} (°C)[c]
					GPP	GPR	KGE/D	
Col	21	23	24		2	1	2	38.2
C877	22	19	22	9/24	3	2	1	37.4
rndCol	21	23	24	11/22	0	2	0	37.6
rvCol	21	23	24	6/22 (14/22)[d]	2	2	0	36.6

[a]Including residues Ile, Leu, Val, Met, Gln, Asn, and Ala.

[b]The number of identical Gly-Xaa-Yaa tripeptides compared to Col domain. All domains consist of 36 Gly-Xaa-Yaa tripeptides; the 36 tripeptides consist of 22 different varieties (different sequences) except that of the C877 domain, which consist of 24 different tripeptides.

[c]The T_m calculated using the web-tool at http://compbio.cs.princeton.edu/csc/, based on the algorithm by Persikove et al.[24]

[d]The number of identical Gly-Xaa-Yaa tripeptides compared to rndCol domain.

have the same periodic placement as they do in Col108. Some sequence elements of Col108, such as the C-terminal foldon domain, the C-terminal (Gly-Pro-Pro)$_4$ sequence (GPP4), the Cys-knots, and the GPP4 linkers between the sequence units are retained in both Col877 and Col108rr. These elements are included to increase the stability and/or to facilitate the correct folding of the triple helices. Our previous work has suggested that these elements are not the determinant factors for the formation of the *d*-period mini-fibrils (more in Section 3).[22]

The variations of the amino acid compositions and sequence homology of the new sequence domains are summarized in Table 2. The 108 residues (36 Gly-Xaa-Yaa triplets) of the original Col domain are comprised of 22 different varieties of Gly-Xaa-Yaa tripeptides (Table 1). In comparison, the C877 domain is made of 24 different Gly-Xaa-Yaa tripeptides, among which nine are identical to those of the Col domain. The C877 also has fewer charged residues and a lower content of Pro residues. The rndCol is composed of 22 different Gly-Xaa-Yaa triplets, and half of them are different from those of the Col domain. The rvCol also has 22 different Gly-Xaa-Yaa tripeptides, and only six of them are identical to those of the Col domain. There appears to be more similarities between the rndCol and rvCol domains: 14 of the 22 non-identical constituent tripeptides are common between the two domains.

However, other than on one occasion (Table 1), the identical triplets are situated in different sequence contexts; the actual amino acid sequences of rndCol and rvCol are more different than the ratio 14/22 may suggest.

It is important that the designed peptides form stable triple helices, which is a structural prerequisite of the fibrillogenesis of collagen.[15,25] The thermal stability of a triple helix depends on the residues in the X and Y positions and on the specific combinations of the residues in the Gly-Xaa-Yaa triplet.[24] In general, a high content of imino acids is considered favorable for reducing conformational entropy because of their high propensity of forming a polyproline II conformation. Sequences of Lys-Gly-Glu (KGE) or Lys-Gly-Asp (KGD) are significant stabilizing factors due to their potential for forming inter-chain salt bridges in the folded triple helix.[24,26] The T_m^{cal} values calculated using the stability calculator[24] (http://compbio.cs.princeton.edu/csc/) were used to obtain a qualitative estimate of the relative stability of the three new sequence domains. The 108 residues of the Col domain were selected from segments of the amino acid sequence of the α1 chain of human type I collagen for their high propensity for triple-helix conformation,[21] and it is not surprising the T_m^{cal} of 38°C is the highest among the domains (Table 2). The T_m^{cal} value of the C877 domain is

just about 1°C lower. The T_m^{cal} rvCol is about 1.5°C lower than that of the Col domain and is the lowest among the four domains. The first returned randomized sequences for the rndCol domain had T_m^{cal} values around 34°C. Considering the already relatively low T_m^{cal} value of the rvCol domain, we considered it necessary to adjust the sequence of the rndCol to bring its T_m^{cal} value close to that of the Col domain to ensure the thermal stability of peptide Col108rr would be within the target range for fibrillogenesis studies. Adjustments were carried out by switching the residues in the X and Y positions in a few tripeptides. After such adjustments, the T_m^{cal} of rndCol was made to be less than half a degree lower than that of the Col domain (Table 2). Because it is not clear how the T_m^{cal} values of one or two domains of a given peptide would affect the overall stability and the melting temperature of the peptide, we chose to accept the small variations in T_m^{cal} values of the domains as an indication of similar stability of the peptides, which was later confirmed by thermal melt experiments as shown below.

The possible interactions of the associating helices in a unit-staggered arrangement of Col877 and of Col108rr were evaluated using calculated interaction curves—an approach used in the original design of Col108[21] and devised by Hulmes and colleagues in their study of type I collagen.[20] The interaction curves were calculated by summing up the total *number* of hydrophobic and electrostatic interactions between two adjacent triple helices, arranged in parallel, as a function of chain stagger—the residue shifting values.[21] In the cases of type I collagen and Col108, a set of strong peaks associated with a staggering number of multiples of 234 and 123 residues, respectively, was used to demonstrate the selectively imposed favorable interactions of the specific arrangements observed during the self-assemblies. As expected, the interaction curves of Col877 showed the same feature as that of Col108: a set of peaks with shift values of multiples of 123 residues dominate over other arrangements (Figure 2a). These peaks indicate the interactions between the associating helices were maximized when having a mutual staggering of one or multiple sequence units. The 123-residue shift value was favored by both the total hydrophobic interactions and the total electrostatic interactions. There also appears to be a set of peaks with shift values in multiples of 40 residues in both hydrophobic and electrostatic interactions but much smaller in amplitude. By shifting multiples of 40 residues, certain interactions may be encouraged between the associating helices; however, whether such a 40-residue staggered conformation will emerge from the self-association process depends on both the absolute stability of the conformation and its relative stability to other possible conformations including that of the unit-staggered arrangement.

The features of the interaction curve of Col108rr are rather different. There is no dominating peak at a shifting

value of 123; furthermore, neither the hydrophobic residues nor the charged residues confer substantial interactions at this specific shifting value (Figure 2b). The modified amino acid sequences of the rvCol and rndCol domains from that of the Col domain in Col108rr appear to effectively abolish the in-register alignments of the interacting residues in the one-unit staggered arrangement (Figure 2b, the drawing). Interestingly, there appears to be considerable interactions when the shifting value is 246, that is, close to a mutual staggering of two sequence units of associating helices. This 246-peak is mainly due to the hydrophobic interactions and has little contribution from the electrostatic interactions (Figure 2b). The 246-peak emerged from focusing too exclusively on the identity of the residues and the triple helix propensity and not enough on the common properties shared by some of the residues during the design of Col108rr. However, the magnitude of a given interaction curve reflects only the number of potential interactions and neither the actual interactions per se nor the strength of these interactions; it is not clear if the conformation having a mutual shifting value of 246 residues would have enough stabilizing interactions to form during the self-association. Besides, the possibility of creating mini-fibrils having exclusively 2-unit staggered arrangement represents an interesting perspective of its own. Our earlier modeling showed that an exclusively 2-unit staggered arrangement will give rise to a different axially repeating structure: a *d*-period having an axial periodicity of 70 nm consisting of a 25 nm gap and 55 nm overlap regions;[21] in the event such mini-fibrils do form, they can be readily recognized and be distinguished from the 35 nm *d*-period mini-fibrils using TEM and/or AFM. We, therefore, proceeded to study Col108rr without further adjustment of the sequence.

2.2 | Conformation of Col877 and Col108rr

The CD spectra of both Col877 and Col108rr in 5 mM HAc are characteristic of a collagen triple helix, showing a positive peak at 225 nm and a deep negative peak between 197 and 200 nm (Figure 3a). The spectra are also highly similar to that of Col108. The apparent melting temperature (T_m'), determined as the temperature when the fraction of folded is 50%, of both Col877 and Col108rr is ∼39°C, a few degrees lower than the 42°C of Col108 (Figure 2b). The variation of the T_m' values can be understood from the changes of the amino acid composition and/or the amino acid sequence of the peptides (Table 2). The different T_m' values between Col108 and Col108rr are particularly interesting considering both peptides have an identical amino acid composition. The salt-bridge sequences KGE and KGD appear to be rather effective stabilizing factors for the higher stability of Col108. Overall, the differences in T_m' of all three peptides are relatively minor. In general, the thermal stability
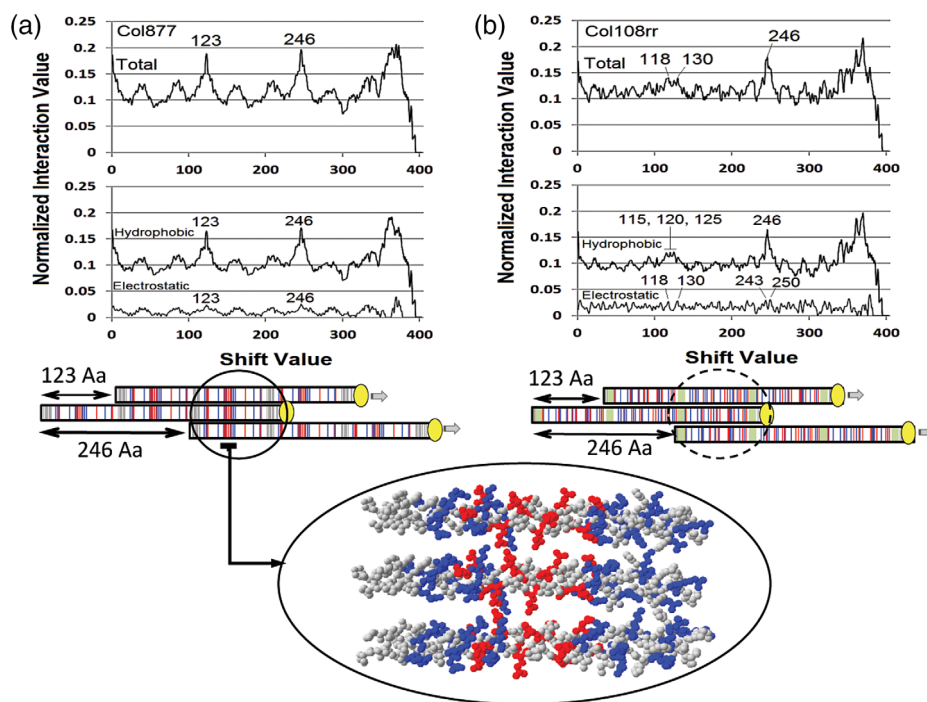
FIGURE 2 The interaction curves. The interaction curves of Col877 (a) and Col108rr (b) are shown for the total interactions for parallel chain arrangement (upper panels) and the constituent hydrophobic and electrostatic interactions (lower panels). Schematic drawings of helices having shifting values of 123 residues or 246 residues of the two peptides are shown, respectively, beneath their interaction curves. The red and blue vertical bars in the drawings mark, respectively, the locations of negatively charged and positively charged residues, and the yellow circle is the foldon domain. The black circle in (a) highlights the in-register alignment of the interacting residues between three neighboring helices of Col877; an expanded view of the section marked by the black bar is shown using the ball-and-stick structural model generated using Swiss-PDB-viewer (DeepView), in which the basic residues are in blue, acidic residues in red, and others in gray. The dotted circle in (b) highlights the lack of residue alignment in Col108rr

of collagen and collagen-like peptides having more than 100 residues is less sensitive to the constituent amino acid sequences than that of shorter peptides. Other factors, such as the foldon domain and the Cys-knots, may also contribute to the relative comparable thermal stability of the three peptides. The three melting curves differ slightly in shape. Many factors can potentially affect the shape of the curve because the thermal melt experiment of collagen triple helices is often not a true equilibrium process.[27–30] Within each peptide, the melting curve is reproducible.

2.3 | Self-assembly of Col877 and Col108rr

The fibril assembly of the peptides was initiated by mixing the peptide solutions in 5 mM HAc at ∼10°C with an equal volume of double strength fibril forming buffer at pH 7 and incubating the solutions in a water bath set at 26 or 37°C. The Col877 triple helix formed mini-fibrils having the characteristic d-period striation when examined under electron microscopy (Figure 4b–e). The mini-fibrils are smooth with tapered ends, similar to those observed for Col108 and 2U108. The diameter is about 50–90 nm at the center, and the length ranges from 600 nm to 1 μm, both comparable to

that of Col108 and 2U108 mini-fibrils.[21,22] Sometimes the mini-fibrils appear to be bundled together (Figure 4c), although it is not clear if this represents two or more fibrils coiled together or is simply overlapping images of individual mini-fibrils. The axial d-periodicity defined by a pair of dark and white bands on the electron micrographs was estimated to be 32 ± 1.4 nm based on measurements from 16 different grids, prepared using samples from several different purification batches. The mini-fibrils were readily observable after 6 hr of incubation at 37°C or after 24 hr of incubation at 26°C. The overall quality of the TEM images is improved compared to the previous work, which may lead to the more precise estimation of the d-period compared to the previous estimation of 35 ± 7.4 nm for Col108.[21] Alternatively, the slight discrepancies in the returned d-periods may reflect the variations of the backbone conformation of the triple helix with the constituent amino acid sequence. Taking the generally accepted variations of the helical rise as being ∼0.8–0.9 nm per Gly-Xaa-Yaa tripeptide, a 123-residue sequence unit is expected to have a length between 33 and 37 nm. The estimated d-periods of both Col877 and Col108 are consistent with the size of one sequence unit.

The images of Col877 in HAc solution before being transferred to fibril forming buffer provide a striking contrast
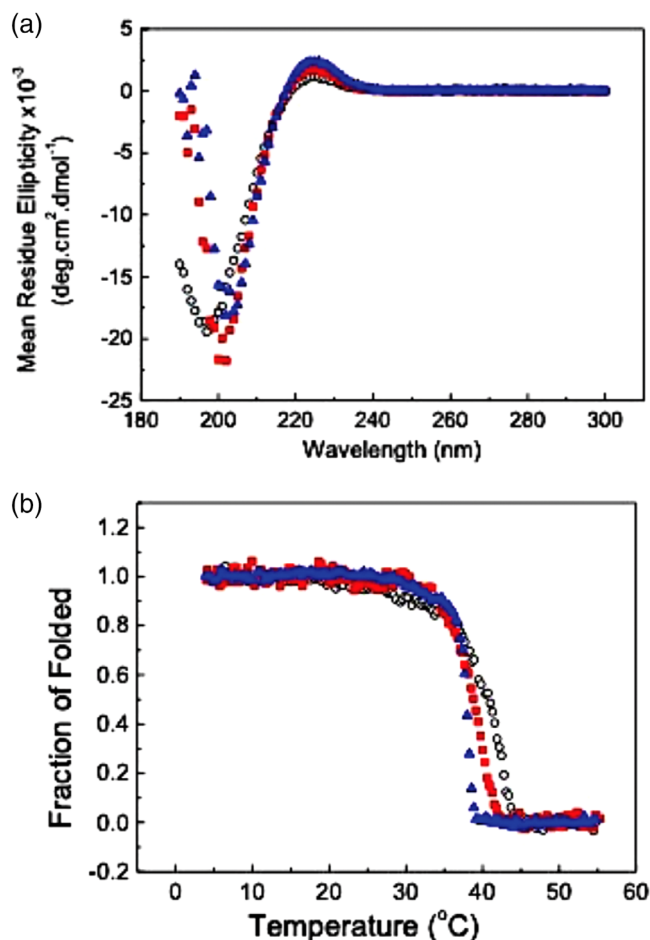
(a)

(b)

FIGURE 3 The triple helices of Col877, and Col108rr. (a) The CD scan of Col877 (red square), Col108rr (blue triangle), and Col108 (black circle), all at 0.2 mg/mL in 5 mM acetic acid (pH 5) and 4°C. (b) The temperature melting curves of Col877 (red square), Col108rr (blue triangle), and Col108 (black circle), all at concentration of 1 mg/mL in 5 mM acetic acid

(Figure 4a). The Col877peptides are largely dispersed triple-helix monomers and return a size consistent with the predicted size of ∼120 nm long and a few nanometers in diameter. It is quite remarkable to see the individual triple helices of Col877 in the image; in their rather rigid conformation the triple helices have the appearance of tiny straws, although some show slight bending or curving. There appeared to be some degree of self-association forming aggregates several times larger than a triple-helix monomer. It is not clear if the aggregates are formed in solution or caused by drying on the grids. The aggregates are significantly different in size and appearance than that of the mini-fibrils.

The TEM images of Col108rr peptides under the same experimental conditions, and under the same magnification of the electron microscope look clearly different (Figure 5a–d). After incubation at 37°C for 12–24 hr, samples appear to be largely dispersed. In a background of triple-helix monomers having a similar size to those of Col877, as shown in

Figure 4a, there appears to be several aggregates judging by their length and especially by their greater diameters. Occasionally, there are fibril-like assemblies (Figure 5b), but these fibril-shaped aggregates are shorter in length and much thinner in diameter than the mini-fibrils observed for Col877 and have no discernable structures. The presence of the aggregates seems to indicate the Col108rr molecules have the potential to interact with one another—to stick together—yet, either because the interactions between the helices are not strong enough or there lacks a mechanism to support further growth of the aggregates, no mini-fibrils having identifiable striation emerge from the self-association. Concerning the relatively lower thermal stability of Col108rr, several studies of the fibrillogenesis of Col108rr at 26°C were carried out to reduce any possible effects of partial thermal unfolding during fibril formation. After 24 hr incubation, no obvious aggregations were observed (Figure 5e). In fact, the images at 26°C looked very much like those of Col108rr in HAc (Figure 5f). We could not find any assemblies among the grids examined, at 37 or 26°C, to have the *d*-period, or to resemble the purported structure of the 2-unit staggered fibrils. If the 2-unit staggered assemblies ever formed, they did not grow to a size observable under TEM, and/or did not emerge as the dominating conformation.

3 | DISCUSSION

The identical *d*-period of Col877 mini-fibrils to those found in the fibril assemblies of peptides Col108 and 2U108, together with the lack of such in Col108rr, provides a convincing demonstration that the tandem repeat of sequence units is an effective strategy to design fibril-forming triple-helical peptides, provided there are enough interacting residues in the sequence units to stabilize the mini-fibril structure. From a protein design point of view, maximizing the interactions of the desired structure functions to both stabilize the desired conformation and differentially favor it over other competing ones. We argued in the previous work that two factors of the repeating sequence units—the in-register alignment of the interacting residues of the neighboring helices and the reiteration and, thus, magnification of these interactions propagated across repeating units—work synergistically to provide the conformational bias for the unit-staggered mini-fibrils.[21] The involvements of other structural elements have yet to be fully evaluated. The foldon domain may prove to be a necessary structural element to eliminating the end-on-end stacking of helices.[22] However, the possible functions of the GPP4 linkers, other than stabilizing the triple-helix conformation, remain speculative. Consistent with our previous observations,[22] the study of Col108rr demonstrates that the foldon and the GPP4 linkers alone are not enough to lead to the formation of the *d*-period mini-fibrils.

FIGURE 4 The mini-fibrils of Col877. The TEM images of peptide Col877 in acetic acid (a), in fibril forming buffer after 24 hr incubation at 37°C (b, c), and after 24 hr incubation at 26°C (d, e). The scale bars are equal to 200 nm in all images
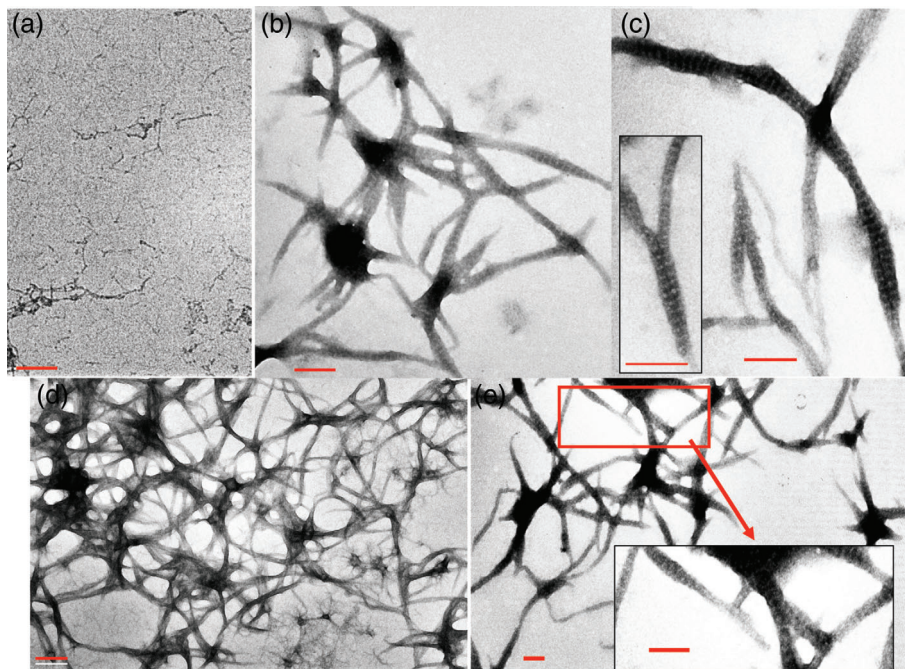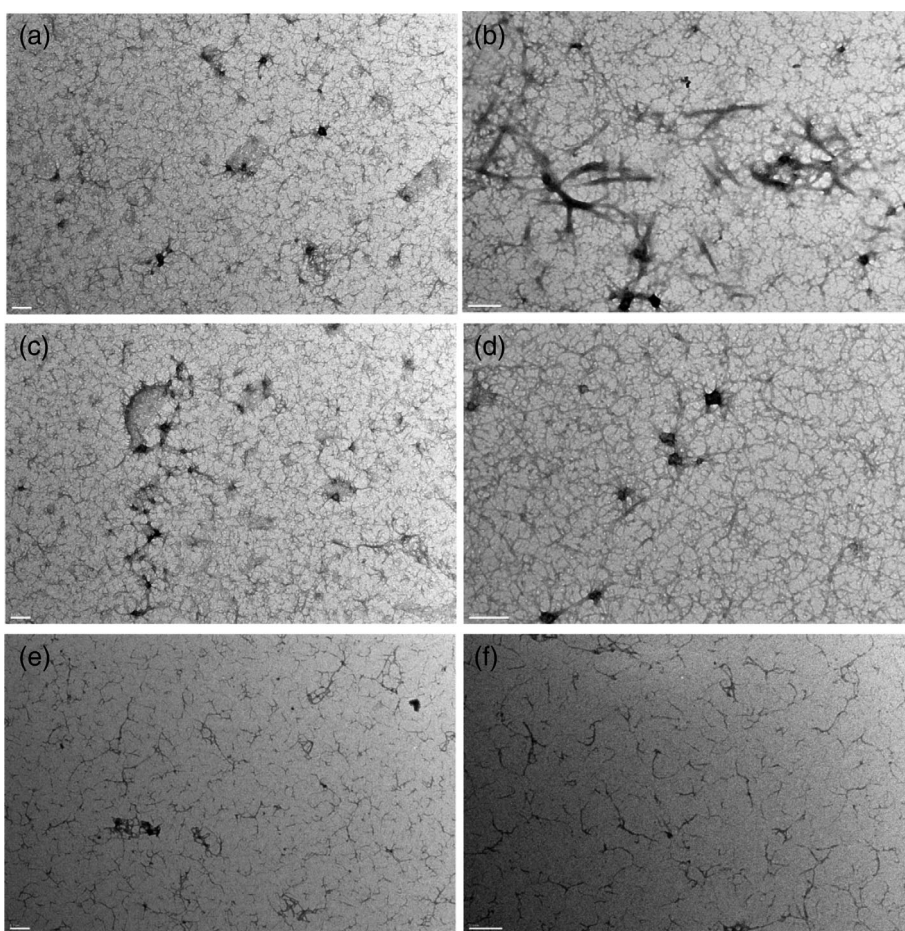

FIGURE 5 The nonspecific aggregates of peptide Col108rr. TEM images of Col108rr in fibril forming buffer were taken after 12 hr incubation at 37°C (a, b), after 24 hr incubation at 37°C (c, d), and after 24 hr incubation at 26°C (e). The peptide in acetic acid before transferring to the fibril-forming buffer is shown in (f). The scale bars are 200 nm in all images

We also conclude from the lack of any mini-fibril assembly in Col108rr that the sequence periodicity is a necessary condition for the formation of *d*-period mini-fibrils. Simply including residues that have high potential to interact with

neighboring helices, and/or stabilize the triple-helix conformation itself, does not facilitate the formation of any specific conformation during the self-assembly. However, instead of strict regular placement of identical residues, the periodicity

should be for residues having similar properties. Such a periodicity of like amino acid residues in the first-order structure of type I collagen was attributed as the factor for the D-period of collagen fibrils.[20,23] It remains to be demonstrated experimentally if such a modification of the sequences of the three sequence units of Col108 or Col877 will form the d-period mini-fibrils. On this account, the interaction curve, however elementary in its algorithm, may prove to be a helpful tool to guide the design of future peptides. As shown in the case of the current study of Col108rr, by focusing on disrupting the periodicity of identical amino acid residues, we overlooked the emergence of a new possible interaction pattern associated with the positions of all hydrophobic residues. It is not clear why the 2U-staggered Col108rr mini-fibrils theoretically revealed by the interaction curve were not observed; low stability and/or lack of the involvement of the interactions of charged residues are among the likely factors. It will be interesting to find out if the 2-unit staggered Col108rr fibrils could form in different buffers, at different incubation temperatures, with increased concentrations, or under other conditions that significantly promote the potential hydrophobic interactions stabilizing this specific conformation.

Despite the incredible tensile strength of connective tissues that utilize collagen fibrils as the molecular scaffold, the triple-helix conformation itself is actually not a very stable one. The melting temperatures of natural collagens are close to the physiological temperature of warm-blooded animals or to ambient temperature in the cases of cold-blooded animals. One study reported the equilibrium melting temperature of the triple helix was even a few degrees lower than physiological temperature.[31] The *in vitro* fibrillogenesis is faster at temperatures lower than that of the melting temperature.[1] Thus, from the point of protein design, it is advantageous to start with a triple helix that is thermally stable for the fibril formation as well as for potential biomedical applications and processing. The thermal stability of short peptides of 30–45 amino acid residues having an unusually high content of Hyp in the Y-position can have an unfolding temperature as high as ∼65°C. However, the effects of Hyp quickly wear off in longer peptides, presumably due to the trade off of the loss of conformational entropy with the gain of enthalpy during the folding of large peptides. The unfolding temperatures of peptides having 100–300 residues are in the range of 36–42°C and are remarkably indifferent to the variations of amino acid composition, sequence, and length.[22,32] Factors such as the imino acid content, Pro in the X or Y positions and especially Hyp in the Y-position,[33] the content of charged residues, especially in the sequences of KGE or KGD,[26] and covalent linking of the three polypeptide chains using disulfide cross-links[34,35] are generally considered factors to be included to increase the stability of a triple helix. In the current study, we found the thermal stability is more closely related to the special salt-bridge sequences of KGE and KGD sequences instead of the absolute number of the charged residues. Having the same number of charged residues but lacking the two salt-bridge sequences, the apparent T_m of Col108rr is about 3°C lower than that of Col108. Inclusion of a nucleation domain can also increase the melting temperature, but only by a few degrees.[32] The overwhelming importance of a nucleation domain such as the foldon domain is to facilitate the in-register alignment of the three polypeptide chains during folding; without the nucleation domain, peptides having more than 100 residues could not fold in a timely manner. The folding of collagens *in vivo* relies on the 200-residue C-propeptide, which is also crucial for the chain selection of collagens consisting of three different polypeptide chains.[36] But, the C-propeptide has to be removed before fibrillogenesis. Mutations that impair the removal of the C-propeptide are a known pathogenic condition associated with the connective tissue disease Osteogenesis Imperfecta (brittle bone disease).[37] In our design, the nucleation domain—the foldon domain—appeared to be accommodated into the mini-fibrils to a certain degree, because of its small size and the unique threefold symmetry in its conformation.[22] Although more studies are underway to define the precise role of the foldon in the self-assembly, any other nucleation domain bulkier in size than the foldon would likely have to be removed enzymatically before designed triple helices could self-assemble to form fibrils.

One other factor that is yet to be evaluated for the potential of broad applications of the design strategy is the hydroxyproline residues in the Y-position. The Hyp is introduced *in vivo* by posttranslational modification in the endoplasmic reticulum before the secretion of the fully folded triple helix. The presence of Hyp is often considered the hallmark of collagens. Yet, more studies have emerged that suggest collagen triple helix without Hyp uphold some of the crucial properties of collagen such as the triple-helix folding,[38–40] the fibril-assembly,[21,41,42] the interactions with metalloproteinase,[43] and the binding of cell receptors and/or other extracellular matrix proteins.[44,45] It is the ultimate goal of our research to eventually incorporate the Hyp into the mini-fibrils using eukaryotic expression systems and/or other approaches. Nevertheless, the ability to form collagen-like staggered min-fibrils by design, albeit without Hyp, sets the stage to advance collagen research to the fibril level. Knowledge of the molecular properties and the mechanical properties of the mini-fibrils can provide insight into the biological functions and properties of collagen fibrils, complementing the peptide studies of triple-helix monomers, and the studies of collagens isolated from tissues.

4 | CONCLUSION

We demonstrated a strategy utilizing long-range sequence periodicity of a peptide to achieve structural periodicity in a higher-order protein molecular assembly. Akin to the heptad repeats of the α-helix coiled coil, a 123-residue, long-range sequence periodicity superimposed on the characteristic Gly-X-Y repeating sequences of collagen triple helix was used to direct the further self-assembly of the triple helices to form supramolecular structures bearing the *D*-period like, axially repeating structure of fibrillar collagens. Further refinement of the design rule will pave the way to develop tailored functional collagen mimetic biomaterials by protein design.

5 | MATERIAL AND METHODS

5.1 | Expression plasmids

The genes of Col877 and Col108R were obtained using the Gene Synthesis service by GenScript Corporation. The codons were optimized for bacterial expression by GenScript. The genes were designed to have two restriction enzyme sites BamHI and EcoRI, respectively, at the 5′ and 3′ ends. The synthesized genes were directly cloned into a modified pET32a(+) expression vector.[40] The sequences of the genes were confirmed by gene sequencing. The Col877 plasmid was sequenced using the forward primer of the T7 promotor and the reverse primer of the T7 terminator (provided by Genewiz Corp.). The Col108rr plasmid was purified and sequenced using two forward primers: primer 5′-ATCCGTGGTATCCCGACTCT-3′ targeting the thioredoxin region and primer 5′-GGGCGATACCGGCCCGGA-3′ targeting the second repeating sequence unit; both primers were synthesized by Macrogen Corporation, and the sequencing was done by Macrogen. The DNA sequencing data were analyzed using the Bioedit software. The products of the genes were fusion proteins in the form of His-tagged thioredoxin-Col877 (Trix-Col877) or His-tagged thioredoxin-Col108rr (Trix-Col108rr). There is a thrombin cleavage site between the thioredoxin and the triple-helix domain for the enzymatic cleavage and removal of the His-tagged Trix.

5.2 | Peptide purification and characterization

The same procedures described in our previous study[21] were used for the expression and purification of the peptides. Briefly, the host cell BL21(DE3) was used for expression. The transformed cells were grown in LB medium containing 0.05 mg/mL ampicillin and induced using 0.2 mM IPTG after the optical density (OD) at 600 nm reached 0.4–0.6. After induction, the cells were grown at ~20°C in a shaker cooled with ice at 200 rpm for ~16 hr. (The shaker does not have a cooling unit; large chunks of ice were added to the incubator at the beginning to lower the temperature to 15°C; at the end of the growth period the temperature was about ~20°C). The fusion protein was first purified using Ni-NTA metal affinity resin (Qiagen Cat# 30210). The His-tagged thioredoxin was subsequently removed by thrombin cleavage and separated from the triple-helical peptides by reverse phase HPLC. The purified peptides were stored as lyophilized powder at 4°C until use.

5.3 | Characterization of the triple helix

The lyophilized protein powder was dissolved in 5 mM acetic acid (pH 4) to a final concentration of 1 mg/mL. The concentration was determined using a NanoDrop 1000 Spectrophotometer with an extinction coefficient of 0.23 at 280 nm for both Col877 and Col108rr. The theoretical extinction coefficients of the peptides were calculated using the ProtParam online tool (https://web.expasy.org/protparam/). After dissolving, the samples were equilibrated at 4°C for ~7 days.

All CD experiments were performed using an AVIV CD spectrometer (AVIV Biomedical, Model 202-01) with a temperature control system (a thermal controller Thermo Neslab Merlin M33 connected to a water bath) and quartz cuvettes with a 1 mm optical path. Spectra were taken at 4°C, in the wavelength range 190–300 nm. Ellipticity measurements were corrected against a buffer baseline using the same cuvette. CD data were analyzed using the Origin software. The raw CD data (in millidegree) were normalized to mean residue molar ellipticity (MRE):

$$[\theta] = \frac{\theta \times m}{c \times l \times n_r}$$

where θ is the ellipticity in millidegree, m is the molecular weight in g/mol; c is the concentration in mg/mL, l is the path length of the cuvette in cm, and n_r is the number of amino acid residues in the peptide.

Thermal stability was determined by temperature melt experiments. The thermal melting profiles were obtained by monitoring the CD signal at 225 nm as the temperature was increased from 4 to 65°C. The equilibration time was 2 min at each temperature (equivalent to an average heating rate of 0.3°C/min). The fraction of folded peptide was calculated based on the equation:

$$\text{Fraction of folded} = \frac{\theta_{observed} - \theta_{monomer}}{\theta_{trimer} - \theta_{monomer}};$$

where $\theta_{observed}$ is the observed signal in millidegree; θ_{trimer} is the extrapolated value using the native baseline, and $\theta_{monomer}$ is the extrapolated value using the denatured baseline in the high temperature region.

5.4 | Fibrillogenesis and TEM imaging

To start fibril formation, protein samples at 1 mg/mL in pH 4 buffer at 4°C were mixed with an equal volume of pre-chilled, double strength neutralization buffer (60 mM TES, 60 mM Na₂HPO₄, and 135 mM NaCl, pH 7.4). Thus, the final concentration of the peptides was 0.5 mg/mL, and the final composition of the fibrillogenesis buffer after mixing was 2.5 mM acetic acid, 30 mM TES, 30 mM Na₂HPO₄, and 67.5 mM NaCl, pH 7.4 ($I = 0.09$). After transferring to the TES buffer, the samples were incubated in a water bath at 37 or 26°C for 6–24 hr. All solutions and buffers were made using ultrapure water.

To prepare for TEM analyses, 3 μL of peptide samples in the specified buffer were placed on a 400 mesh, formvar carbon-coated copper grid (Electron Microscopy Sciences, Cat # FCF400-Cu). After 1 min, the remaining liquid was wicked away slowly with filter paper. Six microliters of 1% sodium phosphotungstate (the staining solution) were immediately added to the grid; after 4 min of staining, the excess staining solution was removed using filter paper. The grid was then rinsed with deionized water. The grid was air-dried for at least 1 hr and then examined under a Zeiss 902 electron microscope or JEM-2100 electron microscope (Jeol Corp.).

5.5 | Interaction curves

The method utilized in calculating the interaction curves has been previously described.[21] In brief, an in-house Perl script was utilized which compared two identical amino acid sequences as one was held constant, and the other was moved by one residue (one shift value) each calculation cycle. For each calculation cycle, the hydrophobic, electrostatic, and total interaction values were returned for directly opposed and immediately flanking neighbor residues. Hydrophobic amino acids considered were V, M, I, L, F, and P; positive amino acids were K and R; negative amino acids were D and E. For the foldon domain, only the residues extending along the linear trajectory from the triple helix were considered: GYIPEAPRDDGEW. The "Normalized Interaction Values" in the interaction curve (Figure 2) are the number of calculated interactions normalized by the total number of residues in contact between two helices at each calculation cycle.

5.6 | Structure modeling

To generate the structural model of the in-register alignment for sequence PQGPRGPRGDKGETGEQGDRGIKGHRG of Col877 in three neighboring triple helices, the coordinate file of peptide (Pro-Pro-Gly)₁₀ (PDB ID 1BKV) was used to, first, "create" the triple helix using the Mutation and Energy Minimization functions of Swiss-PDB-viewer DeepView. The corresponding residues of (Pro-Pro-Gly)₁₀ were substituted according to the desired sequence one residue at a time. A round of energy minimization was conducted after the same residues of the three chains of a triple helix were all mutated. The model was then generated by loading three identical triple helices into the View window. The helices were placed close to each other within the limit of the van der Waals hard-sphere with minimal twisting and rotating of individual helices.

ORCID

Yujia Xu https://orcid.org/0000-0001-6060-1121

REFERENCES

1. Kadler KE, Hojima Y, Prockop DJ. Assembly of collagen fibrils de novo by cleavage of the type I pC-collagen with procollagen C-proteinase. Assay of critical concentration demonstrates that collagen self-assembly is a classical example of an entropy-driven process. J Biol Chem. 1987;262:15696–15701.
2. Fleischmajer R, Perlish JS, Faraggiana T. Rotary shadowing of collagen monomers, oligomers, and fibrils during tendon fibrillogenesis. J Histochem Cytochem. 1991;39:51–58.
3. Cisneros DA, Hung C, Franz CM, Muller DJ. Observing growth steps of collagen self-assembly by time-lapse high-resolution atomic force microscopy. J Struct Biol. 2006;154:232–245.
4. Orgel JP, Irving TC, Miller A, Wess TJ. Microfibrillar structure of type I collagen in situ. Proc Natl Acad Sci U S A. 2006;103:9001–9005.
5. Traub W, Piez KA. The chemistry and structure of collagen. Adv Prot Chem. 1971;25:243–352.
6. Smith JW. Molecular pattern in native collagen. Nature. 1968;219:157–158.
7. Kadler KE, Holmes DF, Trotter JA, Chapman JA. Collagen fibril formation. Biochem J. 1996;316:1–11.
8. Holmes DF, Kadler KE. The 10+4 microfibril structure of thin cartilage fibrils. Proc Natl Acad Sci U S A. 2006;103:17249–17254.
9. Kadler KE, Baldock C, Bella J, Boot-Handford RP. Collagens at a glance. J Cell Sci. 2007;120:1955–1958.
10. Brodsky B, Kaplan DL. Shining light on collagen: Expressing collagen in plants. Tissue Eng A. 2013;19:1499–1501.
11. Raines RT. Stronger and (now) longer synthetic collagen. Adv Exp Med Biol. 2009;611:xci–xcviii.
12. Sherman VR, Yang W, Meyers MA. The materials science of collagen. J Mech Behav Biomed Mater. 2015;52:22–50.
13. Bella J, Eaton M, Brodsky B, Berman HM. Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. Science. 1994;266:75–81.

14. Jones EY, Miller A. Analysis of structural design features in collagen. J Mol Biol. 1991;218:209–219.

15. Gelman RA, Williams BR, Piez KA. Collagen fibril formation. Evidence for a multistep process. J Biol Chem. 1979;254: 180–186.

16. Bachinger HP, Bruckner P, Timpl R, Prockop DJ, Engel J. Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Eur J Biochem. 1980;106:619–632.

17. Nestler FH, Hvidt S, Ferry JD, Veis A. Flexibility of collagen determined from dilute solution viscoelastic measurements. Biopolymers. 1983;22:1747–1758.

18. Privalov PL, Tictopulo EI, Tischenko VM. Stability and mobility of the collagen structure. J Mol Biol. 1979;127:203–216.

19. Prockop DJ, Fertala A. The collagen fibril: The almost crystalline structure. J Struct Biol. 1998;122:111–118.

20. Hulmes DJ, Miller A, Parry DA, Piez KA, Woodhead-Galloway J. Analysis of the primary structure of collagen for the origins of molecular packing. J Mol Biol. 1973;79:137–148.

21. Kaur PJ, Strawn R, Bai H, et al. The self-assembly of a mini-fibril with axial periodicity from a designed collagen-mimetic triple helix. J Biol Chem. 2015;290:9251–9261.

22. Strawn R, Chen F, Jeet Haven P, et al. To achieve self-assembled collagen mimetic fibrils using designed peptides. Biopolymers. 2018;109:e23226.

23. Hulmes DJ, Miller A, Parry DA, Woodhead-Galloway J. Fundamental periodicities in the amino acid sequence of the collagen alpha1 chain. Biochem Biophys Res Commun. 1977;77:574–580.

24. Persikov AV, Ramshaw JAM, Brodsky B. Prediction of collagen stability from amino acid sequence. J Biol Chem. 2005;280: 19343–19349.

25. Piez KA. Structure and assembly of the native collagen fibril. Connect Tissue Res. 1982;10:25–36.

26. Persikov AV, Ramshaw JA, Kirkpatrick A, Brodsky B. Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. Biochemistry. 2005;44:1414–1422.

27. Miles CA. Kinetics of collagen denaturation in mammalian lens capsules studied by differential scanning calorimetry. Int J Biol Macromol. 1993;15(5):265–271.

28. Engel J, Bachinger HP. Cooperative equilibrium transitions coupled with a slow annealing step explain the sharpness and hysteresis of collagen folding. Matrix Biol. 2000;19:235–244.

29. Engel J, Chen HT, Prockop DJ, Klump H. The triple helix in equilibrium with coil conversion of collagen-like polytripeptides in aqueous and nonaqueous solvents. Comparison of the thermodynamic parameters and the binding of water to (L-Pro-L-Pro-Gly)n and (L-Pro-L-Hyp-Gly)n. Biopolymers. 1977;16:601–622.

30. Persikov AV, Xu Y, Brodsky B. Equilibrium thermal transitions of collagen model peptides. Protein Sci. 2004;13:893–902.

31. Leikina E, Mertts MV, Kuznetsova N, Leikin S. Type I collagen is thermally unstable at body temperature. Proc Natl Acad Sci U S A. 2002;99:1314–1318.

32. Yu Z, Brodsky B, Inouye M. Dissecting a bacterial collagen domain from *Streptococcus pyogenes*: sequence and length-dependent variations in triple helix stability and folding. J Biol Chem. 2011;286: 18960–18968.

33. Berg RA, Prockop DJ. Purification of (14C) protocollagen and its hydroxylation by prolyl-hydroxylase. Biochemistry. 1973;12: 3395–3401.

34. Bachinger HP. The influence of peptidyl-prolyl cis-trans isomerase on the in vitro folding of type III collagen. J Biol Chem. 1987;262: 17144–17148.

35. Davis JM, Bachinger HP. Hysterisis in the triple helix-coil transition of type III collagen. J Biol Chem. 1993;268:25965–25972.

36. Kielty CM, Hopkinson I, Grant ME. The collagen family: Structure, assembly, and organization in the extracellular matrix. In: Royce PM, Steinmann B, editors. Connective tissue and its heritable disorders. New York, NY: Wiley-Liss, 1993; p. 103–147.

37. Lindahl K, Barnes AM, Fratzl-Zelman N, et al. COL1 C-propeptide cleavage site mutations cause high bone mass osteogenesis imperfecta. Hum Mutat. 2011;32:598–609.

38. Toman PD, Chisholm G, McMullin H, et al. Production of recombinant human type I procollagen trimers using a four-gene expression system in the yeast *Saccharomyces cerevisiae*. J Biol Chem. 2000;275:23303–23309.

39. Yu Z, An B, Ramshaw JA, Brodsky B. Bacterial collagen-like proteins that form triple-helical structures. J Struct Biol. 2014;186:451–461.

40. Xu K, Nowak I, Kirchner M, Xu Y. Recombinant collagen studies link the severe conformational changes induced by osteogenesis imperfecta mutations to the disruption of a set of interchain salt bridges. J Biol Chem. 2008;283:34337–34344.

41. Olsen DR, Leigh SD, Chang R, et al. Production of human type I collagen in yeast reveals unexpected new insights into the molecular assembly of collagen trimers. J Biol Chem. 2001;276:24038–24043.

42. Perret S, Merle C, Bernocco S, et al. Unhydroxylated triple helical collagen I produced in transgenic plants provides new clues on the role of hydroxyproline in collagen folding and fibril formation. J Biol Chem. 2001;276:43693–43698.

43. Yu Z, Visse R, Inouye M, Nagase H, Brodsky B. Defining requirements for collagenase cleavage in collagen type III using a bacterial collagen system. J Biol Chem. 2012;287:22988–22997.

44. An B, Abbonante V, Xu H, et al. Recombinant collagen engineered to bind to discoidin domain receptor functions as a receptor inhibitor. J Biol Chem. 2016;291:4343–4355.

45. An B, Abbonante V, Yigit S, Balduini A, Kaplan DL, Brodsky B. Definition of the native and denatured type II collagen binding site for fibronectin using a recombinant collagen system. J Biol Chem. 2014;289:4941–4951.