Winter 1-14-2021

Design of Fibril Forming Collagen Mimetic Peptides: Heterotrimers and Nucleation Domains

Sally Tan
CUNY Hunter College

DESIGN OF FIBRIL FORMING COLLAGEN MIMETIC PEPTIDES: HETEROTRIMERS AND NUCLEATION DOMAINS

by

Sally Tan

Submitted in partial fulfillment
of the requirements for the degree of
Master of Arts Biochemistry, Hunter College
The City University of New York

2021

01/14/2021	Dr. Yujia Xu
Date	Thesis Sponsor
01/14/2021	Dr. Nancy Greenbaum
Date	Second Reader

Abstract

Collagen is the most abundant protein in the human body and plays an important role in extracellular matrix organization, support, and elasticity. Of the different types of collagens, Type I Collagen makes up 90% of the human body's collagen content. It is a heterotrimer with two identical polypeptide chains (α_1(I) chain) and one polypeptide chain with different amino acid sequence (α_2(I) chain). Naturally, Type I Collagen is also the most studied collagen. Many Collagen Mimetic Peptides use sequences derived from human Type I Collagen to explore its folding, structure, and bioactivity. In the bioengineering of collagen mimetic peptides for biomaterial, there is a great interest in forming peptides with the heterotrimeric and fibril forming characteristics of Type I Collagen. The previously F877 homotrimer includes a collagen domain from the α_1(I) chain and forms stable triple helices. In hopes of creating a heterotrimeric triple helix where the collagen domain mimics that of Type I Collagen, we designed new mimetic peptides based on F877 using collagen sequences from the α_2(I) chain. We report two new constructs α_1TRX^- and α_2TRX^- that successfully formed triple helices and can be used in future designs for homo- or heterotrimers. We show that triple helix folding is dependent on the triple helical region amino acid sequence, the location of the cross-linking domains as well as the nucleation domain. We have also completed a preliminary study using the NC2 domain of Type IX Collagen in an attempt to form heterotrimeric peptides having the desired chain register.

Introduction

Type I Collagen

Collagen, being the most abundant protein in the human body, plays a fundamental role in extracellular matrix organization, support, and elasticity. Each collagen unit encompasses three individual polyproline II-like peptide chains that form a triple helix with a one residue stagger (Brodsky and Ramshaw, 1997; Engel and Bachinger, 2005). The triple helical structure of the protein is attributed to the distinctive $(Gly\text{-}X\text{-}Y)_n$ amino acid repeats of collagen peptides (Canty and Kadler, 2005). When folded, the X and Y amino residues of the triplet, commonly proline and hydroxyproline, become solvent exposed while glycine residues are buried within the helix (Boudko et al., 2012; Hulmes et al., 1973). Collagen chains are initially synthesized as procollagen with non-triple helical C- and N-terminal propeptide regions. In particular, the C-propeptide domain is a key player in chain selection and chain register, as well as serving as a nucleation domain for triple helix folding (Bulleid, 1996; Doyle and Smith, 1998). The globular propeptide regions are proteolyzed upon secretion from the cell, allowing individual collagen triple helices to further self-associate into supramolecular structures (Canty and Kadler, 2005). This fibrillar form of collagen allows for its crucial load-bearing role in bones, tendons, and ligaments (Varma et al., 2016). Individual collagen trimers self-associate in a parallel bundle with a characteristic D-Period of 67 nm (Petruska & Hodge, 1964; Orgel et al., 2006). The collagen fibril periodicity is visualized with negative staining of TEM (Transmission Electron Microscopy) methods, revealing the 27 nm (0.4 D) overlap region and 40 nm (0.6 D) gap region (Piez and Miller, 1974). Such a structure reveals a series of light (overlap) and dark (gap) striation patterns in TEM images.

Of the 28 types of collagens known today, Type I Collagen is the most common collagen found in the human body, comprising almost 90% of the body's collagen content (Kadler, 2007). Encoded by the COL1A1 and COL1A2 genes, Type I Collagen is made up of two α_1 chains and one α_2 chain that parallelly twist together in triple helical formation (Varma et al., 2016). While normal Type I Collagen is found in the heterotrimeric form, its homotrimeric (α_1(I)) isoform also occurs naturally. It is noted that while the α_1(I) chain forms collagen homotrimers, the predominant consensus in the field is that α_2(I) chains do not self-associate in triple helices. However, there is little understanding to why α_2 homotrimers do not naturally occur given the extremely high sequence identity and similarity between the two chains (Xu and Kirchner, 2021). A study suggests that charge repulsion between Lys129 of one chain and Lys247 of the adjacent chain in the α_2(I) C-propeptide Chain Recognition Sequence (CRS) forbids the formation of α_2(I) homotrimers (Sharma et al., 2017).

The α_1(I) homotrimer is often associated with cellular abnormalities and disease, such as fibrosis and cancer (Chang et al., 2012). The isoform is also observed in Brittle Bone Disease (Osteogenesis Imperfecta), where mutations in the COL1A1 or COL1A2 gene results in the production of homotrimeric Type I Collagen over normal heterotrimeric Type I Collagen (Xu et al., 2008). Homotrimeric Type I Collagen shows increased resistance to MMP-1 (Matrix Metalloprotease 1) although the reasoning is still not fully understood (Chang et al., 2012). Some attribute increased collagenase resistance to the increased stability of the homotrimer as it lacks the α_2 chain (Han et al., 2010). This is because the α_2 chain disrupts hydroxylation and hydrogen binding patterns present in homotrimeric Type I Collagen. (Nerenberg and Collin, 2008) Thus, providing flexibility for local unfolding that precedes proteolysis at the MMP-1 cleavage site (Mekkat et al., 2018).

Collagen Mimetic Studies

In addition to the role of collagen in diseases and aging, collagen is also a promising biomaterial scaffold used in many medical treatments. To better study the mechanics and function of collagen, there has been an increase of collagen mimetic studies over the years. Such a field encompasses the expression of natural collagen in non-human expression systems as well as the bioengineering of both native collagen and collagen-like sequences (Brodsky and Ramshaw, 2017; Xu and Kirchner, 2021). As many of collagen's dynamic roles are due to its fibrillar structure, it is of great interest to synthesize fibril forming biomaterial. Likewise, with the most abundant collagen protein being Type I Collagen also brings about interest in synthesizing heterotrimeric mimetic peptides. Mimetic peptides allow researchers to explore the folding mechanism, structure, and stability of collagen and collagen fibrils with the end-goal of engineering peptides to desired parameters. The use of bioengineered collagens deviates from the dependence of animal sources which are not applicable for biomedical usage due to contamination during the purification process.

F877

The Xu lab has been designing collagen mimetics peptides using native collagen sequences in bacterial expression systems. The laboratory hopes to synthesize heterotrimeric fibril forming collagen mimetic peptides akin to Type I Collagen. In past research, they have successfully formed triple helices without hydroxyproline as well as mini-fibrils showcasing 35 nm d-periodicity (Chen et al., 2019; Kaur et al., 2014; Strawn et al., 2018). They show a minimum of two identical sequence unit (collagen domain) repeats are sufficient to result in mini-fibrils with periodicity (Chen et al., 2019). The one sequence unit construct, F877, while unable to form fibrils, forms a stable triple helix with a melting temperature ~38℃ (Xu et al.,

2018). This is comparable with the melting temperature of native Type I Collagen, which lies just under that of physiological temperature (Leikina et al., 2002). The structural motifs and amino acid sequence of F877 are shown below in Figure 1.

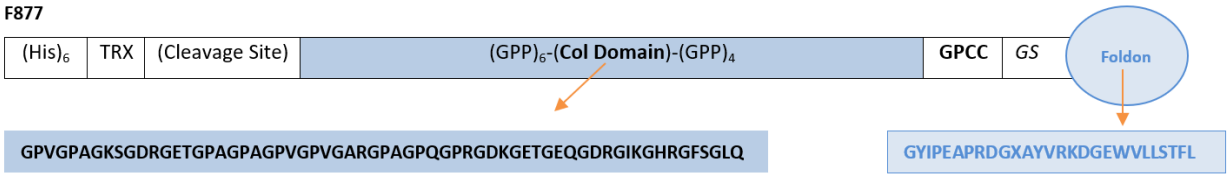


Figure 1. F877 Design. *The structural motifs of fusion protein F877 from the N-terminus (left) to C-terminus (right). His-tagged thioredoxin is followed by the thrombin cleavage site. The collagen domain is a 63 amino acid long sequence from Human Type I Collagen α_1 chain (Residues 877-940). Gly-Ser sequences are introduced into the peptide from restriction enzymes, and one is placed before the Foldon for flexibility. A cysteine knot is formed in the homotrimer due to the GPCC (bolded) at the C-terminal end.*

The collagen domain is flanked on both sides with $(GPP)_n$ repeats for stabilization factors and a GPCC sequence at the C-terminal forms a cysteine-knot to aid trimerization (Xu et al., 2008). The fusion peptide includes a thioredoxin domain to increase protein solubility in *E. coli* cytoplasm, preventing the formation of inclusion bodies (LaVille et al., 1993; Yasukawa et al., 1995). This His-tagged thioredoxin calls for further purification steps due to its interactions with the triple helical domain. The domain is removed by thrombin at the cut site on the c-terminal side of the His-tagged thioredoxin. However, this purification step runs into a common problem of the target protein precipitating out as thrombin cleavage requires dialyzing the protein into neutral pH buffer. Removal of the enzyme and cleaved fragment by High Performance Liquid Chromatography (HPLC) subsequently also lowers the final yield.

This brings up the notion whether the thioredoxin domain can be removed from the peptide design. The construct α_1TRX^- is identical to that of F877 sans the thioredoxin sequence. As the bulk of protein yield loss occurs during dialysis to neutral pH and thrombin cleavage,

bypassing these steps should greatly enhance the purification yield. Using F877 as a positive control, if the α_1TRX$^-$ peptide can successfully trimerize into triple helical formation, it can both greatly shorten the purification process as well as increase overall purification yield.

Foldon

The nucleation domain that is incorporated in the design of F877 is the T4 Fibritin Foldon. The highly stable β-hairpin homotrimer, having a melting temperature ~60℃ has shown to be a successful trimerization domain (Frank, 2001; Meier et al., 2004). This is supported by the 2008 study by Xu et al., where homotrimeric F877 forms a stable triple helix. However, in attempt to synthesize a Type I Collagen mimetic peptide brings up the consideration of chain selection and chain register. The Foldon domain is then limited by its homotrimeric nature and is not applicable in designing heterotrimers. Furthermore, its bacteriophage origin becomes a hindrance in designing collagen mimetic peptides for biomedical purposes. Finding a viable nucleation domain from human protein sequences greatly broadens the application of synthetic collagen studies.

Type IX Collagen NC2 Domain

Recent studies on the non-collagenous (NC) domains of different types of collagens bring promising candidates for trimerization domains. A 2010 study by Boudko et al. found the NC2 domain of the heterotrimeric Collagen IX can be used as a triple helix nucleation domain. Collagen IX is a fibril-associated collagen with interrupted triple helices (FACIT) and a heterotrimer of three distinct polypeptide chains (α_1, α_2 and α_3). The NC2 region (~38 amino acids) itself is also a heterotrimer with three distinct alpha chains (Figure 2).

NC2α$_1$	GRAPTDQHIKQV**C**MRVIQEHFAEMAASLKRPDSGAT
NC2α$_2$	GRDATDQHIVDVALKMLQEQLAEVAVSAKREALGAV
NC2α$_3$	GKEASEQRIREL**C**GGMISEQIAQLAAHLRKPLAPGSI

Figure 2. Sequence of Type IX Collagen NC2 Domain. *The amino acid sequence of the non-collagenous domain 2 in Type IX Collagen. The three chains are ~38 amino acids long and come together to form an alpha-helix coiled coil. The cysteine residues that form disulfide linkage between the NC2α$_1$(IX) and NC2α$_3$(IX) chains are bolded.*

In 2012, Boudko and Bachinger expanded on the earlier experiment to show that the NC2 domain not only functions as a trimerization domain but also selects for chain register. This greatly opens possibilities in synthetic collagen design—ultimately allowing for a "toolkit" to assemble trimers in any desired chain selection and register. In theory, both homo- and hetero-trimers can be assembled using the NC2 domain. The human origin of the NC2 domain would also be preferred over the bacteriophage derived Foldon in application to biomaterial. The NC2 domain provides a promising alternative nucleation domain to the Foldon.

Boudko et al. designed peptides with a short collagenous domain (12 residues) using sequences from Human Type I Collagen α_1 and α_2 chains. The collagen domain was flanked by $(GPP)_4$ on the N-terminal side and $(GPP)_3$ on the C-terminal side. The Type IX Collagen NC2 domain (either α_1, α_2, and α_3 chains) was appended at the C-terminal end. A GPCC sequence was placed at the N-terminus for formation of a cysteine knot to increase trimer stability. We replaced the short collagen sequence with the 63 amino long sequence used in F877 to synthesize peptides α_1-NC2α_1, α_1-NC2α_2 and α_1-NC2α_3. The corresponding sequence (residues 877-940) from Human Type I Collagen α_2 chain was used to synthesize α_2-NC2α_1, α_2-NC2α_2 and α_2-NC2α_3.

We hypothesize that the NC2 domain can be used as the nucleation domain to develop heterotrimeric collagen mimetic peptides; furthermore, because it is more native-like, the heterotrimeric mini-fibril will expand our scope of collagen research and ability to create new biomaterials. In order to achieve the heterotrimer peptides at a high expression yield, we further modified the existing expression construct. We designed the construct α_1TRX$^-$ using F877 as a positive control to gauge whether stable triple helices could be formed without thioredoxin. Six constructs were constructed to using native sequences from the α_1 and α_2 chains of Type I Collagen to each of the three alpha chains of the NC2 domain. We learned that the folding and stability of the designed peptides depends on several factors, including the amino acid residues of the triple helix domain, the location of the GPCC cross-linking domain, and the nucleation domain.

Materials and Methods

Construction of expression plasmids

DNA sequences for the designed peptides (α_1TRX$^-$, α_2TRX$^-$, α_2TRX$^-$-N, α_1-NC2α_1, α_1-NC2α_2, α_1-NC2α_3, α_2-NC2α_1, α_2-NC2α_2 and α_2-NC2α_3) were codon optimized and subcloned into a modified pET32a (+) vector using GenScript® Gene Synthesis Services. Constructs were inserted between the EcoRI and NcoI restriction enzyme sites, with the T7 promoter upstream of the N-terminus and the T7 terminator downstream of the C-terminus. Each peptide included a poly-histidine tag at the N-terminus and a nucleation domain at the C-terminus. α_1TRX$^-$, α_2TRX$^-$ and α_2TRX$^-$-N contain the Foldon domain while α_1-NC2α_1, α_1-NC2α_2, α_1-NC2α_3, α_2-NC2α_1, α_2-NC2α_2 and α_2-NC2α_3 include the NC2 domain.

Transformation into Competent Cells

The peptides were expressed using Invitrogen One Shot™ BL21(DE3)pLysS Chemically Competent *E. coli* cells (Cat. #C606003) optimized for T7-promoter expression systems. Each construct was separately transformed into competent cells by heatshock technique indicated in the product manual. Transformation solution was then plated onto LB Agar plates with 100 µg/mL Ampicillin to select for antibiotic resistance. Plasmid DNA purification was performed on transformed colonies using Promega Wizard® Plus SV Minipreps DNA Purification System (Cat. #A1330) as instructed by product protocol. DNA concentrations were quantified using Thermo Scientific™ NanoDrop™ One Microvolume UV-Vis Spectrophotometer. Gene sequences were then confirmed with Sanger Sequencing services provided by GENEWIZ, Inc

before proceeding onto expression and purification protocols. The T7 Promoter and T7 Terminator primers were used to sequence the plasmid DNA samples.

Expression and Purification of Target Constructs

Transformed colonies were grown separately in LB Media (10 g Tryptone, 5 g Yeast Extract, 10 g NaCl, 100 µg/mL Ampicillin per liter) under sterile conditions. 10 mL seeding cultures were incubated overnight at 37°C, 225 RPM shaking settings before inoculation into growth media (1 L of 100 µg/mL Ampicillin LB Media) the following day. Cells were grown at 37°C, 225 RPM until optical density at 600 nm reached 0.5 AU. Expression of the targeted protein was initiated upon the addition of 0.1 mM final concentration of Isopropyl β-D-1-thiogalactopyranoside (IPTG). Samples were induced overnight in an incubator (16°C, 225 rpm) for 18 hours. The NC2 containing constructs, α_1-NC2α_1, α_1-NC2α_2, α_1-NC2α_3 (will now be mentioned as α_1-NC2) and α_2-NC2α_1, α_2-NC2α_2, α_2-NC2α_3 (will now be mentioned as α_2-NC2) were mixed in equal volume ratios after induction and co-purified.

The growth media were centrifugated to collect the cells before resuspension in Lysis Buffer (50 mM Tris, 300 mM NaCl, pH 7) with added DNAse, lysozyme and phenylmethylsulfonyl fluoride (PMSF). Cells were then lysed by sonication before another round of centrifugation to remove cell debris. The supernatant containing the targeted protein was bound to 2 mL slurry of Thermo Scientific™ HisPur™ Ni-NTA Resin (Cat. #PI88221). The bound resin was loaded onto a column and washed first with 10 mL 20 mM Imidazole Lysis Buffer (50 mM Tris, 300 mM NaCl, 1 mM PMSF, pH 7) and subsequently 10 mL 50 mM Imidazole Lysis Buffer (50 mM Tris, 300 mM NaCl, pH 7). The protein constructs were eluted with 500 mM Imidazole Lysis Buffer (50 mM Tris, 300 mM NaCl, pH 7) in 1 mL elution

fractions. Elution fractions were run on a 17% poly-acrylamide gel under reducing conditions. Sample migration was monitored with GenScript® Broad Multi Color Pre-Stained Protein Standard (Cat. #M00624). The SDS-PAGE gel was then stained with Coomassie Blue dye for visualization.

Sample migration was also confirmed using Western Blot for the His-tag epitope. The proteins were transferred onto Invitrogen's PDVF (polyvinylidene difluoride) membrane (Cat. #LC2002) after SDS-PAGE. The membrane was then blocked with 5% milk on a shaker at room temperature for an hour. Proteins were probed with Invitrogen's HRP-conjugated 6x-His Tag Monoclonal Antibody (Cat. #MA1-21315-HRP) in 1:2000 dilution overnight with slight agitation at 5°C. It was subsequently washed with 1x-TBS-T (8.8 g NaCl, 0.2g KCl, 3g Tris Base, 500 µL Tween-20 per liter) on a shaker at room temperature. Chemiluminescence was detected using MilliporeSigma™ Immobilon™ Western Chemiluminescent HRP Substrate (ECL) (Cat. #WBKLS0050). The Western Blot films was obtained with a Konika SRX-101A X-Ray Film Developer.

Collected elution fractions were then dialyzed extensively against 5 mM Acetic Acid (pH 4.5) at low temperature (4°C). The protein was kept incubated at 4°C for at least three days after dialysis to initiate trimerization. Protein concentrations were obtained using a DU800 Spectrophotometer (Beckman Coulter Inc, USA). The mg/mL concentration values were calculated using absorbance values at 280 nm and Abs 0.1% extinction coefficients provided by ExPASy ProtParam (Table 1).

Characterization using Circular Dichroism

Circular dichroism spectroscopy data was obtained using Chirascan™ V100 (Applied Photophysics Ltd, UK). Data parameters were manipulated using the Pro-Data Chirascan™ program and experimental temperatures were monitored with the Chirascan™ CS/PSM Turret T1 temperature probe. Concentrations of samples measured were determined by the Beckman Coulter DU800 Spectrophotometer. Samples for CD experiments remained in 5 mM Acetic Acid (pH 4.5). Scans of constructs were taken from 190-260 nm in 1 mm pathlength quartz cuvettes at 5°C. Thermal melt experiments were conducted from 5-60°C increasing one degree per minute. A scan was taken from 190-260 nm at every one-degree increment, with an equilibration time of one minute at every temperature point. Sample spectra was corrected using baseline subtraction of the buffer spectra before being normalized for concentration and converted into mean residual ellipticity. Normalized circular dichroism data was analyzed with the open-source software SciDavis, producing a melting curve by monitoring CD values at the negative peak wavelength per construct. The thermal transition profiles (temperature vs. fraction folded) were fitted to a Boltzmann's distribution curve and melting temperature (T_m) was determined where the protein was 50% folded (Fraction Folded = 0.5)

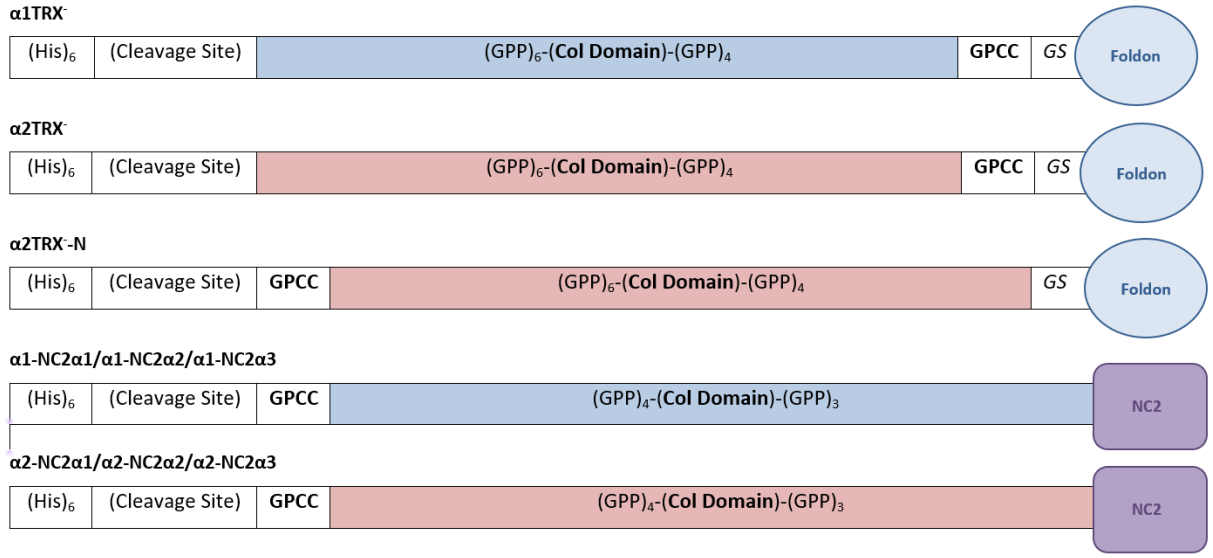
Results

Peptide Design

In designing heterotrimeric peptides, our first goal was to synthesize a F877 heterotrimer where the collagen domain mimics that of Type I Collagen. To maximize purification yield, we optimized the F877 design by removing the thioredoxin domain. α_1TRX^- tested whether the peptides could form triple helices without the domain. While F877 and α_1TRX^-'s collagenous domain derives from the α_1 chain of Human Type I Collagen, we synthesized two new constructs replacing the collagen domain with the corresponding sequence from the α_2 chain of Type I Collagen. The two α_2-Foldon constructs differ only in the placement of the GPCC sequence, one on the N-terminal side of the triple helical region and the other on the C-terminal side. From here on, the constructs will be referred to as α_2TRX^--N and α_2TRX^-.

We then designed peptides where the Foldon was replaced by the NC2 domain. Sequences from the α_1(I) and α_2(I) chains were separately linked to each of the three alpha chains of the NC2 domain, yielding six different constructs. Once the homotrimers form stable triple helices, these constructs can used to form heterotrimers. We aimed to utilize the NC2 domain to create a trimerization "toolkit", where we can form homotrimers in any chain composition and chain register we desired. A single tyrosine was included in the NC2 domain containing peptides to monitor absorbance at 280 nm.

A



B

α1TRX-	HHHHHH*GS*GLVPR*GS*GPPGPPGPPGPPGPPGPPGPPGPVGPAGKSGDRGETGPAGPAGPVGPVGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQGPPGPPGPP GPP**GPCC***GGS*GYIPEAPRDGXAYVRKDGEWVLLSTFL
α2TRX-	HHHHHH*GS*GLVPR*GS*GPPGPPGPPGPPGPPGPPGPPGPVGPAGKHGNRGETGPSGPVGPAGAVGPRGPSGPQGIRGDKGEPGEKGPRGLPGLKGHNGLQGPPGPPGPP GPP**GPCC***GS*GYIPEAPRDGQAYVRKDGEWVLLSTFL
α2TRX-N	HHHHHH*GS*GLVPR*GS***GPCC**GPPGPPGPPGPPGPPGPPGPPGPVGPAGKHGNRGETGPSGPVGPAGAVGPRGPSGPQGIRGDKGEPGEKGPRGLPGLKGHNGLQGPPGP PGPPGPP*GS*GYIPEAPRDGQAYVRKDGEWVLLSTFL
α1-NC2α1	HHHHHH*GS*GLVPR*GS*Y**GPCC**GPPGPPGPPGPPGPPGPVGPAGKSGDRGETGPAGPAGPVGPVGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQGPPGPPGPPG RAPTDQHIKQVCMRVIQEHFAEMAASLKRPDSGAT
α1-NC2α2	HHHHHH*GS*GLVPR*GS*Y**GPCC**GPPGPPGPPGPPGPPGPVGPAGKSGDRGETGPAGPAGPVGPVGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQGPPGPPGPPG RDATDQHIVDVALKMLQEQLAEVAVSAKREALGAV
α1-NC2α3	HHHHHH*GS*GLVPR*GS*Y**GPCC**GPPGPPGPPGPPGPPGPVGPAGKSGDRGETGPAGPAGPVGPVGARGPAGPQGPRGDKGETGEQGDRGIKGHRGFSGLQGPPGPPGPPG KEASEQRIRELCGGMISEQIAQLAAHLRKPLAPGSI
α2-NC2α1	HHHHHH*GS*GLVPR*GS*Y**GPCC**GPPGPPGPPGPPGPPGPVGPAGKHGNRGETGPSGPVGPAGAVGPRGPSGPQGIRGDKGEPGEKGPRGLPGLKGHNGLQGPPGPPGPPG RAPTDQHIKQVCMRVIQEHFAEMAASLKRPDSGAT
α2-NC2α2	HHHHHH*GS*GLVPR*GS*Y**GPCC**GPPGPPGPPGPPGPPGPVGPAGKHGNRGETGPSGPVGPAGAVGPRGPSGPQGIRGDKGEPGEKGPRGLPGLKGHNGLQGPPGPPGPPG RDATDQHIVDVALKMLQEQLAEVAVSAKREALGAV
α2-NC2α3	HHHHHH*GS*GLVPR*GS*Y**GPCC**GPPGPPGPPGPPGPPGPVGPAGKHGNRGETGPSGPVGPAGAVGPRGPSGPQGIRGDKGEPGEKGPRGLPGLKGHNGLQGPPGPPGPPG KEASEQRIRELCGGMISEQIAQLAAHLRKPLAPGSI

Figure 3. The design of peptides. *(**A**) The structural motifs of the separate constructs starting from the N-terminal (left) to the C-terminal (right). Trimers of each construct will come together with a one residue-stagger. Collagen domain from α_1(I) are highlighted in blue while collagen domain*

*from α_2(I) are highlighted in red. The Foldon domain is depicted as a blue circle while the NC2 domain is depicted as a purple box. Gly-Ser (italicized) are introduced by restriction enzymes. A cysteine knot is formed due to the Gly-Pro-Cys-Cys (bolded) sequence. A Gly-Ser is inserted before the Foldon for flexibility. (**B**) The protein sequences of the separate constructs with the respective collagen domain sequence for each peptide is underlined.*

α_1TRX$^-$ and the α_1NC2 constructs contains the 63 amino acid long fragment (residues 877-940) from Human Type I Collagen α_1 chain. The α_2-NC2 and α_2TRX$^-$ constructs, on the other hand, contain the corresponding peptide sequence from Human Type I Collagen α2 chain. The collagen sequence in each peptides nests between $(GPP)_n$ sequence repeats. Peptide motifs and amino acid sequences are found in Figure 3. Molecular weights and extinction coefficients for each construct were obtained via the ExPASy ProtParam Tool hosted by the SIB Swiss Institute of Technology (https://web.expasy.org/protparam/) and are listed in Table 1.

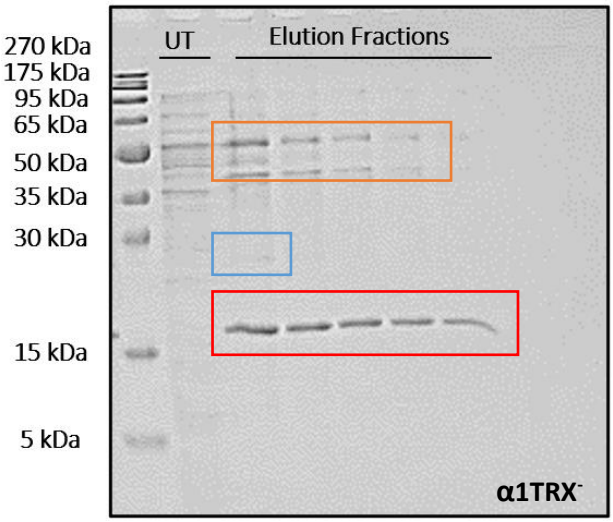
Construct	**Molecular Weight (Da)**	**Extinction Coefficient (M^{-1} cm^{-1})**	**Extinction Coefficient (Abs 0.1% (=1 mg/mL))**
α_1TRX$^-$	13667.31	8480	0.626
α_2TRX$^-$	13636.20	8480	0.622
α_2TRX$^-$-N	13636.20	8480	0.622
α_1-NC2α_1	13791.32	1490	0.108
α_1-NC2α_2	13644.15	1490	0.109
α_1-NC2α_3	13770.37	1490	0.108
α_2-NC2α_1	13800.46	1490	0.108
α_2-NC2α_2	13653.29	1490	0.109
α_2-NC2α_3	13779.51	1490	0.108

Table 1. Molecular Weights and Extinction Coefficients of Constructs. *Listed in the table are the calculated molecular weights and extinction coefficients per single chain of each peptide construct*

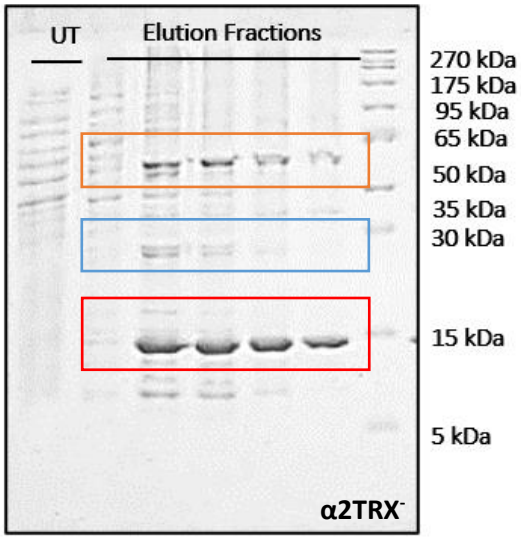
The Expression and Purification of the Peptides

Each construct was purified separately and purified via nickel-affinity resin apart from the two NC2 containing homotrimers (α_1-NC2, α_2-NC2) where three individual chains of the desired homotrimer were mixed and co-purified. Collected elution fractions were run on a 17% poly-acrylamide gel during SDS-PAGE. The respective monomer, dimer and trimer are highlighted for each construct. All the constructs appeared as expected on the SDS-PAGE results as seen in Figure 4. All five constructs all have a molecular weight ~14 kDa and appeared close to the 15 kDa band as indicated by the molecular weight marker. Under 1x SDS reducing agent conditions, there were still presence of the dimer and trimer configurations. The dimer (~28 kDa) appeared just below the 30 kDa marker while the trimer (~42 kDa) appeared just below the 50 kDa. These respective bands are not seen in the untransformed cell (negative control). The monomer, dimer and trimer bands were also confirmed by the western blot using antibodies against the His-tag.

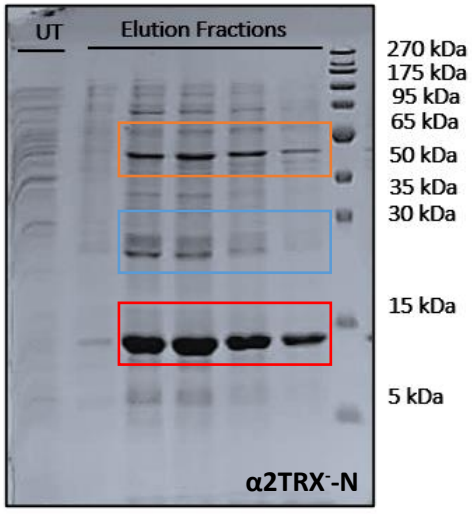
A


UT
Elution Fractions
270 kDa
175 kDa
95 kDa
65 kDa
50 kDa
35 kDa
30 kDa
15 kDa
5 kDa
α1TRX-

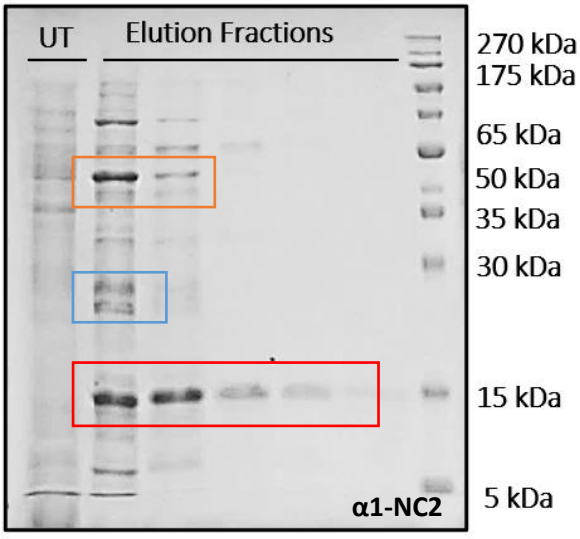
B


UT
Elution Fractions
270 kDa
175 kDa
95 kDa
65 kDa
50 kDa
35 kDa
30 kDa
15 kDa
5 kDa
α2TRX-

C


UT
Elution Fractions
270 kDa
175 kDa
95 kDa
65 kDa
50 kDa
35 kDa
30 kDa
15 kDa
5 kDa
α2TRX-N

D


UT
Elution Fractions
270 kDa
175 kDa
65 kDa
50 kDa
35 kDa
30 kDa
15 kDa
5 kDa
α1-NC2

E

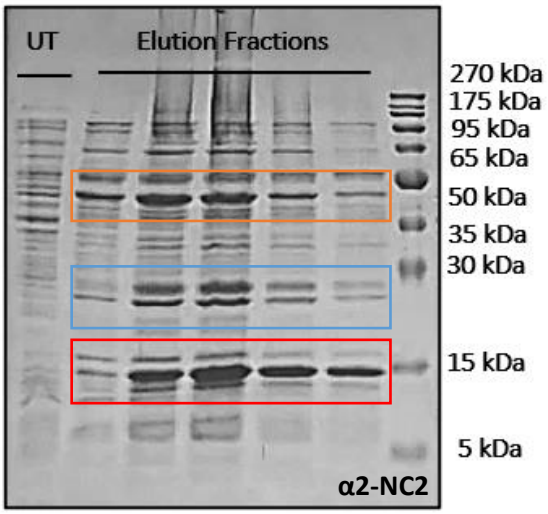


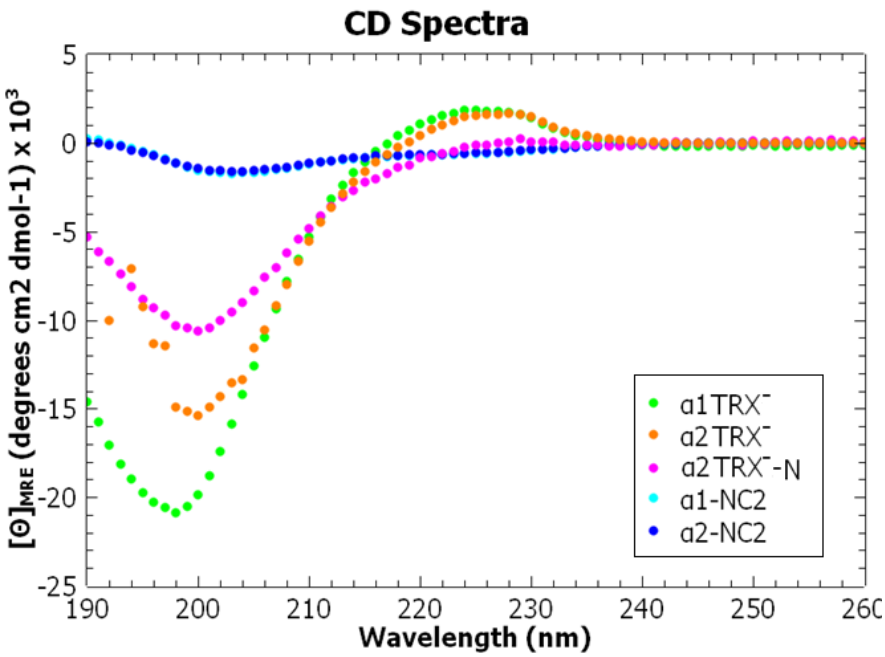
Figure 4. SDS-PAGE of column chromatography elution fractions. *Samples were run under reducing conditions on a 17% poly-acrylamide gel. Collected elution fractions after Ni-NTA affinity resin chromatography were run alongside an untransformed cell sample (negative control). The separate constructs shown in the figure are* ***(A)*** *α_1TRX$^-$* ***(B)*** *α_2TRX$^-$* ***(C)*** *α_2TRX$^-$-N* ***(D)*** *α_1-NC2 and* ***(E)*** *α_2-NC2. Monomers, dimers and trimers of each construct are highlighted in red, blue, orange respectively. Protein migration was confirmed by Western Blot for His-tag.*

The Folding and the Stability of the Designed Peptides

Protein aggregation was seen during dialysis against 5 mM Acetic Acid (pH 4.5) though not as aggressive as when dialyzed against Lysis Buffer (pH 7). As switching to neutral pH buffer resulted in a great loss of protein due to precipitation, experiments on all fusion proteins were performed at low temperatures and acidic pH to maximize yield. Circular dichroism spectra were taken at 5°C with samples in 5 mM Acetic Acid (pH 4.5). The CD spectra obtained for each construct all showcase a negative peak around 198 nm and positive peak around 225 nm (Figure 5a). The α_1-NC2 and α_2-NC2 homotrimers show almost identical wavelength scan profiles, with the lowest magnitude of positive and negative peaks. In comparison, the α_1TRX^-, α_2TRX^- and α_2TRX^--N have slight positive peaks and a strong negative peak. The negative peak of α_1TRX^- show the strongest magnitude, though it is still half of that reported by F877 (Xu et al., 2008). It is noted that the positive peaks of α_1-NC2, α_2-NC2 and α_2TRX^--N do not exceed 0 degrees cm^2 $dmol^{-1}$.

With the CD spectra of the pure NC2 trimer generously provided by Dr. Sergei Boudko's lab, the wavelength scan profile of the NC2 containing trimers sans the contribution of the alpha-helica; NC2 domain was obtained by equimolar subtraction (Figure 5b). The deconvoluted wavelength scan reveal what seems to be an inverted alpha-helix CD profile. This shows that the triple helical region has very little contribution to the wavelength profile of the NC2 containing constructs. The CD spectra of Foldon containing constructs were not deconvoluted as previous works of the Xu Lab show minimal contribution of the Foldon domain.

A



B

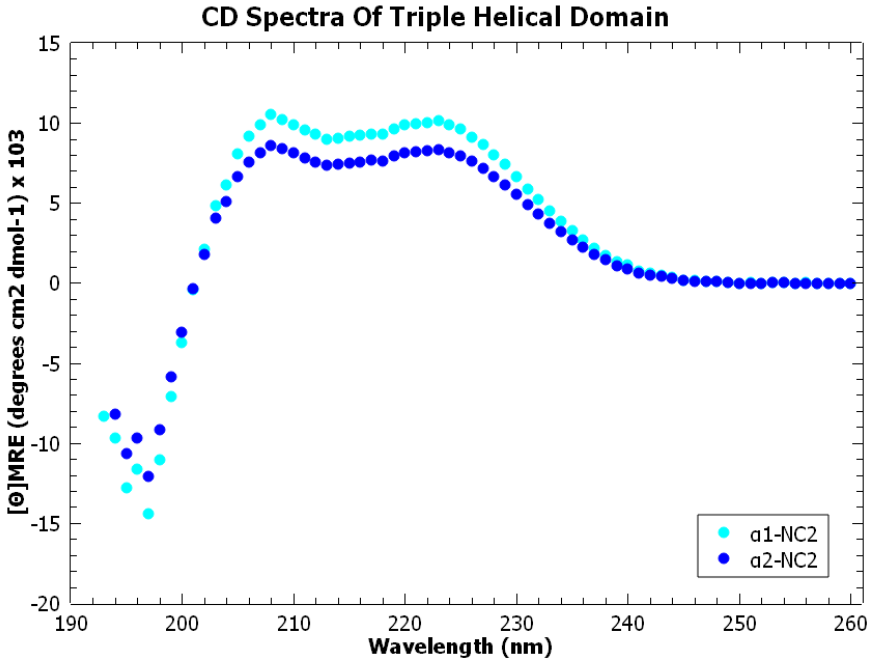


Figure 5. Circular Dichroism Spectra. ***(A)*** *Wavelength scans from 190-260 nm of α_1TRX^- (green), α_2TRX^- (orange), α_2TRX^--N (pink), α_1-NC2 (turquoise) and α_2-NC2 (blue), in 5 mM Acetic Acid (pH 4) at 5°C. Sample concentrations were 0.31 mg/mL, 1 mg/mL, 1 mg/mL, 0.91 mg/mL and 0.34 mg/mL respectively. CD data was normalized to mean residual ellipticity.* ***(B)*** *Wavelength scans of α_1-NC2 and α_2-NC2 without the NC2 domain. Equimolar subtraction of the pure NC2 CD spectra (provided by Sergei Boudko) to remove the contribution of the alpha-helical nucleation domain to reveal that of the triple helical region.*

With the CD spectras obtained, the Rpn (ratio of positive (225 nm) to negative (198 nm) peak) of the separate constructs were determined (Table 2). The Rpn of a pure triple helix is 0.12, providing a value to assess the triple-helical content of the constructs (Feng, 1996). However, this Rpn value is from natural collagen and collagen mimetic sequences with high hydroxyproline content. Of the constructs, α_1TRX^- and α_2TRX^- show comparable Rpn values (Rpn ~0.1) suggesting that they formed stable triple helices. The deviation from pure triple collagen may be due to presence of the His-tagged nucleation domain, indicating lower triple helical content, as well as the lack of hydroxyproline. In comparison, the α_2TRX^--N exhibit a extremely low Rpn value. The two NC2 containing peptides have a Rpn value of 0.87 and agrees with the deconvoluted CD spectra, the folding of the peptides is not comparable to that of a triple helix.

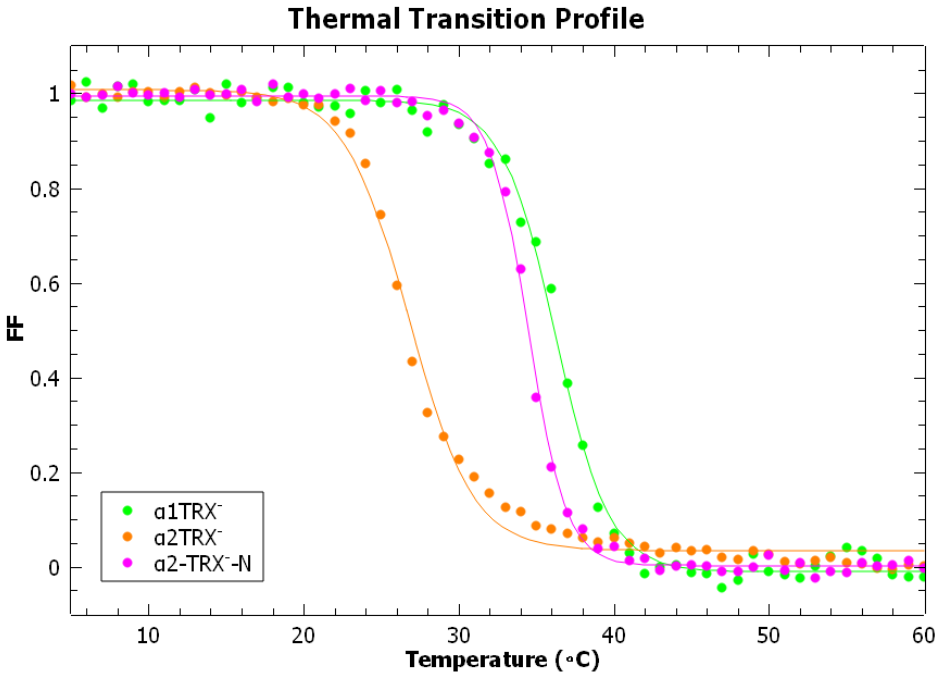
Construct	Rpn
α_1TRX^-	0.09
α_2TRX^-	0.10
α_2TRX^--N	0.02
α_1-NC2	0.87
α_2-NC2	0.87

Table 2. Rpn values of constructs. *Peak values were determined from the wavelength scans of each peptide. The Rpn value of NC2 containing domains used the corrected wavelength scan after equimolar subtraction of the NC2 spectra to see the triple helical contribution of the peptide.*

Temperature melt experiments were conducted on the constructs. As temperature increased, the negative peak and positive peak of the wavelength scans (190-260 nm) lowered in magnitude. The end temperature of the thermal melt experiment (60°C) is the reported melting temperature for the Foldon domain, indicating that the only thermal transition curve in the data is from the triple helical region. The change in CD signal is due to the transition of the triple helix

to that of a random coil. Fraction folded (FF) values were calculated using the circular dichroism data. The thermal transition profile of the constructs were obtained by plotting temperature against the fraction folded values (Figure 6). From the wavelength scan and Rpn values obtained, α_1TRX$^-$ and α_2TRX$^-$ had formed a triple-helical trimer (Rpn ~ 0.10). These two constructs exhibit ordered transition profiles (Figure 6a). Contrarily, the thermal transition profiles for the NC2 containing constructs support the notion that they were not properly folded (Figure 6b). The experimentally determined melting temperatures (T_m) of α_1TRX$^-$ and α_2TRX$^-$ are ~36°C and ~27°C respectively, both lower than the reported T_m of F877 (~38°C).

A



B

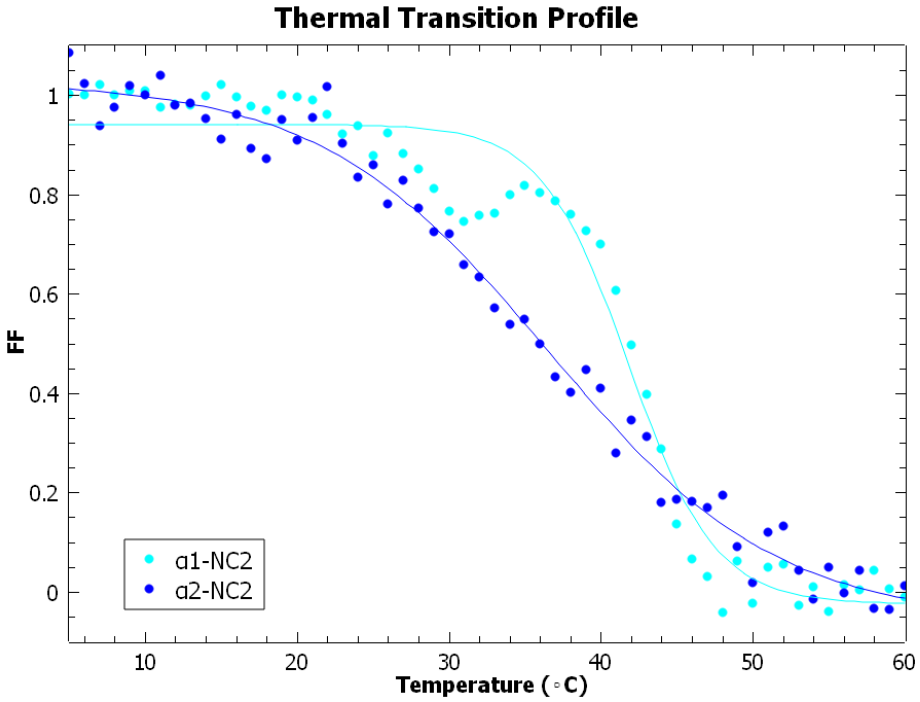


Figure 6. Unfolding Transition Profiles. *Temperature melt experiments were performed to assess the stability of the peptides. Samples were in 5 mM Acetic Acid (pH 4.5). The melting curve was produced monitoring the CD value at 198 nm (negative peak) from 5-60°C. (**A**) Thermal transition profile of α_1TRX^- (green), α_2TRX^- (orange) and α_2TRX^--N (pink) (**B**) Thermal transition profile of α_1-NC2 (turquoise) and α_2-NC2 (blue)*

Discussion

SDS-PAGE results showed that all constructs appeared within the predicted molecular weight ranges. The bands corresponding to the fusion proteins were not present in the negative control, confirming successful transformation. Even under reducing conditions (1x SDS), dimeric and trimeric forms were still present due to the GPCC sequence in the constructs. A disulfide bond is also formed between the NC2α_1 and NC2α_3 chains in the six NC2 containing proteins. Boudko et al. reported in 2010 that the NC2α_1 fragment has been shown to homotrimerize, making it difficult to pinpoint the identity of the dimeric and trimeric forms present in the α_1-NC2 and α_2-NC2 elution fractions. This is not an issue for the other constructs as the Foldon trimer consists of three identical subunits.

The circular dichroism wavelength scan profiles confirmed that α_1TRX$^-$ and α_2TRX$^-$ formed stable triple helix. We show that it is possible to form stable helices without the thioredoxin domain. This not only shortens the purification process but also avoids the steps where we've previously encountered the greatest loss in yield. We were also able to form a homotrimer derived from the α_2(I) chain comparable to its α_1(I) counterpart previously published from the Xu Lab. While the wavelength scan profile of the α_2 constructs do slightly differ from α_1 constructs, the reasoning is not fully understood. In the future, mutagenesis studies can be performed to see why α_2(I) homotrimers seem to fold less stably than α_1(I) homotrimers.

However, the lower Rpn value indicate that α_1TRX$^-$ and α_2TRX$^-$ have lower triple helical content than F877. This may be due to the presence of the His-tag, which was cleaved off with thioredoxin for the F877 fusion peptide. A 2014 study by Acevedo-Jake reported that the incorporation of amino acids rare to native collagen disrupts the triple helical formation. Though the study only reported for the incorporation of one histidine residue at the N-terminus, the

bundle of 12 negatively charged histidine residues in a trimer may adversely interrupt the triple helical region of the fusion protein.

The presence of the poly-histidine tag on the α_1-NC2 and α_2-NC2 constructs may also disrupt proper folding of the protein. As indicated by the deconvoluted CD spectra, the nucleation domain successfully folded while the triple helical region did not. It is also noted that the triple helical region in the six NC2 chains are much longer than those originally designed by Boudko and colleagues. It may be that the NC2 domain is not a sufficient trimerization domain for long collagenous domains. The presence of a C-terminal Cys-knot may also be an important stabilizing factor as it is found in both constructs that successfully folded. In the future, we can try incorporating this motif into the NC2 constructs.

Both α_1TRX^- and α_2TRX^- also show a shallower negative peak at 198 nm in comparison to F877. The loss of CD signal may indicate lower stability of the triple helical region for reasons as stated above. It is difficult to tell whether the signal indicates a loosely formed triple helix or small interruptions of the triple helix fold. However, while the negative peak ellipticity values are more positive than that of F877, they are comparable to previous reported collagen mimetic peptides (Col877, Col108) from the Xu lab (Chen et al., 2019; Kaur et al., 2014). The thermal stability experiments agree with the CD Spectra. Both α_1TRX^- and α_2TRX^- show lower melting temperature compared to F877, Col877 and Col108.

α_1TRX^- folding into stable triple helix conformation indicates that there is no need for the thioredoxin region in future fusion peptide designs. However, as the results do indicate a less stable triple helix, removal of the poly-histidine tag can be explored in future experiments. Perhaps, removal of the his-tag during the initial column chromatography step with enzyme

immobilized resin can be an efficient method. In this case, the original dialysis and thrombin cleavage step can be replaced by two column chromatography steps.

The success of α_2TRX^- forming a triple helical trimer confirms that the α_2(I) fragment can successfully trimerize despite the notion that α_2(I) homotrimers cannot form *in vitro*. In addition, expression of the α_2TRX^- can also be incorporated into heterotrimer studies given that α_1TRX^- is its α_1(I) counterpart. With the two successfully formed homotrimers, we can build a heterotrimer that mimics Type I Collagen. Designing constructs with two and more sequence units based on the two peptides can also allow us to study heterotrimeric fibrils.

References

Acevedo-Jake, Amanda M., et al. "Comparative NMR Analysis of Collagen Triple Helix Organization from N- to C-Termini." *Biomacromolecules*, vol. 16, no. 1, 10 Dec. 2014, pp. 145–155., doi:10.1021/bm501281a.

Boudko, Sergei P et al. "The crucial role of trimerization domains in collagen folding." *The international journal of biochemistry & cell biology* vol. 44,1 (2012): 21-32. doi:10.1016/j.biocel.2011.09.009

Boudko, Sergei P., et al. "The NC2 Domain of Collagen IX Provides Chain Selection and Heterotrimerization." *Journal of Biological Chemistry*, vol. 285, no. 31, 30 July 2010, pp. 23721–23731., doi:10.1074/jbc.m110.128405.

Boudko, Sergei P., and Hans Peter Bächinger. "Structural Insight for Chain Selection and Stagger Control in Collagen." *Scientific Reports*, vol. 6, no. 1, 29 Nov. 2016, doi:10.1038/srep37831.

Brodsky, Barbara, and John A. Ramshaw. "Bioengineered Collagens." *Subcellular Biochemistry*, 2017, pp. 601–629., doi:10.1007/978-3-319-49674-0_18.

Boudko, Sergei P., and Hans Peter Bächinger. "The NC2 Domain of Type IX Collagen Determines the Chain Register of the Triple Helix." *Journal of Biological Chemistry*, vol. 287, no. 53, 28 Dec. 2012, pp. 44536–44545., doi:10.1074/jbc.m112.417543.

Brodsky, Barbara, and John A.M. Ramshaw. "The Collagen Triple-Helix Structure." *Matrix Biology*, vol. 15, no. 8-9, 1997, pp. 545–554., doi:10.1016/s0945-053x(97)90030-5.

Bulleid, Neil J. "Novel Approach to Study the Initial Events in the Folding and Assembly of Procollagen." *Seminars in Cell & Developmental Biology*, vol. 7, no. 5, 1996, pp. 667–672., doi:10.1006/scdb.1996.0081.

Canty, Elizabeth G, and Karl E Kadler. "Procollagen trafficking, processing and fibrillogenesis." *Journal of cell science* vol. 118,Pt 7 (2005): 1341-53. doi:10.1242/jcs.01731

Chang, Shu-Wei, et al. "Structural and Mechanical Differences between Collagen Homo- and Heterotrimers: Relevance for the Molecular Origin of Brittle Bone Disease." *Biophysical Journal*, vol. 102, no. 3, 2012, pp. 640–648., doi:10.1016/j.bpj.2011.11.3999.

Chen, Fangfang, et al. "The Predominant Roles of the Sequence Periodicity in the Self-Assembly of Collagen-Mimetic Mini-Fibrils." *Protein Science*, vol. 28, no. 9, 29 July 2019, pp. 1640–1651., doi:10.1002/pro.3679.

Doyle, Sharon A., and Barbara D. Smith. "Role of the pro-α2(i) COOH-Terminal Region in Assembly of Type I Collagen: Disruption of Two Intramolecular Disulfide Bonds in pro-

α2(i) Blocks Assembly of Type I Collagen." *Journal of Cellular Biochemistry*, vol. 71, no. 2, 1 Nov. 1998, pp. 233–242., doi:10.1002/(sici)1097-4644(19981101)71:2<233::aid-jcb8>3.0.co;2-w.

Engel, Jürgen, and Hans Peter Bächinger. "Structure, Stability and Folding of the Collagen *Triple Helix." Topics in Current Chemistry*, 2005, pp. 7–33., doi:10.1007/b103818.

Feng, Yangbo, et al. "Collagen-Based Structures Containing the Peptoid Residue n-Isobutylglycine (Nleu): Synthesis and Biophysical Studies of Gly-Pro-Nleu Sequences by Circular Dichroism, Ultraviolet Absorbance, and Optical Rotation." *Biopolymers*, vol. 39, no. 6, 6 Dec. 1998, pp. 859–872., doi:10.1002/(sici)1097-0282(199612)39:6<859::aid-bip10>3.0.co;2-z.

Frank, Sabine, et al. "Stabilization of Short Collagen-like Triple Helices by Protein Engineering." *Journal of Molecular Biology*, vol. 308, no. 5, 18 May 2001, pp. 1081–1089., doi:10.1006/jmbi.2001.4644.

Han, Sejin, et al. "Molecular Mechanism of Type I Collagen Homotrimer Resistance to Mammalian Collagenases." *Journal of Biological Chemistry*, vol. 285, no. 29, 16 July 2010, pp. 22276–22281., doi:10.1074/jbc.m110.102079.

Hulmes, D J et al. "Analysis of the primary structure of collagen for the origins of molecular packing." *Journal of molecular biology* vol. 79,1 (1973): 137-48. doi:10.1016/0022-2836(73)90275-1

Kadler, Karl E et al. "Collagens at a glance." *Journal of cell science* vol. 120,Pt 12 (2007): 1955-8. doi:10.1242/jcs.03453

Kaur, Parminder Jeet, et al. "The Self-Assembly of a Mini-Fibril with Axial Periodicity from a Designed Collagen-Mimetic Triple Helix." *Journal of Biological Chemistry*, vol. 290, no. 14, 11 Feb. 2015, pp. 9251–9261., doi:10.1074/jbc.m113.542241.

LaVallie, Edward R., et al. "A Thioredoxin Gene Fusion Expression System That Circumvents Inclusion Body Formation in the E. Coli Cytoplasm." *Nature Biotechnology*, vol. 11, no. 2, 1 Feb. 1993, pp. 187–193., doi:10.1038/nbt0293-187.

Leikina, E., et al. "Type I Collagen Is Thermally Unstable at Body Temperature." *Proceedings of the National Academy of Sciences*, vol. 99, no. 3, 5 Feb. 2002, pp. 1314–1318., doi:10.1073/pnas.032307099.

Meier, Sebastian, et al. "Foldon, the Natural Trimerization Domain of T4 Fibritin, Dissociates into a Monomeric A-State Form Containing a Stable β-Hairpin: Atomic Details of Trimer Dissociation and Local β-Hairpin Stability from Residual Dipolar Couplings." *Journal of Molecular Biology*, vol. 344, no. 4, 2004, pp. 1051–1069., doi:10.1016/j.jmb.2004.09.079.

Mekkat, Arya, et al. “Effects of Flexibility of the α2 Chain of Type I Collagen on Collagenase Cleavage.” *Journal of Structural Biology*, vol. 203, no. 3, 12 May 2018, pp. 247–254., doi:10.1016/j.jsb.2018.05.002.

Nerenberg, Paul S., and Collin M. Stultz. “Differential Unfolding of α1 and α2 Chains in Type I Collagen and Collagenolysis.” *Journal of Molecular Biology*, vol. 382, no. 1, 11 July 2008, pp. 246–256., doi:10.1016/j.jmb.2008.07.009.

Orgel, J. P., et al. “Microfibrillar Structure of Type I Collagen in Situ.” *Proceedings of the National Academy of Sciences*, vol. 103, no. 24, 2006, pp. 9001–9005., doi:10.1073/pnas.0502718103.

Petruska, J A, And A J Hodge. “A Subunit Model For The Tropocollagen Macromolecule.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 51,5 (1964): 871-6. doi:10.1073/pnas.51.5.871

Piez, Karl A., and Andrew Miller. “The Structure of Collagen Fibrils.” *Journal of Supramolecular Structure*, vol. 2, no. 2-4, 1974, pp. 121–137., doi:10.1002/jss.400020207.

Sharma, Urvashi, et al. “Structural Basis of Homo- and Heterotrimerization of Collagen I.” *Nature Communications*, vol. 8, no. 1, 10 Mar. 2017, doi:10.1038/ncomms14671.

Strawn, Rebecca, et al. “To Achieve Self-Assembled Collagen Mimetic Fibrils Using Designed Peptides.” *Biopolymers*, vol. 109, no. 7, 14 Apr. 2018, doi:10.1002/bip.23226.

Xu, Ke, et al. “Recombinant Collagen Studies Link the Severe Conformational Changes Induced by Osteogenesis Imperfecta Mutations to the Disruption of a Set of Interchain Salt Bridges.” *Journal of Biological Chemistry*, vol. 283, no. 49, 8 Oct. 2008, pp. 34337–34344., doi:10.1074/jbc.m805485200.

Xu, Yujia, and Michele Kirchner. “Collagen Mimetic Peptides.” *Bioengineering*, vol. 8, no. 1, 5 Jan. 2021, p. 5., doi:10.3390/bioengineering8010005.

Yasukawa, Takashi, et al. “Increase of Solubility of Foreign Proteins in Escherichia Coli by Coproduction of the Bacterial Thioredoxin.” *Journal of Biological Chemistry*, vol. 270, no. 43, 27 Oct. 1995, pp. 25328–25331., doi:10.1074/jbc.270.43.25328.