Alpha-helical coiled-coil oligomerization domains are almost ubiquitous in the collagen superfamily*

Audrey McAlinden¹, Thomasin A. Smith², Linda J. Sandell¹, Damien Ficheux³, David A.D. Parry² and David J.S. Hulmes³

¹Department of Orthopedic Surgery, Washington University School of Medicine, Barnes-Jewish Hospital, St Louis, MI 63110, USA; ²Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand; ³Institut de Biologie et Chimie des Protéines, CNRS UMR 5086, Université Claude Bernard Lyon 1, 69367 Lyon cedex 7, France.

Corresponding author :
Dr D.J.S. Hulmes, IBCP, 7 passage du Vercors, 69367 Lyon cedex 07, France.
Tel : (33) (0)4 72 72 26 67; Fax : (33) (0)4 72 72 26 04; e-mail : d.hulmes@ibcp.fr

Running Title : Coiled-coil domains in the collagen superfamily

*This work was supported by the CNRS (Programme “Protéomique et Ingeniérie des Protéines”) and National Institutes of Health Grants AR48250-01 (to A.M.) and AR-36994 (to L.S.).
Summary

Alpha-helical coiled-coils are widely occurring protein oligomerization motifs. Here we show that most members of the collagen superfamily contain short, repeating heptad sequences typical of coiled coils. Such sequences are found at the N-terminal ends of the C-propeptide domains in all fibrillar procollagens. When fused C-terminal to a reporter molecule containing a collagen-like sequence that does not spontaneously trimerize, the C-propeptide heptad repeats induced trimerization. C-terminal heptad repeats were also found in the oligomerization domains of the multiplexins (collagens XV and XVIII). N-terminal heptad repeats are known to drive trimerization in transmembrane collagens, while fibril associated collagens with interrupted triple helices (FACITs), as well as collagens VII, XIII, XXIII and XXV, were found to contain heptad repeats between collagen domains. Finally, heptad repeats were found in the von Willebrand factor A domains known to be involved in trimerization of collagen VI, as well as in collagen VII. These observations suggest that coiled-coil oligomerization domains are widely used in the assembly of collagens and collagen-like proteins.
Introduction

The mechanisms controlling chain oligomerization in extracellular matrix and related proteins are currently topics of active research (1, 2). In the case of the collagen superfamily, which includes both collagens and collagen-like proteins (3-5), a number of structural features have been identified that are essential for bringing together component polypeptide chains with the correct stoichiometry, thus leading to folding of the triple-helix. Early studies on the procollagen precursors of the fibrillar collagens (types I, II, III, V and XI) (6) showed that the C-propeptide regions are necessary to direct correct chain association, which is followed by zipper-like folding of the triple-helix in the C- to N-terminal direction (7). This concept was subsequently extended to basement membrane collagen type IV (8-10), microfibril forming collagen VI (11, 12), members of the C1q family (13) including collagens VIII and X (14-17) and the emilins (18), and the FACITs (19-21).

Alpha-helical coiled coils have been shown to be oligomerization domains in both collagens and collagen-like proteins. The paradigm here is the collectin family (22), which includes the lung surfactant proteins (SP-A, SP-D), mannan binding proteins (MBP-A, MBP-C), conglutinin and collectin-43. Collectins are trimeric molecules in which each chain contains a collagen-like domain followed by an α-helical coiled-coil region and then a carbohydrate recognition domain (CRD). The amino acid sequence of the coiled-coil domain is characterized by up to four heptad repeats (a-b-c-d-e-f-g) in which hydrophobic amino acid residues occur at positions $a$ and $d$ (23). Three dimensional structures of the coiled-coil and CRD regions in MBP and SP-D show that chains within the trimer are associated via a three-stranded coiled coil, with the three CRDs arranged as distinct lobes (24-26). Several studies have shown that the coiled-coil region is both necessary and sufficient for trimerization of SP-D (27, 28). In a recent study (29), the first two heptad repeats of SP-D were shown to be

---

1 The abbreviations used are: CHO, Chinese hamster ovary; CPII, C-propeptide region of procollagen II; CRD, carbohydrate recognition domain; FACITs, fibril associated collagens with interrupted triple-helices; SP-D,
sufficient to drive trimerization and folding when fused C-terminal to the N-propeptide region of procollagen IIA, which also contains a collagen triple-helix but which alone does not spontaneously trimerize.

Coiled-coil oligomerization domains have also been shown to be essential for trimerization and folding of the membrane associated collagens XIII (30, 31) and XVII (32). Similar domains occur in collagen XXIII (33), the Alzheimer amyloid plaque component precursor CLAC-P/collagen XXV (34) and the collagen-like membrane proteins MARCO and ectodysplasin-A1 (31). In all of these proteins, the transmembrane domain is found close to the N-terminus and is followed immediately by a heptad repeat region at the beginning of the large ectodomain. Once trimerized via the coiled-coil domains, folding of the C-terminal collagen helix then proceeds in the N- to C-terminal direction. N-terminal coiled-coil oligomerization domains are thought to play similar roles in the assembly of macrophage scavenger receptor proteins (35), which also contain collagen-like regions. Several coiled-coil domains, as well as collagen-like regions, are also present in the emilins (36).

Here we show that putative α-helical coiled-coil oligomerization domains are present in most members of the collagen superfamily, located either before, after or between collagen-like regions, suggesting a general role in triple-helical assembly.

Experimental Procedures

*Sequence analysis* - All sequences were obtained from SWISSPROT, SPTREMBL or REMTREMBL databases. Potential coiled-coil sequences in all polypeptide chains of the collagen superfamily were searched initially using the program COILS2 (37). Since the reliability of this computational approach is very limited, particularly for short (14 residue) sequences, all "hits" (regardless of score) were also examined by informed eye in order to

*Surfactant protein-D.*
assess the likelihood of coiled coil formation. Experience has shown that this approach can more easily recognize heptad discontinuities and individual or small groups of residues that are generally disruptive to coiled-coil formation, hence allowing elimination of some regions from further consideration.

Although sequences were analyzed for their potential to form either two or three-stranded helices using SCORER (38), their very short lengths (2-3 heptads in general) rendered the results statistically uncertain. The presence of three-chain structures in the proteins reported here would nonetheless strongly favor the formation of three-stranded rather than two-stranded coiled-coils.

To test the plausibility of a three-stranded coiled-coil geometry for the heptad containing region of type I procollagen C-propeptides, a model three-stranded coiled-coil was built based on the homotrimeric GCN4-pII leucine zipper structure (PDB code 1gcm). Appropriate residue replacements were made, placing the pro\(\alpha\)1 chains onto chains A and B and the pro\(\alpha\)2 chain onto chain C of GCN4-pII, in such a way as to ensure the heptad repeat patterns were in phase. The packing of the core residues was then checked manually using Turbo-Frodo 5.5 (http://afmb.cnrs-mrs.fr/TURBO_FRODO), noting particularly the positions of the \(a\) and \(d\) residues relative to those in the GCN4-pII structure. Automated energy minimization was done in a multi-step process using conjugate gradients followed by simulated annealing with CNSsolve 0.5 (39). In the conjugate gradient minimization, the main-chain atoms were first held fixed followed by harmonically restraining them to their initial positions with an energy constant of 10 kcal/mol.A\(^2\). Annealing using torsion dynamics was employed with the main-chain atoms again harmonically restrained.

*Synthesis of chimeric IIA N-pro/CPII cDNA fusion constructs -* cDNA fusion constructs encoding the full-length human type IIA procollagen N-propeptide linked to 18, 25, or 31 amino acids of the coiled-coil domain present at the beginning of the type II procollagen C-
propeptide (IIA N-pro/CP II; Fig. 3) were synthesized by overlap extension PCR. Oligonucleotide primers used for PCR and overlap extension PCR procedures were synthesized by Invitrogen. One PCR was carried out to amplify type IIA N-propeptide cDNA using primers 1 and 2 (Table II) for 30 cycles (95°C, 30s; 55°C, 30s; 72°C, 50s). Here, IIA N-pro/SP-D cDNA in pGEM3Z was used as a substrate (28). Another set of PCRs was done to amplify cDNA encoding the 18, 25 or 31 amino acid sequence of the C-propeptide (plus the terminal alanine residue of the telopeptide at the procollagen C-proteinase cleavage site) corresponding to amino acid numbers 1172-1189, 1172-1196 or 1172-1202, respectively (Gene Bank™ Accession number X16468). Table II shows the sequence of the PCR primer pairs used to amplify the cDNA fragments encoding each of the three C-propeptide regions. Each of these PCRs was carried out for 30 cycles (95°C, 30s; 55°C, 30s; 72°C, 25s) and oligonucleotides encoding amino acid residues 1172-1189, 1172-1196 or 1172-1202 of the C-propeptide were used as substrates. Overlap extension PCRs were then prepared by mixing cDNA encoding the IIA N-propeptide with cDNA encoding one of the three C-propeptide cDNA fragments. To promote overlap extension, a preliminary PCR was carried out for 5 cycles in the absence of primers (95°C, 30s; 52°C, 30s; 72°C, 45s). Primer pairs were then added to each PCR (Table II) and IIA N-pro/CP II cDNA fusion constructs were amplified for 30 cycles (95°C, 30s; 55°C, 30s; 72°C, 55s). cDNA fusion constructs were purified from 1% agarose gels, digested with EcoRI and subcloned into pcDNA3 (Invitrogen). DNA sequencing using the T7 and sp6 promoter primers confirmed correct orientation of the inserted cDNA fragments.

**Transient transfections** - Chinese Hamster Ovary (CHO) cells were plated overnight in 6-well tissue culture dishes at a density of $3 \times 10^5$ cells/ml. Cells were transiently transfected with pcDNA3 alone or pcDNA3 containing each of the three cDNA fusion constructs (IIA N-pro/CP II aa 1172-1189/1196/1202) using FuGENE 6 reagent (Roche Molecular
Biochemicals), in the presence of 50 µg/ml ascorbate. Cells transfected to synthesize full-length IIA N-pro/SP-D fusion protein (28) or just the type IIA N-propeptide were included as controls: IIA N-pro/SP-D forms trimers in solution while IIA N-pro alone remains monomeric. Conditioned medium (12-36 ml per fusion construct) was harvested after 2 days in culture and clarified by centrifugation. PMSF (1mM) was added to the medium and proteins were precipitated overnight at 4°C with 33% ammonium sulfate. Precipitated proteins were collected by centrifugation and washed three times in saturated ammonium sulfate. Protein pellets were re-suspended in PBS (1-3 ml) and dialyzed overnight against PBS at 4°C.

Chemical crosslinking and detection of chimeric fusion proteins - Covalent crosslinking of fusion proteins was done using bis-(sulfo succinimidyl)suberate (BS3; Pierce). Increasing amounts of BS3 (0, 0.1, 0.5, 1 or 2mM final concentration) prepared in 5mM sodium citrate, pH 5, were added to dialyzed conditioned medium for 1h at room temperature. Addition of SDS-PAGE loading buffer containing Tris-HCl (0.5M) inhibited the reaction. Samples were boiled for 5 min prior to SDS-PAGE which was carried out without sulfydryl reduction. Proteins were transferred to nitrocellulose membranes by Western blotting and the presence of cross-linked dimers or trimers was identified by immunolocalization using anti-IIA polyclonal antisera (40).

Peptide analysis - Peptides corresponding to GCN4-pII (RMKQIEDKIEEILSKYHILENEIARIKKKLGER-NH2) and residues 1172-1189 (ADQAAGGLRQHDAEVDAT-NH2), 1172-1196 (ADQAAGGLRQHDAEVDATLKLSSNQ-NH2) and 1172-1202 (ADQAAGGLRQHDAEVDATLKLSSNQIESIRS-NH2) from human procollagen II were synthesized on a Milligan 9050 instrument with Fmoc/DIC/HOBt chemistry. They were then cleaved using TFA with classical scavengers and precipitated in diethyl ether, followed by centrifugation. Pellets were then dissolved in water and lyophilized. The crude peptides were then re-
dissolved in water and purified on a Vydac column (C18, 5 µm, 25 x 1 cm) with an appropriate gradient of acetonitrile in 0.1 % TFA. They were then characterized by electrospray mass spectrometry (SCIEX API 165) and by HPLC (HP1100) using an analytical Vydac C18 column with a 30 min gradient from 7 % to 63 % acetonitrile.

Far UV (180 - 260 nm) circular dichroism measurements were carried out using thermostated 0.2 mm path length quartz cells in a Jobin-Yvon CD6 instrument, calibrated with aqueous d-10-camphorsulfonic acid. Peptides (200 µM - 400 µM) were analyzed at 25 °C in 20 mM KH₂PO₄/NaOH, 150 mM NaF, pH 7.2, in the presence or absence of 50 % TFE (41). Spectra were measured with wavelength increment 0.2 nm, integration time 1 sec and bandpass 2 nm. For cross-linking analysis with BS₃, conditions were the same as those used for the fusion proteins (see above), except that peptide concentrations were as for the circular dichroism.

Results

Coiled coils in fibrillar procollagen C-propeptides - Fig. 1A shows an amino acid sequence alignment of the C-propeptide domains of the human fibrillar procollagens. From the pattern of repeating hydrophobic residues at positions a and d, and the frequent occurrence of charged residues at positions e and g, it can be seen that there are up to four heptad repeats, indicative of coiled-coils, beginning soon after the procollagen C-proteinase cleavage site. In all chains, the g position at the end of the fourth heptad repeat is a proline which therefore defines the end of the α-helical region. The extent of the heptad repeat pattern is similar to that found in the collectins (Fig. 1B), again C-terminal to the collagenous region. This suggests the existence of a coiled-coil structure near the N-terminus of all fibrillar C-propeptide trimers. Putative coiled-coil regions were also found C-terminal to the triple helix in the fibrillar procollagens XXIV (42) and XXVII (43, 44) (Fig. 1A).
When the C-propeptide heptad repeats were further analyzed using SCORER (38), a preference for three-stranded, as opposed to two-stranded, coiled coils was evident (data not shown). Modeling of the three-dimensional structure of the type I procollagen C-propeptide di-heptad sequence spanning the \(d\) positions of heptads two and four (Fig. 1A), based on that adopted by the mutated form of the GCN4 leucine zipper (GCN4-pII, in which residues at the \(a\) and \(d\) positions are isoleucines) (45) was consistent with a three-stranded coiled-coil conformation free of steric clashes (Fig. 2). Possible ionic interactions stabilizing the structure are listed in Table I. Based on this analysis, and in view of the recent observation that a di-heptad repeat in SP-D is sufficient for trimerization (29), this suggests that the heptad repeats in the fibrillar procollagen C-propeptides could also act as oligomerization domains.

**Coiled coils in procollagens as oligomerization domains** - In order to test whether the heptad repeats in the procollagen C-propeptides can act as oligomerization domains, we adopted the approach described by McAlinden et al (29) in which putative trimerization domains were fused C-terminal to the type IIA procollagen N-propeptide and expressed in CHO cells. The IIA N-propeptide contains a short, interrupted \((GXY)_n\) region (25 triplets) which alone does not spontaneously trimerize. When fused to at least the first two heptad repeats of the neck region of surfactant protein-D (SP-D), however, trimerization ensues, as revealed by electrophoretic analysis after stabilization of trimers by cross-linking. Using the same IIA N-propeptide as a trimerization reporter molecule, we made fusion constructs using sequences from the C-propeptide of human type II procollagen, beginning at the alanine residue in the procollagen C-proteinase cleavage site, up to the end of the second, third or fourth heptad repeats (Fig. 3). As shown in Fig. 4A, positive controls with the IIA N-propeptide fused to the entire C-terminal region of SP-D (beginning at the first heptad repeat) readily formed dimers and trimers, as expected (29), while in the absence of an oligomerization domain, only monomers were observed (Fig. 4B). In contrast, when the IIA
N-propeptide was fused to the N-terminal sequence of the type II procollagen C-propeptide, up to position f in the fourth heptad repeat, cross-linkable dimers and trimers were again observed (Fig. 4C). Similar results were obtained with fusion constructs containing only the first three (Fig. 4D) or two (Fig. 4E) heptad repeats.

The data obtained with the fusion constructs show that sequences from the N-terminal region of the procollagen II C-propeptide can trimerize when linked to the IIA N-propeptide. To determine whether these sequences are capable of trimerizing independently, the corresponding peptides were synthesized along with GCN4-pII as a positive control. When analyzed by circular dichroism (Fig. 5A), GCN4-pII gave a spectrum characteristic of an almost fully α-helical coiled coil, as shown by the positions and heights of the peaks at 192, 208 and 222 nm, with a ratio of mean residue ellipticities $\theta_{220}/\theta_{208}$ of 1.039. In the presence of 50 % TFE (Fig. 5B), which disrupts coiled coils and stabilizes α-helices (46), this ratio decreased to 0.84, due mainly to changes in the peak at 208 nm, as expected. In contrast, all three peptides from the N-terminal region of the procollagen II C-propeptide yielded spectra characteristic of random coils, in the absence of TFE. Therefore these peptides do not form stable coiled coils independently. In the presence of 50 % TFE, spectra clearly characteristic of α-helices were obtained with the two longer peptides (residues 1172-1196 and 1172-1202), while the shorter peptide (residues 1172-1189) showed some signs of α-helical structure. Therefore all three peptides are capable of adopting an α-helical fold.

The results obtained by circular dichroism were confirmed by cross-linking analysis using BS³, followed by SDS-PAGE, which showed trimerization of GCN4-pII but not of a similar length peptide (residues 1172-1202) from the human procollagen II C-propeptide (data not shown).

_C-terminal coiled coils in other members of the collagen superfamily_ - Alpha-helical coiled-coil heptad repeats were also found in the C-propeptides of known invertebrate fibrillar
procollagens, including the sponge *Ephydatia muelleri*, hydra, annelids (*alvinella, arenicola, riftia*), mollusc (abalone) and sea urchin (Fig. 1C) (47, 48). Both the length and distance of the heptad repeats from the end of the collagen triple-helical region were similar to those in the vertebrate fibrillar collagens (Fig. 1A), though in hydra and annelid fibrillar procollagens the intervening sequence was relatively long (Fig. 1C). Similar C-terminally located heptad sequences were also found in some bacteriophage tail fiber collagen-like proteins (data not shown).

As pointed out by Beck and Brodsky (49), a possible coiled-coil oligomerization domain C-terminal to the triple helix is also found in the collagen-like tail subunit of acetylcholinesterase (ColQ) (50). The sequence (Fig. 1D) consists of 2.5 heptad repeats and is found about 30 residues C-terminal to the last GPP triplet, similar to that in the procollagen C-propeptides. Probable trimerization domains have also been identified in the multiplexins (collagens XV and XVIII) (51, 52), immediately after the most C-terminal triple-helical domain. Examination of the corresponding sequences showed that these also contain heptad repeats indicative of coiled coils, albeit with a relative paucity of charged residues in positions e and g (Fig. 1D). Finally, sequence analysis of the recently described collagen XXVI (53) also revealed a putative coiled-coil region (31) in the C-terminal non triple-helical domain (Fig. 1D). This consists of 3 perfect heptad repeats and is located a similar distance from the end of the collagen triple-helical regions to those in ColQ and collagens XV and XVIII.

*N-terminal coiled coils* - Coiled-coil motifs N-terminal to collagen sequences have been shown to be necessary for trimerization of the transmembrane collagens XIII (30, 31) and XVII (32). These are shown in Fig. 1E, along with homologous sequences in collagen XXIII (33), CLAC-P/collagen XXV (34), MARCO and ectodysplasin-A. Following the suggestion of Balding et al (54), the heptad repeat is shown extending into the transmembrane region.
Coiled coils N-terminal to collagen triple-helices have also been described in the scavenger receptors (35, 55) and the emilins (36). To add to this sub-group, sequence analysis revealed a perfect 3 heptad substructure (31) about 45 residues N-terminal to the first triple-helical domain in the recently described collagen XXVI (Fig. 1F) (53). Finally, N-terminally located heptad repeats were found in *c. elegans* cuticle collagens as well as streptococcus and bacteriophage tail fiber collagen-like proteins (data not shown).

**Inter triple-helical domain coiled coils** - Putative coiled-coil oligomerization domains were also found in the fibril associated collagens with interrupted triple-helices (FACITs). Unlike the previous examples, however, these were found between (Gly-X-Y)$_n$ regions, in the NC2 domain, between triple-helical domains COL1 and COL2 (Fig. 1G). A further distinguishing feature is that the coiled-coil regions in FACITs are discontinuous, consisting of two partially overlapping di-heptad repeats.

Coiled-coil sequences separating triple-helical regions were also found in collagens VII, XIII, XXIII and XXV (Fig. 1H,I). The NC3 domains of collagens XIII and XXV, which are relatively highly conserved, consist of 2.5 heptad repeats, while that of collagen XXIII is much shorter (31). In collagen VII, the non triple-helical interruption in the long collagen domain contains 3 heptad repeats.

**Intra-domain coiled coils** - Finally, putative coiled-coil motifs were also found within von Willebrand factor A (VWA) domains. These occur N-terminal to the triple-helical region in collagen VII, and C-terminal to the triple-helical regions in all three chains of collagen VI, in the latter in domains known be required for trimerization (11, 12). The corresponding sequence alignment is shown in Fig. 1J. Secondary structure prediction on the NPS@ server (56); www.ibcp.fr) indicated an $\alpha$-helical conformation for all these sequences. Since all known three-dimensional structures of VWA domains (57) conform to a Rossman type fold with internal $\beta$-sheets and exposed $\alpha$-helices, we predict that these putative coiled-coil
sequences in collagens VI and VII would also be surface located. In this way, three adjacent VWA domains might trimerize via coiled-coil interactions.

**Discussion**

Here we show that most members of the collagen superfamily, including collagen triple-helix containing proteins not normally classified as collagens, contain short, 2-4 heptad repeat \(\alpha\)-helical coiled-coil domains. These include all fibrillar procollagens (types I, II, III, V, XI, XXIV and XXVII), all FACITs (collagens IX, XII, XIV, XIX, XX and XXIII), all transmembrane collagens (types XIII, XVII, XXIII and XXV, MARCO, ectodysplasin, scavenger receptors), the collectins, the multiplexins (collagens XV and XVIII), the emilins, the collagen-like tail subunit of acetylcholinesterase, and collagens VI, VII and XXVI. The only exceptions were basement membrane collagen IV, collagen XXII, the ficolins (58) and non-emilin members of the C1q family. The widespread occurrence of these domains suggests a general role in the assembly of collagen triple helices.

Cooperativity between collagen-like and coiled-coil domains - The observation that the heptad repeat containing peptides from the N-terminal region of the procollagen II C-propeptide do not form \(\alpha\)-helical coiled coils independently shows that trimerization is a cooperative process involving both coiled-coil and collagen-like regions. The absence of coiled-coil formation with the isolated peptides is not surprising, as it is generally agreed that the minimum length for the formation of a coiled-coil as an autonomously folding unit is three heptad repeats or 21-23 residues (46). Even the longest peptide examined (residues 1172-1202) contains only two perfect heptad repeats, given the presence of the aspartate residue at position \(a\) in heptad repeat 2 (Fig. 1A). In addition to the collagen-like region, cooperativity probably also exists between folding of the coiled-coil region and of the rest of the C-propeptide (see below).
Implications for the structures of procollagen C-propeptide regions - These observations give insights into the three-dimensional structure of the procollagen C-propeptides. Recent studies using low angle X-ray scattering have shown that the low-resolution structure of the isolated C-propeptide trimer from type III procollagen consists of three major lobes together with one minor lobe (59). We hypothesized that each major lobe corresponds to the loop containing the intra-chain disulfide bonds in each of the three component chains, while the minor lobe corresponds to a putative junction region. Here we show that the junction region most likely begins with a three-stranded coiled coil. The three-dimensional structure of the C-propeptide trimer therefore shows some similarities to the C-terminal regions of the collectins (24-26), where the CRD domains of each of the three chains also appear as independent lobes, connected by a coiled-coil junction region.

Coiled-coils and procollagen assembly - What is the role of the coiled-coil region in fibrillar procollagen chain trimerization? The C-propeptides of all the fibrillar procollagen chains (each ~250 residues in length) are highly conserved throughout most of their length (4, 60). Early observations on naturally occurring mutants of type I procollagen have shown the importance of the C-propeptides in trimerization (61). Frameshift mutations resulting in deletion and replacement of sequences C-terminal to the coiled-coil regions in both the proα1(I) and proα2(I) chains either prevent or delay trimerization (62-64). More recently, by site directed mutagenesis, it has been shown (65) that the presence of the last ten amino acids in the C-propeptide of the proα2(I) chain is essential for formation of the [proα1(I)]₂proα2(I) heterotrimer. These results and the experiments of Bulleid and colleagues which resulted in the identification of the “molecular recognition sequence” (66) approximately half way through the C-propeptide sequence suggest that the coiled-coil region is not sufficient for trimerization.
In contrast, the results reported here, using the procollagen IIA N-propeptide fused to the beginning of the procollagen II C-propeptide domain, show that the heptad repeat containing region of the latter is indeed sufficient to induce trimerization. Previous results with IIA N-pro fused to similar heptad repeats in SP-D have shown that these are also sufficient to induce correct folding of the IIA N-pro triple helix (29). These results are analogous to those of Bulleid et al (67) who showed that procollagen III molecules continue to trimerize when the C-propeptides are replaced by the 30-residue transmembrane domain of influenza virus haemagglutinin. We therefore suggest that the C-propeptide coiled-coil region alone would be sufficient to drive trimerization of fibrillar procollagen molecules lacking the rest of the C-propeptide. When followed by the remainder of the C-propeptide sequence, however, the coiled coil might be de-stabilized by unfavorable interactions associated with incorrect chain stoichiometry or the presence of mutations. Only when chains assemble with the correct stoichiometry, and in the absence of unfavorable mutations, would formation of the coiled-coil region be possible, leading to nucleation and folding of the adjacent collagen-triple helix. We predict that mutations within the coiled-coil region would prevent trimerization, especially if these were to occur at the a and d, or possibly e and g, positions. No such mutations have been described.

**Heterotrimerization versus homotrimerization of fibrillar collagen α2 chains** - Detailed examination of the beginning of the heptad repeat region of the procollagen C-propeptides reveals a possible explanation for the known tendency of α2 chains (in types I, V and XII collagens) not to form heterotrimers. As shown in Fig. 1A, these sequences are distinguished from all fibrillar procollagen α1 chains by the presence of proline residues within the first or second heptad repeat. Since proline residues cannot be incorporated into α-helices, other than in the first turn, the presence of prolines is likely to impede homotrimer assembly of α2 chains, thus favoring the assembly of heterotrimers in the presence of α1
chains. The fact that Lees et al (66) were able to obtain homotrimerization of proα2(I) chains, after having replaced their "molecular recognition sequence" by that of the proα1(III) chain, shows however that homotrimerization of α2(I) chains is possible, when α1(I) chains are absent.

*Inter triple-helical coiled coils* - The presence of discontinuous heptad repeats in the inter triple-helical NC2 domain of FACITs was surprising and unusual. Skips and stutters, corresponding to an insertion of one residue and a deletion of three residues, respectively, in an otherwise continuous heptad repeat have been described in a number of coiled-coil structures (68). In this case, however, there is an effective deletion of two residues between the overlapping heptad repeats. This is uncommon, and suggests either that there is a major perturbation of the structure at the site of the two-residue deletion, or that one of the heptad sequences dominates at the expense of the other. We suggest that these coiled-coil regions in the NC2 domains of the FACITs may be involved in trimerization and subsequent folding of the adjacent collagenous domains, in addition to the NC1 domains and adjoining hydroxyproline rich (GXY)n repeats (19-21). A similar conclusion has recently been reached in the case of the coiled-coil region in the NC3 domain of collagen XIII (31), where it was shown that the adjacent COL2 and COL3 regions can fold even when the N-terminal coiled coil is absent.

*Conclusion* - Heptad repeat sequences were not found in collagen IV nor in the non-emilin members of the C1q family which includes collagens VIII and X. In the case of type IV collagen, chain association is assured by the C-terminal non-triple helical domains (NC1, ~230 residues), each of which consists of two homologous sub-domains. These associate to form trimers, consisting predominantly of β-strands, with a novel protein fold (9, 10). In members of the C1q family, trimerization is assured by the C-terminal C1q domain (~160 residues), which assembles to form a tightly packed β-sandwich structure related to the tumor
necrosis factor fold (17). Thus different members of the collagen superfamily appear to have evolved three principal structural mechanisms to ensure chain trimerization in molecular assembly, either the type IV collagen NC1 domain, the C1q domain or, as suggested here, the α-helical coiled coil.

Acknowledgements - We thank E. Crouch, J.-Y. Exposito and F. Ruggiero for discussions, and M. Gordon for sharing the sequences of collagen XXIII and XIV prior to publication. The paper describing coiled-coil sequences in collagens XIII, XXIII and XXVI (31) was published while this manuscript was in the process of revision.
References


**Figure Legends**

**Figure 1. Putative α-helical coiled-coil domains in members of the collagen superfamily.**

In all sequences, hydrophobic residues in the $a$ and $d$ positions of the coiled-coil heptad repeats are highlighted. (A) Alignment of the fibrillar procollagen chains, starting with the final GPP triplet near the end of the main triple-helical region (shown in bold) and including the first ~40 residues of the C-propeptides. Known procollagen C-proteinase cleavage sites are underlined. Heptad repeats 1-4 are indicated (above), as are proline residues (white on black) which serve to demarcate the coiled-coil region. Also shown is the sequence of GCN4-pII, showing the numbering system used in Table I. (B) Alignment of the collectins, showing ~50 residues beginning near the end of the collagen-like region. (C) Alignment of invertebrate fibrillar procollagen chains, starting with the final GPP triplet near the end of the main triple-helical region and including the beginning of the C-propeptides. (D) Partial sequence of the human collagen-like tail subunit of acetylcholinesterase (ColQ) beginning near the end of the collagen-like region. Also shown is an alignment of the putative trimerization domains in collagens XV and XVIII, beginning near the end of the most C-terminal $(GXY)_n$ region, as well as an equivalent region in collagen XXVI. (E) Alignment of transmembrane collagens in the region of the putative transmembrane domain (outlined). The $(GXY)_n$ regions begin at least 57 residues C-terminal to the transmembrane regions. (F) Sequence of collagen XXVI N-terminal to the first $(GXY)_n$ region (COL1), part of which is shown (in bold). (G) Alignment of NC2 domains in the FACITs, showing also four GXY triplets in adjacent COL1 and COL2 domains. The discontinuous heptad repeat is indicated. (H) Alignment of NC3 domains in collagens XIII, XXIII and XXV, showing also four GXY triplets in adjacent COL2 and COL3 domains. (I) The triple-helical interruption in collagen VII, with four triplets from adjacent $(GXY)_n$ regions in bold. (J) Heptad repeats in von Willebrand factor A domains in collagens VII (N-terminal to the $(GXY)_n$ region) and VI (C-terminal to the $(GXY)_n$ region). Sequences are identified by name or collagen type/alpha chain, followed by species (ab=abalone, al=alvinella, ar=arenicola, rf=riftia, b=bovine, c=chicken, h=human, m=mouse, hy=hydra,
sp=sponge, su=sea urchin) then residue positions from the start of translation (where available).

**Figure 2. Structural model of the heptad containing region of the type I procollagen C-propeptide trimer based on the GCN4-pII leucine zipper structure.** The three-stranded coiled coil is shown in stereo view, with residues in positions $a$ and $d$ in red. The structure includes two $\alpha 1$ (I) chains and one $\alpha 2$ (I) chain spanning the region between the $d$ positions in heptads two and four (Fig. 1A).

**Figure 3. Amino acid sequences of coiled-coil regions attached to the full-length type IIA procollagen N-propeptide.** Fusion proteins were synthesised containing the human IIA N-propeptide fused to 18, 25 or 31 amino acids of the coiled-coil region from the human type II procollagen C-propeptide (IIA N-pro/CP II amino acids 1172-1189; 1172-1196 or 1172-1202, respectively). As a positive control to demonstrate trimerization, IIA N-pro/SP-D fusion protein was also synthesised which consists of the IIA N-propeptide attached to the coiled-coil neck domain and carbohydrate recognition domain (CRD) of rat surfactant protein D. IIA N-propeptide alone was also synthesised as a negative control. The C-terminal sequence of IIA N-pro is shown, beginning at the end of the short triple-helical region which contains 25 GXY repeats.

**Figure 4. Trimerization of the type IIA procollagen N-propeptide (IIA N-pro) by coiled-coil regions of the type II procollagen C-propeptide (CP II).** Chimeric fusion proteins were precipitated from conditioned CHO cell culture media after transient transfections, then re-suspended and chemically cross-linked with increasing concentrations (0-2 mM) of BS$^3$. Monomers (M) and cross-linked dimers (D) and trimers (T) were detected by SDS-PAGE, Western blotting and immunolabelling using anti-IIA polyclonal antisera. (A) IIA N-pro/SP-D
IIA N-pro attached to the coiled-coil neck domain and carbohydrate recognition domain (CRD) of lung surfactant protein-D) (29), positive trimerization control. (B) IIA N-pro alone, negative control. (C-E) IIA N-pro fused to sequences at the beginning of CPII (plus the terminal alanine residue in the telopeptide) containing approximately (C) four (aa 1172-1202) (D) three (aa 1172-1196) or (E) two (aa 1172-1189) heptad repeats (Figs. 1A and 3).

Figure 5. Analysis of peptides by circular dichroism. Peptides corresponding to GCN4-pII, and residues 1172-1189, 1172-1196 and 1172-1202 in human procollagen II were analyzed at 25 °C in 20 mM KH₂PO₄/NaOH, 150 mM NaF, pH 7.2, in the concentration range 200 to 400 µM, in (A) the absence and (B) the presence of 50 % TFE. Data are expressed as mean residue ellipticities [θ]MRW.
Table I. Possible interactions between charged amino acid residues in the putative coiled regions of the human fibrillar procollagen C-propeptides. Both inter-chain (between positions \(g\) and \(e\)) and intra-chain (separated by 3 or 4 residues) interactions are indicated. Residues are identified according to the GCN4-pII sequence shown in Fig. 1A.

<table>
<thead>
<tr>
<th></th>
<th>(i \rightarrow i+3)</th>
<th>(i \rightarrow i+4)</th>
<th>(g \rightarrow e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{pro}\alpha_1(I))</td>
<td>E15 - R18</td>
<td>D4 - K8</td>
<td></td>
</tr>
<tr>
<td>(\text{pro}\alpha_2(I))</td>
<td></td>
<td>D4 - K8</td>
<td></td>
</tr>
<tr>
<td>(\text{pro}\alpha_1(II))</td>
<td>E15 - R18</td>
<td>D4 - K8</td>
<td></td>
</tr>
<tr>
<td>(\text{pro}\alpha_1(III))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{pro}\alpha_1(V))</td>
<td>E15 - K18</td>
<td>K11 - E15</td>
<td>E13 - K18</td>
</tr>
<tr>
<td>(\text{pro}\alpha_1(V))</td>
<td>E15 - K18</td>
<td>K11 - E15</td>
<td>E13 - K18</td>
</tr>
<tr>
<td>(\text{pro}\alpha_2(V))</td>
<td>E15 - R18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{pro}\alpha_2(XI))</td>
<td>E15 - K18</td>
<td>K11 - E15</td>
<td>D13 - K18</td>
</tr>
<tr>
<td>(\text{pro}\alpha_2(XI))</td>
<td>D8 - R11</td>
<td>R12 - E16</td>
<td>E13 - R18</td>
</tr>
</tbody>
</table>
Table II. Oligonucleotide primers used for amplification of cDNA encoding the type IIA N-propeptide (IIA N-pro) or three regions of the type II procollagen C-propeptide domain (CPII). The three CPII fragments amplified encode the coiled-coil region corresponding to amino acid numbers 1172-1198, 1172-1196 or 1172-1202 (amino acid numbers were obtained from the published sequence: Gene Bank™ Accession number X16468). Primer pairs used for overlap extension PCR to synthesise each of the three fusion constructs (IIA N-pro/CPII aa 1172-1189/1196/1202) are shown. F=forward primer, R=reverse primer. EcoRI sites are highlighted in bold. Double underlined sequence in primer 2 corresponds to the most 5’ region of the CPII coiled-coil domain. Underlined sequence in primer 3 corresponds to the most 3’ region of the IIA N-propeptide.

<table>
<thead>
<tr>
<th>cDNA</th>
<th>Primer pairs for PCR (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIA N-propeptide</td>
<td>F 1: ggtac\textbf{gaattc}catgattgctgctggg</td>
</tr>
<tr>
<td></td>
<td>R 2: \textbf{tgccttgctgcc}catggtccttgcat</td>
</tr>
<tr>
<td>CPII (aa 1172-1189)</td>
<td>F 3: \underline{aggaccaatg}gcaccaggccagccagcc</td>
</tr>
<tr>
<td></td>
<td>R 4: catgg\underline{gaattg}ctcactctggtttgaggga</td>
</tr>
<tr>
<td>CPII (aa 1172-1196)</td>
<td>F 3: \underline{aggaccaatg}gcaccaggccagccagcc</td>
</tr>
<tr>
<td></td>
<td>R 5: catgg\underline{gaattc}ctcactctggtttgaggga</td>
</tr>
<tr>
<td>CPII (aa 1172-1202)</td>
<td>F 3: \underline{aggaccaatg}gcaccaggccagccagcc</td>
</tr>
<tr>
<td></td>
<td>R 6: catgg\underline{gaattc}ctcactctggtttgaggga</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fusion cDNA construct</th>
<th>Primer pairs for overlap extension PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIA N-pro/CPII (aa 1172-1189)</td>
<td>1 + 4</td>
</tr>
<tr>
<td>IIA N-pro/CPII (aa 1172-1196)</td>
<td>1 + 5</td>
</tr>
<tr>
<td>IIA N-pro/CPII (aa 1172-1202)</td>
<td>1 + 6</td>
</tr>
</tbody>
</table>
IAA N-pro/SF-D

...GGPGLGGNFAAAAGGFGDKKASSAGLGVQGPM/DSSALRQCGHKNGGDD PEALSHEKKAAALPFG - CED

IAA N-pro

...GGPGLGGNFAAAAGGFGDKKASSAGLGVQGPM

IAA N-pro/CFII (aa 1172-1189)

...GGPGLGGNFAAAAGGFGDKKASSAGLGVQGPM/ADQQAGGFRQHD AEDAT

IAA N-pro/CFII (aa 1172-1196)

...GGPGLGGNFAAAAGGFGDKKASSAGLGVQGPM/ADQQAGGFRQHD AEDAT/SQESNQ

IAA N-pro/CFII (aa 1172-1202)

...GGPGLGGNFAAAAGGFGDKKASSAGLGVQGPM/ADQQAGGFRQHD AEDAT/SQESNQKERS
Alpha-helical coiled-coil oligomerization domains are almost ubiquitous in the collagen superfamily
Audrey McAlinden, Thomasin A. Smith, Linda J. Sandell, Damien Ficheux, David A. D. Parry and David J. S. Hulmes

J. Biol. Chem. published online August 14, 2003

Access the most updated version of this article at doi: 10.1074/jbc.M302429200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts