Identification of the molecular recognition sequence which determines the type-specific assembly of procollagen

Janice F.Lees, Mohammed Tasab and Neil J.Bulleid[1]

School of Biological Sciences, The University of Manchester, 2.205, Stopford Building, Manchester M13 9PT, UK

[1]Corresponding author

A key question relating to procollagen biosynthesis is the way in which closely related procollagen chains discriminate between each other to assemble in a type-specific manner. Intracellular assembly of procollagen occurs via an initial interaction between the C-propeptides followed by vectorial propagation of the triple-helical domain in the C to N direction. Recognition signals within the C-propeptides must, therefore, determine the selective association of individual procollagen chains. We have used the proα1 chain of type III procollagen [proα1(III)] and the proα2 chain of type I procollagen [proα2(I)] as examples of procollagen chains that are either capable or incapable of self-assembly. When we exchanged the C-propeptides of the proα1(III) chain and the proα(I) chain we demonstrated that this domain is both necessary and sufficient to direct the assembly of homotrimers with correctly aligned triple-helices. To identify the sequences within this domain that determine selective association we constructed a series of chimeric procollagen chains in which we exchanged specific sequences from the proα1(III) C-propeptide with the corresponding region within the proα2(I) C-propeptide (and vice versa) and assayed for the ability of these molecules to form homotrimers. Using this approach we have identified a discontinuous sequence of 15 amino acids which directs procollagen self-association. By exchanging this sequence between different procollagen chains we can direct chain association and, potentially, assemble molecules with defined chain compositions.

Keywords: chain selection/molecular recognition/procollagen

Introduction

Collagens are complex multi-domain proteins composed of three α chains wound into a triple-helix (reviewed by Van der Rest and Garrone, 1991; Kielty *et al.*, 1993; Kadler, 1995). The collagen super-family includes over 30 different gene products that assemble into at least 19 genetically distinct collagen types which may be homotrimeric or heterotrimeric in nature (Prockop and Kivirikko, 1995*)*. The ubiquitous nature of collagen is such that a single cell may co-express several collagen types. For example, skin fibroblasts are able to synthesize six highly homologous but genetically distinct procollagen

chains necessary for the assembly of fibrillar collagen types I, III and V yet, despite the level of identity, they invariably assemble in a collagen type-specific manner. The ability of individual procollagen chains to discriminate between each other is, therefore, a central feature of the assembly process, ensuring that each collagen molecule has the correct complement of proα chains.

It is presumed that procollagen chains first associate through a series of non-covalent interactions between the non-collagenous C-propeptide domains to form a trimer, an interaction that is stabilized by the formation of interchain disulfide bonds (Schofield *et al.*, 1974; Olsen *et al.*, 1976; Bachinger *et al.*, 1980, 1981; Doege and Fessler, 1986; Prockop *et al.*, 1990). Interaction between the three chains via the C-propeptide provides the correct alignment and registration of the chains necessary to nucleate the triple-helix, which is then propagated vectorially in a C- to N-terminal direction (Bachinger *et al.*, 1980, 1981). The mechanism(s) that determines chain selectivity must, therefore, operate at an early stage in the assembly pathway through specific recognition sequences which reside within the C-propeptide domain. Comparative analysis of the C-propeptide domains of the eight procollagen molecules that assemble to form the fibrillar collagens types I, II, III, V and XI reveals a high degree of sequence homology (Dion and Myers, 1987; Kadler, 1995), presumably reflecting the evolution of a common tertiary structure. It seems likely, therefore, that the unique features necessary to ensure chain discrimination reside within one or more of the divergent regions that punctuate the C-propeptide.

Type I collagen exists as a heterotrimer comprising two proα1(I) chains and one proα2(I) chain ([proα1(I)$_2$ proα2(I)]. While proα1(I) chains have the ability to form homotrimers, albeit with low efficiency (Jimenez *et al.*, 1977; Moro and Smith, 1977), proα2(I) homotrimers have not been detected and the inclusion of this chain into a trimer is, therefore, dependent upon its association with proα1(I) chains. Type III collagen is an obligate homotrimer comprising three proα1(III) chains [proα1(III)]$_3$, which are unable to co-assemble with any other procollagen monomer. In order to define the molecular mechanisms underlying chain selectivity we devised an experimental strategy that was based on the assumption that transfer of the C-propeptide (or sequences within the C-propeptide) from the homotrimeric proα1(III) chain to the proα2(I) molecule would be sufficient to direct self-association and assembly into homotrimers. We are able to reconstitute the initial stages in the assembly of procollagen by expressing specific RNAs in a cell-free translation system in the presence of semi-permeabilized cells known to carry out the co- and post-translational modifications required to ensure assembly of a correctly aligned triple-helix (Wilson *et al.*, 1995; Bulleid *et al.*, 1996). By

analysing the folding and assembly pattern of a series of chimeric procollagen chains in which we exchanged specific regions of the proα1(III) C-propeptide with the corresponding region within the proα2(I) chain (and vice versa) we have identified a short discontinuous sequence of 15 amino acids within the proα1(III) C-propeptide which directs procollagen self-association. This sequence is, therefore, responsible for the initial recognition event and is necessary to ensure selective chain association.

Results

Transfer of the proα1(III) C-propeptide to the proα(I)2 chain is sufficient to direct self-assembly

We based our experimental strategy on the assumption that transfer of the C-propeptide from the proα1(III) chain to the proα2(I) chain should be sufficient to direct self-recognition and assembly into homotrimers. Hence, by exchanging different regions within the proα1(III) C-propeptide with the corresponding sequence from the proα2(I) chain we intended to distinguish between sequences that direct the folding of the tertiary structure and those involved in the selection process. To simplify analysis of the translation products we constructed chimeric procollagen molecules from two parental procollagen 'mini-chains', proα1(III)Δ1 and proα2(I)Δ1. These molecules, which have been described previously (Lees and Bulleid, 1994), comprise both the N- and C-propeptides together with truncated triple-helical domains. We tested the initial assumption by analysing the folding and assembly of chimeric procollagen chains in which the C-propeptide of the proα2(I) chain was substituted with the equivalent domain from the proα1(III)Δ1 chain (proα2-(I):(III)CP) and, conversely, where we replaced the C-propeptide of proα1(III) chain with that from proα2(I)Δ1 chain (proα1(III):(I)CP) (Figure 1). The C-propeptide (CP) junction points were determined by the sites of cleavage by the procollagen C-proteinase (PCP) which occurs between Ala and Asp (residues 1119–1120) in the proα2(I) chain (Hojima *et al.*, 1985; Kessler *et al.*, 1996). In the absence of data regarding the precise location of cleavage within the proα1(III) chain (Kadler, 1995), we chose to position the junction between Ala and Pro (residues 1217–1218). However, Kessler and co-workers (1996) have subsequently shown that cleavage by PCP occurs between Gly and Asp (residues 1222–1223), with the consequence that recombinant proα2(I):(III)CP includes an additional four residues derived from the proα1(III) C-telopeptide, whilst the C-telopeptide in construct proα1(III):(I)CP is missing those same four amino acids. RNA transcripts were transcribed *in vitro* and expressed in a cell-free system comprising a rabbit reticulocyte lysate optimized for the formation of disulfide bonds supplemented with semi-permeabilized HT1080 cells (SP-cells), which we have shown previously to carry out the initial stages in the folding, post-translational modification and assembly of procollagen (Bulleid *et al.*, 1996). The C-propeptides of both proα1(III) and proα2(I) chains contain cysteine residues which participate in the formation of interchain disulfide bonds (Schofield *et al.*, 1974; Olsen *et al.*, 1976; Bachinger *et al.*, 1981; Prockop *et al.*, 1990). Translation products were, therefore, separated by SDS–PAGE under reduced and non-reduced conditions in order to detect disulfide-bonded trimers. Translation of the parental molecules proα1(III)Δ1 and proα2(I)Δ1 yielded major products of ~77 kDa and 61 kDa respectively (Figure 2A, lanes 1 and 2), the size differential being accounted for by the relative molecular weights of the N-propeptides and truncated triple-helical domains in each molecule (Lees and Bulleid, 1994). The heterogeneity of the translation products is due to hydroxylation of proline residues in the triple-helical domain that leads to an alteration in electrophoretic mobility (Cheah *et al.*, 1979). The additional lower molecular weight proteins present in lanes 3 and 7 probably represent translation products obtained after initiation of translation at internal start codons. We have previously shown that these minor translation products are not translocated into the endoplasmic reticulum (Lees and Bulleid, 1994). The presence of high molecular weight species under non-reducing but not reducing conditions is indicative of interchain disulfide bond formation. Separation under non-reduced conditions revealed that proα1(III)Δ1, but not proα2(I)Δ1, chains were able to self-associate to form disulfide-bonded trimers (Figure 2A, lanes 5 and 6). A similar examination of chimeric chains proα2(I):(III)CP and proα1(III):(I)CP revealed that only proα2(I):(III)CP chains were able to form disulfide-bonded homotrimers (Figure 2A, lanes 3, 4, 7 and 8) demonstrating that the C-propeptide from type III procollagen is both necessary and sufficient to drive the initial association between procollagen chains.

We have shown previously that proα1(III)Δ1 chains synthesized in the presence of SP-cells were resistant to a combination of pepsin, chymotrypsin and trypsin in a standard assay (Bruckner and Prockop, 1981) used specifically to detect triple-helical procollagen (Bulleid *et al.*, 1996). We confirmed that proα2(I):(III)CP chains had the ability to form a correctly aligned triple-helix by performing a thermal denaturation experiment in which translated material was heated to various temperatures prior to protease treatment (Figure 2B). The results indicate that at temperatures below 35°C a protease-resistant triple-helical fragment is present, but at temperatures above 35°C the triple-helix melts and becomes protease sensitive (Figure 2B, lanes 1–10). The melting temperature (T_m) was calculated to be ~35.5°C after quantification by phosphorimage analysis. The T_m value obtained for proα2-(I):(III)CP is significantly lower than the figure of 39.5°C obtained for proα1(III)ΔI (Bulleid *et al.*, 1996) and probably reflects the percentage of hydroxyproline residues relative to the total number of amino acids in the triple-helical domain (11% and 15% respectively). These results indicate that transfer of the proα1(III) C-propeptide enables us to generate an entirely novel procollagen species comprising three proα2(I) chains that fold into a correctly aligned triple-helix.

Assembly of recombinant procollagen chains with chimeric C-propeptides

Given that the proα2(I):(III)CP hybrid procollagen chain includes all of the information required for self-association we reasoned that progressive removal of the proα1(III) C-propeptide sequence and replacement with the corresponding proα2(I) sequence would eventually disrupt the chain selection mechanism. Conversely, we anticipated that transfer of progressively more proα1(III) C-propeptide
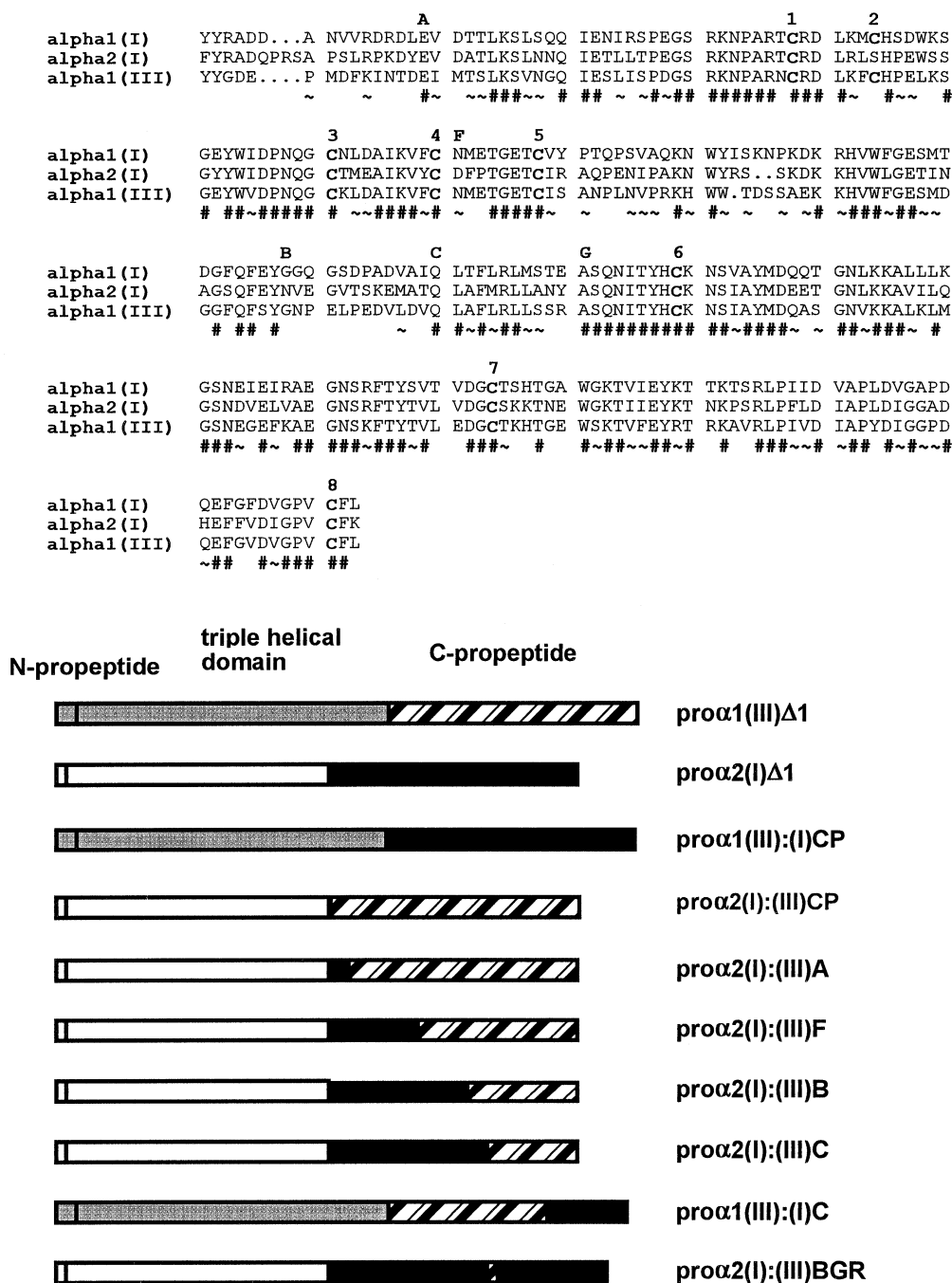
```
                                            A                              1          2
alpha1(I)     YYRADD...A NVVRDRDLEV DTTLKSLSQQ IENIRSPEGS RKNPARTCRD LKMCHSDWKS
alpha2(I)     FYRADQPRSA PSLRPKDYEV DATLKSLNNQ IETLLTPEGS RKNPARTCRD LRLSHPEWSS
alpha1(III)   YYGDE....P MDFKINTDEI MTSLKSVNGQ IESLISPDGS RKNPARNCRD LKFCHPELKS
                 ~          ~        #~    ~~###~~  #  ## ~ ~#~## ###### ### #~  #~~   #

                               3          4    F          5
alpha1(I)     GEYWIDPNQG CNLDAIKVFC NMETGETCVY PTQPSVAQKN WYISKNPKDK RHVWFGESMT
alpha2(I)     GYYWIDPNQG CTMEAIKVYC DFPTGETCIR AQPENIPAKN WYRS..SKDK KHVWLGETIN
alpha1(III)   GEYWVDPNQG CKLDAIKVFC NMETGETCIS ANPLNVPRKH WW.TDSSAEK KHVWFGESMD
              # ##~##### # ~~####~#  ~  #####~   ~   ~~~ #~ #~ ~   ~ ~# ~###~##~~

                         B          C                    G          6
alpha1(I)     DGFQFEYGGQ GSDPADVAIQ LTFLRLMSTE ASQNITYHCK NSVAYMDQQT GNLKKALLLK
alpha2(I)     AGSQFEYNVE GVTSKEMATQ LAFMRLLANY ASQNITYHCK NSIAYMDEET GNLKKAVILQ
alpha1(III)   GGFQFSYGNP ELPEDVLDVQ LAFLRLLSSR ASQNITYHCK NSIAYMDQAS GNVKKALKLM
              # ## #         ~    # #~#~##~~  ########## ##~####~  ~ ##~###~ #

                                               7
alpha1(I)     GSNEIEIRAE GNSRFTYSVT VDGCTSHTGA WGKTVIEYKT TKTSRLPIID VAPLDVGAPD
alpha2(I)     GSNDVELVAE GNSRFTYTVL VDGCSKKTNE WGKTIIEYKT NKPSRLPFLD IAPLDIGGAD
alpha1(III)   GSNEGEFKAE GNSKFTYTVL EDGCTKHTGE WSKTVFEYRT RKAVRLPIVD IAPYDIGGPD
              ###~ #~ ## ###~###~# ###~  #  #~##~~##~# #  ###~~# ~## #~#~~#

                         8
alpha1(I)     QEFGFDVGPV CFL
alpha2(I)     HEFFVDIGPV CFK
alpha1(III)   QEFGVDVGPV CFL
              ~## #~### ##
```


Fig. 1. Alignment plot of the C-propeptides of the proα-chains from type I and III procollagen (top) and a schematic representation of the chimeric of procollagen chains described in the text (bottom). The C-proteinase cleavage site is located between Ala and Asp (residues 4–5) for the type I procollagen chains and between Gly and Asp (residues 3–4) for proα1(III) chain. The alignment shows amino acids which are identical (#) or those with conserved side-chains (~). The conserved cysteine residues are numbered 1–8, while letters A, B, C, F and G denote the first amino acid at the junctions between proα1(III) and proα2(I) sequences in the hybrid molecules.

sequence to the proα1(IIII):(I)CP chimeric chain would yield a molecule which was capable of self-assembly. We constructed a series of procollagen chains with chimeric C-propeptide domains and assessed the ability of individual chains to form homotrimers with stable triple-helical domains. A schematic representation of these recombinants is presented in Figure 1, with the letters A, B, C, F and G denoting the position of each junction. It should be noted that the proα1(III) and proα2(I) C-propeptides differ in their complement of cysteine residues, with proα2(I) lacking the Cys2 residue. Our

previous data suggest that interchain disulfide bonds within the C-propeptide of type III procollagen form exclusively between Cys2 and 3 (Lees and Bulleid, 1994). However, interchain disulfide bonding, between either the C-propeptides or C-telopeptides is not required for chain association and triple-helix formation (Bulleid *et al.*, 1996), therefore, it is possible that homotrimers may form between chimeric procollagen chains which lack either the C-propeptide Cys2 residue or the C-telopeptide cysteine [only found in the triple-helical domain of proα1(III)]. These molecules will not, however, contain interchain disulfide bonds and,
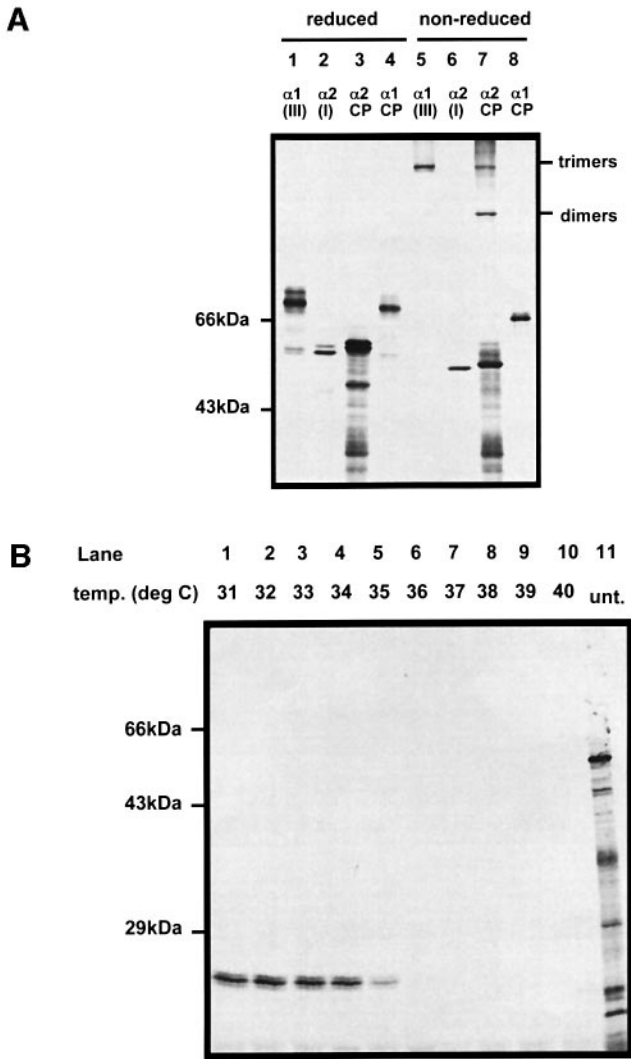
Fig. 2. Interchain disulfide bond formation among chimeric procollagen chains in which the C-propeptide domains were exchanged. (**A**) Parental molecules, proα1(III)Δ1 [α1(III)], proα2(I)Δ1 [α2(I)] and hybrid chains proα2(I):(III)CP [α2:CP], proα1(III):(I)CP [α1:CP] were expressed in a rabbit reticulocyte lysate in the presence of semi-permeabilized (SP) HT1080 cells. SP-cells were isolated by centrifugation, solubilized and the translation products separated by SDS–PAGE through a 7.5% gel under reducing (lanes 1–4) or non-reducing conditions (lanes 5–8). (**B**) Heat denaturation of proα2(I):(III)CP triple-helix. Proα2(I):(III)CP RNA was translated in the presence of SP-cells. SP-cells were isolated by centrifugation, solubilized and treated with pepsin (100 µg/ml). The reaction mixture was neutralized, diluted in chymotrypsin/trypsin digest buffer and divided into aliquots, each aliquot being heated to a set temperature prior to digestion with a combination of trypsin (100 µg/ml) and chymotrypsin (250 µg/ml). The samples were analysed by SDS–PAGE through a 12.5% gel under reducing conditions (lanes 1–10). Lane 11 (unt.) contains translation products which have not been treated with proteases.

Fig. 3. Trimerization and triple-helix formation among chimeric procollagen chains. (**A**) Parental chains proα1(III)Δ1, proα2(I)Δ1, and hybrids proα2(I):(III)CP, A, F, F^{s-c}, B^{s-c}, C^{s-c}, proα1(III):(I)C (α2CP, A, F, F^{s-c}, B^{s-c}, C^{s-c}, α1C) were translated in a rabbit reticulocyte lysate in the presence of SP-cells. The SP-cells were isolated by centrifugation, solubilized and a portion of the translated material separated by SDS–PAGE under non-reducing conditions through a 7.5% gel (lanes 1–9). (**B**) The remainder of the samples were treated with pepsin (100 µg/ml) prior to neutralization and digestion with a combination of trypsin (100 µg/ml) and chymotrypsin (250 µg/ml). The proteolytic digestion products were analysed by SDS–PAGE through a 12.5% gel under reducing conditions (lanes 1–9).

as a consequence, will not appear as oligomers after analysis under non-reducing conditions. To circumvent this problem, where appropriate, we generated our hybrid chains from a recombinant proα2(I)Δ1^{s-c} (Lees and Bulleid, 1994) in which the existing serine residue was substituted for cysteine, thus restoring the potential to form trimers stabilized by interchain disulfide bonds. It should also be noted that whilst proα1(III):(I)CP lacks Cys2, it does still retain the potential to form disulfide-bonded trimers by virtue of the two cysteine residues

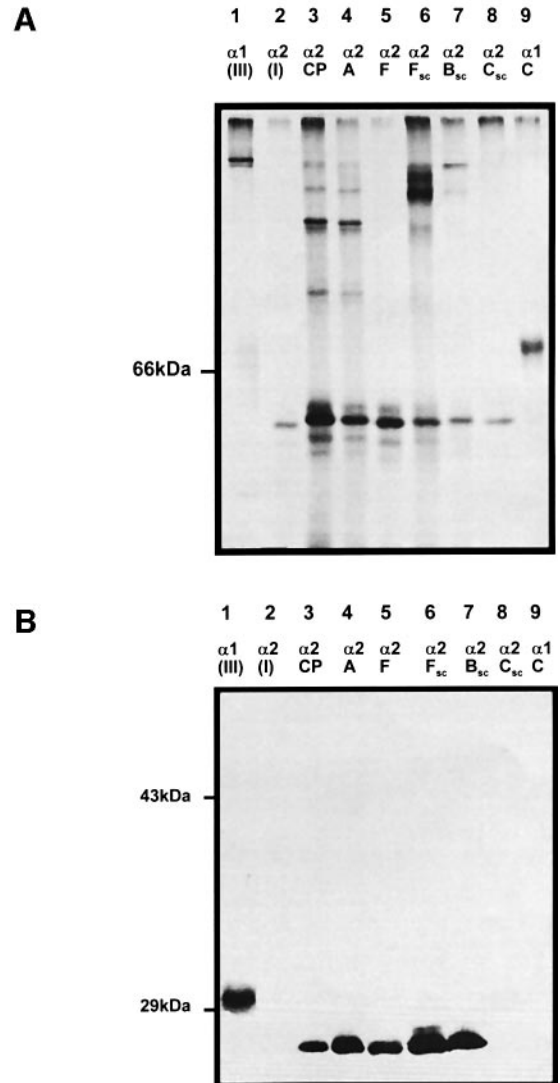located at the junction of the triple-helical domain and the C-telopeptide. Parental chains proα2(I)Δ1 and hybrids proα2(I):(III)CP, A, F, F^{s-c}, B^{s-c}, C^{s-c}, proα1(III):(I)C were translated in the presence of SP-cells and the products separated by SDS–PAGE under non-reducing conditions (Figure 3A). The results demonstrate that recombinants proα1(III)Δ1, proα2(I):(III)CP, A, F^{s-c}, B^{s-c} (Figure 3A, lanes 1, 3, 4, 6 and 7) are able to form interchain disulfide-bonded trimers and dimers while proα2(I)Δ1, proα2(I):(III)F, C^{s-c} and proα1(III):(I)C (Figure 3A, lanes 2, 5, 8 and 9) remain monomeric. We have already demonstrated that interchain disulfide bonding is not a prerequisite for triple-helix formation (Bulleid *et al.*, 1996), therefore, the inability to form disulfide-bonded

trimers does not preclude the possibility that the molecules assemble to form a triple-helix. To ascertain whether the chimeric chains had the ability to fold into a correctly aligned triple-helix, we treated translation products with a combination of pepsin, chymotrypsin and trypsin and analysed the digested material under reducing conditions by SDS–PAGE. As shown in Figure 3B, recombinants proα1(III)Δ1, proα2(I):(III)CP, A, F^{s-c}, F, B^{s-c} (Figure 3B, lanes 1, 3, 4, 5, 6 and 7) all yielded protease-resistant fragments. The size differential reflects the relative lengths of the triple-helical domains in each of the parental molecules [proα2(I)Δ1–185 residues and proα1(III)Δ1–192 residues]. The ability of proα2(I):(III)F to form a stable triple-helix confirms our earlier data that interchain disulfide bonding is not necessary for triple-helix folding. Thus, hybrid molecules containing sequences from the proα2 C-propeptide between the C-propeptide cleavage site and the B-junction are able to form homotrimers with stable triple-helical domains and, therefore, contain all of the information necessary to direct chain self-assembly. These results indicate that the signal(s) which controls chain selectivity must be located between the B-junction and the C-terminus of the C-propeptide. Neither proα2-(I):(III)C^{s-c} nor proα1(III):(I)C chains are able to fold into a triple-helix. The inability of these reciprocal constructs to self-associate suggests that chain selectivity is mediated, either by a co-linear sequence that spans the C-junction or by discontinuous sequence domains located on either side of the C-junction.

Identification of a sequence motif from the proα1(III) C-propeptide which directs chain self-assembly

Procollagen chain selectivity is probably mediated through one or more of the variable domains located within the C-propeptide. The sequence between the B- and C-junctions is one of the least conserved among the procollagen C-propeptides (Figure 1), yet we have demonstrated that inclusion of this domain, in the absence of proα1(III) sequence distal to the C-junction, is not sufficient to direct chain assembly. To ascertain whether we had indeed interrupted the recognition sequence for chain recognition we generated a further recombinant, proα2-(I):(III)BGR^{s-c} (B–G replacement), which contained all of the proα2(I)Δ1 sequence apart from the Ser→Cys mutation at Cys2 and a stretch of 23 amino acids derived from the type III C-propeptide which spans the C-junction from points B to G, the B–G motif: ^{b}GNPEL-PEDVLDV^{c}QLAFLRLLSSR^g (underscoring indicates the most divergent residues, see Figure 1). The location of the G-boundary in the replacement motif allowed for the inclusion of the first non-conserved residues after the C-junction (SR). When expressed in the presence of SP-cells the chimeric proα2(I):(III)BGR^{s-c} chains were able to form inter-chain disulfide-bonded molecules (Figure 4A, lane 6) demonstrating that the C-propeptides were capable of self-association. Furthermore, this hybrid was able to fold and form a stable triple-helix as judged by the formation of a protease-resistant fragment (Figure 4B, lane 3). Proα2(I):(III)BGR^{s-c} contains a Ser→Cys substitution which enabled us to assay for the formation of disulfide-bonded trimers. Our previous data demonstrated that this substitution alone does not enable
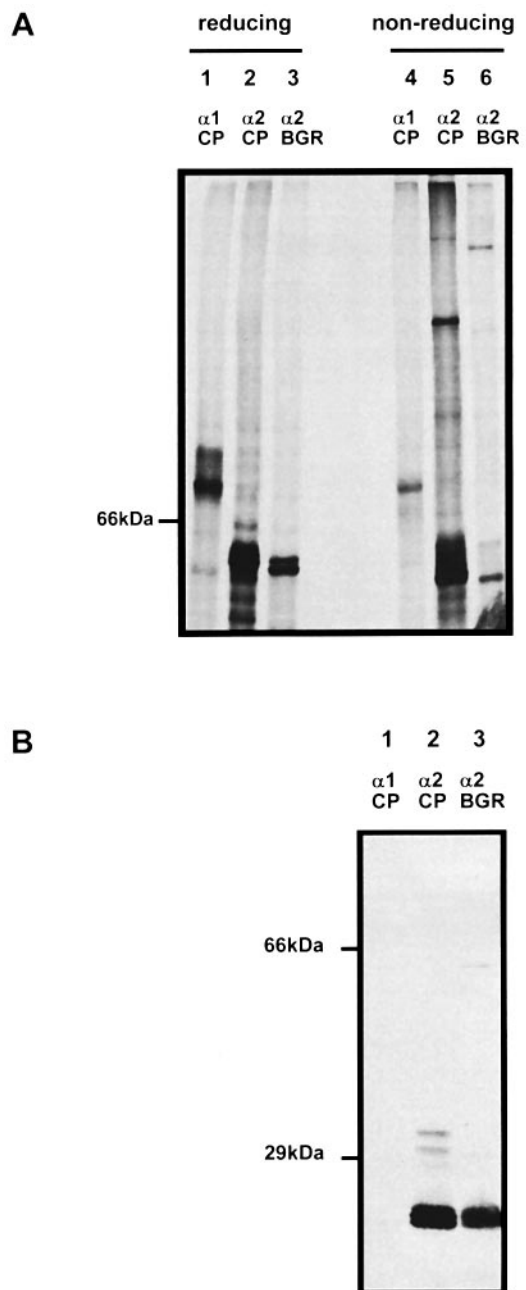

Fig. 4. Trimerization and triple-helix formation among chains containing the 23 amino acid B–G motif. (**A**) Recombinant procollagen chains proα1(III):(I)CP, proα2(I):(III)CP and proα2(I):(III)BGR^{s-c} were expressed in a reticulocyte lysate supplemented with SP-cells. The SP-cells were isolated by centrifugation, solubilized and a portion of the translated material separated by SDS–PAGE through a 7.5% gel, under reducing (lanes 1–3) or non-reducing conditions (lanes 4–6). (**B**) The remainder of the samples were treated with pepsin (100 µg/ml) prior to neutralization and digestion with a combination of trypsin (100 µg/ml) and chymotrypsin (250 µg/ml). The proteolytic digestion products were analysed by SDS–PAGE through a 12.5% gel under reducing conditions (lanes 1–3).

wild-type proα2(I)Δ1 chains to form homotrimers (Lees and Bulleid, 1994). Nevertheless, to eliminate the possibility that this mutation influences the assembly pattern we created a revertant proα2(I):(III)BGR^{c-s} which contains the wild-type complement of Cys residues. As expected proα2(I):(III)BGR^{c-s} was unable to form disulfide-bonded
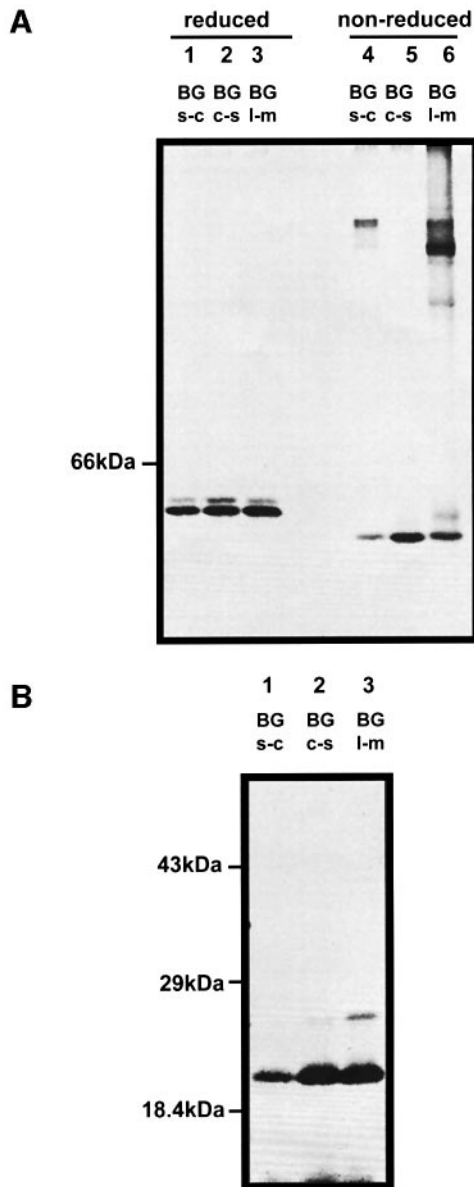
A

B

Fig. 5. The effect of Cys–Ser reversion and Leu–Met mutation on the assembly of proα2(I):(III)BGR chains. (**A**) Recombinant procollagen chains proα2(I):(III)BGR^{s-c} proα2(I):(III)BGR^{c-s}, proα2(I):(III)BGR^{l-m} were translated in a reticulocyte lysate supplemented with SP-cells. The cells were isolated by centrifugation, solubilized and a portion of the translated material separated by SDS–PAGE through a 7.5% gel, under reducing (lanes 1–3) or non-reducing conditions (lanes 4–6). (**B**) The remainder of the samples were treated with pepsin (100 μg/ml) prior to neutralization and digestion with a combination of trypsin (100 μg/ml) and chymotrypsin (250 μg/ml). The proteolytic digestion products were analysed by SDS–PAGE through a 12.5% gel under reducing conditions (lanes 1–3).

trimers (Figure 5A, lane 5) but did assemble correctly into a protease-resistant triple-helix (Figure 5B, lane 2). Thus, the 23-residue B–G motif contains all of the information required to direct procollagen self-assembly.

The ability of the proα2(I):(III)BGR^{s-c} chains to form interchain disulfide bonds suggests that this molecule is able to associate via its C-propeptide. However, to confirm that this is indeed the case we carried out a collagenase digestion of the products of translation (Figure 6). Bacterial collagenase specifically digests the triple-helical domain, leaving both the N- and C-propeptides intact. The
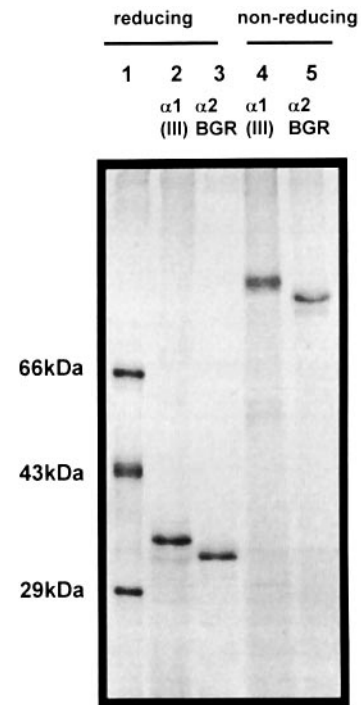

Fig. 6. Inter-chain disulfide bonds form between proα2(I):(III)BGR C-propeptides. Recombinant procollagen chains proα1(III)Δ1 and proα2(I):(III)BGR were translated in a reticulocyte lysate supplemented with SP-cells. The cells were isolated by centrifugation, solubilized and digested with 1.5 units of bacterial collagenase. The products of digestion were analysed by SDS–PAGE through a 10% gel under reducing (lanes 2 and 3) or non-reducing (lanes 4 and 5) conditions.

N-propeptides of both chains do not contain any methionine residues and as a consequence, the only radiolabelled product remaining after digestion is the C-propeptide. Comparison of the samples separated under reducing and non-reducing conditions demonstrated that inter-chain disulfide-bonded trimers were formed within the C-propeptides of proα1(III)Δ1 and proα2(I):(III)BGR^{s-c} chains (Figure 6, lanes 2 and 4, and 3 and 5). This demonstrates that these chains do indeed associate via their C-propeptides.

The effect of Leu→Met substitution on proα2(I):BGR assembly

Analysis of the 23 amino acid B–G motif from the proα1(III) and proα2(I) chains (Figure 7) indicates that residues 13–20 (QLAFLRLL) are identical with the exception of position 17, Leu (L) in proα1(III) and Met (M) in proα2(I). Using site-directed mutagenesis we substituted the existing Leu residue with Met to create proα2(I):(III)BGR^{l-m} and monitored the effect of this mutation on chain assembly. The Leu→Met mutagenesis was performed using recombinant proα2(I):(III)BGR^{s-c} as the parental molecule, so the mutant chains retained the potential to form disulfide-bonded trimers. As illustrated in Figure 5, both proα2(I):(III)BGR^{s-c} and proα2(I):(III)BGR^{l-m} were able to form interchain disulfide-bonded molecules when analysed under non-reducing conditions (Figure 5A, lanes 4 and 6) and both constructs formed protease-resistant triple-helical domains (Figure 5B, lanes 1 and 3). The Leu→Met substitution did not, therefore, disrupt the process of chain selection nor did it

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
alpha1(I)	G	G	Q	G	S	D	P	A	D	V	A	I	Q	L	T	F	L	R	L	M	S	T	E
alpha2(I)	N	V	E	G	V	T	S	K	E	M	A	T	Q	L	A	F	M	R	L	L	A	N	Y
alpha1(II)	G	D	D	N	L	A	P	N	T	A	N	V	Q	M	T	F	L	R	L	L	S	T	E
alpha1(III)	G	N	P	E	L	P	E	D	V	L	D	V	Q	L	A	F	L	R	L	L	S	S	R
alpha1(V)	V	D	A	E	G	N	P	V	G	V	.	V	Q	M	T	F	L	R	L	L	S	A	S
alpha2(V)	G	D	H	Q	S	P	N	T	A	I	.	T	Q	M	T	F	L	R	L	L	S	K	E
alpha1(XI)	L	D	V	E	G	N	S	I	N	M	.	V	Q	M	T	F	L	K	L	L	T	A	S
alpha2(XI)	V	D	S	E	G	S	P	V	G	V	.	V	Q	L	T	F	L	R	L	L	S	V	S

Fig. 7. Sequence alignment of the chain selectivity recognition domains in other fibrillar procollagens. Sequence homology within the 23 residue B–G motif is illustrated, the shaded regions indicating the position of the unique 15 residue sub-domain which directs procollagen chain discrimination.

prevent the formation of a correctly aligned triple-helix. Our observations led us to conclude that a discontinuous sequence of 15 amino acids: (GNPELPED-VLDV........SSR) contains all of the information necessary to allow procollagen chains to discriminate between each other and assemble in a type-specific manner.

Discussion

The molecular mechanism which enables closely related procollagen chains to discriminate between each other is a central feature of the assembly pathway. The initial interaction between the C-propeptides both ensures that the constituent chains are correctly aligned prior to nucleation of the triple-helix and propagation in a C- to N-direction, and that component chains associate in a collagen type-specific manner (Bachinger et al., 1981). As a consequence, recognition signals which determine chain selectivity are assumed to reside within the primary sequence of this domain, presumably within a region(s) of genetic diversity. By generating chimeric procollagen molecules from parental 'mini-chains' proα1(III)Δ1 and proα2(I)Δ1 we have demonstrated that transfer of the proα1(III) C-propeptide to the naturally heterotrimeric proα2(I) molecule was sufficient to direct formation of homotrimers. Furthermore, analysis of a series of molecules in which we interchanged specific sequences from proα1(III) and proα2(I) C-propeptides allowed us to identify a discontinuous sequence of 15 amino acids (GNPELPEDVLDV........SSR) within the proα1(III) C-propeptide, which, if transferred to the corresponding region within the proα2(I) chain, enabled those chains to trimerize. Transfer of the proα1(III) recognition motif to the proα2(I) chain did not appear to have an adverse effect on chain alignment, allowing the triple-helical domains to fold into a protease-resistant conformation. This sequence motif is, therefore, both necessary and sufficient to ensure that procollagen chains discriminate between each other and assemble in a type-specific manner.

In order to establish a structure–function relationship for the chain recognition domain, we examined the hydropathy profile and secondary structure potential of the 23-residue B–G sequence: GNPELPEDVLDVQLAFLRLLSSR (Chou and Fasman, 1978). The data indicate that the 15-residue chain recognition motif: GNPEL-

PEDVLDV........SSR is markedly hydrophilic, in contrast to the hydrophobic properties of the conserved region: QLAFLRLL. These features are entirely consistent with a potential role for this motif in mediating the initial association between the component procollagen monomers. An examination of the 15-residue recognition motif from other fibrillar procollagens predicts that they are all relatively hydrophilic and probably assume a similar structural conformation, regardless of the degree of diversity in the primary sequence (Figure 7). It is, presumably, the nature of the amino acid changes which provides the distinguishing topographical features necessary to ensure differential chain association. An examination of the B–G sequence alignment (Figure 7) indicates that residues 1, 2, 12 and 21 are more tightly conserved than amino acids 3–11, 22 and 23, suggesting that the latter may form a core recognition sequence that is of critical importance in the selection process. We do not know whether the other four residues participate directly in chain discrimination but this can be tested experimentally by site-directed mutagenesis.

The initial association between procollagen chains has been attributed to a series of ill-defined non-covalent interactions between the C-propeptides. In the absence of three-dimensional crystalographic data we can only speculate as to the nature of the interactive forces involved. We have identified the functional domain which determines chain selectivity and propose that trimerization is initiated via an interaction(s) between these recognition sequences. It is unclear, however, whether the interactions which determine chain composition are the same as those which allow productive association and stabilization of the trimer. It is also possible to envisage an alternative scenario in which assembly occurs by default and that the presence of an inappropriate recognition sequence prevents productive association from occurring. The nature of potential stabilizing interactions is uncertain, but recent data (Bulleid et al., 1996) indicate that, for type III procollagen at least, the formation of interchain disulfide bonds does not play a direct role in procollagen assembly. It has also been postulated that a cluster of four aromatic residues, which are conserved in the fibrillar collagens, collagens X, VIII and collagen-like complement factor C1q, may be of stategic importance in trimerization (Brass et al., 1992).

The C-telopeptides were originally proposed to have a role in both procollagen assembly (Doege and Fessler,

1986) and chain discrimination (Dion and Myers, 1987), the latter by virtue of the level of sequence diversity between various procollagen chains. However, we have recently demonstrated (Bulleid *et al.*, 1996) that the C-telopeptides of type III collagen do not interact prior to nucleation of the triple-helix, ruling out a role for this peptide sequence in the initial association of the C-propeptides. Data obtained from the assembly of hybrid chains indicates that the ability to discriminate between chains does not segregate with the species of C-telopeptide, lending support to this assertion.

Using this approach we have been able to synthesize an entirely novel procollagen species comprising three proα2(I)Δ1 chains [proα2(I)Δ1]$_3$. Throughout this study we used procollagen 'mini-chains' with truncated triple-helical domains; however, we have recently demonstrated that full-length proα2(I) chains containing the 15-residue proα1(III) recognition sequence also self-associate into a triple-helical conformation (data not shown). Thus, the ability to introduce the chain recognition sequence into different procollagen chains should provide the means to design novel collagen molecules with defined chain compositions. This, in turn, introduces the possibility of producing collagen matrices with defined biological properties, such as enhanced or differential cell-binding or adhesion properties. Furthermore, the identification of a short peptide sequence which directs the initial association between procollagen chains may provide a target for therapeutic intervention allowing for the modulation or inhibition of collagen deposition.

Materials and methods

Construction of recombinant plasmids

Recombinants pα1(III)Δ1 and pα2(I)Δ1 have been described previously (Lees and Bulleid, 1994). Chimaeric molecules were generated by PCR overlap extension using the principles outlined by Horton (1993). PCRs (100 μl) comprised template DNA (500 ng), oligonucleotide primers (100 pmol each) in 10 mM KCl, 20 mM Tris–HCl pH 8.8, 10 mM (NH₄)₂SO₄, 2 mM MgSO₄, 0.1% (v/v) Triton X-100, 300 μM each dNTP. Ten rounds of amplification were performed in the presence of 1 unit Vent DNA poymerase (New England Biolabs, MA). Recombinants pα2(I)Δ1:(III)CP, A, F, F$^{s–c}$, B$^{s–c}$, C$^{s–c}$ were generated using a 5′ oligonucleotide primer (5′AGATGGTCGCACTGGACATC 3′) complementary to a sequence 70 bp upstream of an *Sfi*I site in pα2(I)Δ1 and a 3′ oligonucleotide primer (5′ TCGCAGGGATCCGT-CGGTCACTTGCACTGGTT 3′) complementary to a region 100 bp downstream of the stop codon in pα1(III)Δ1. A *Bam*HI site was introduced into this primer to facilitate subsequent sub-cloning steps. Pairs of internal oligonucleotides, one of which included a 20 nucleotide overlap, were designed to generate molecules with precise junctions as delineated in Figure 1. Overlap extension yielded a product of ~990 bp which was purified, digested with *Xho*I and *Bam*HI and ligated into pα2(I)Δ1 from which a 1080 bp *Xho*I–*Bam*HI fragment had been excised. Recombinants pα1(III)Δ1:(I)CP,C were synthesized in a similar manner using a 5′ oligonucleotide (5′ AATGGAGCTCCTGGACCCATG 3′) complementary to a sequence 100 bp upstream of an *Xho*I site in pα1(III)Δ1 and a 3′ amplification primer (5′CTGCTAGGTACCA-AATGGAAGGATTCAGCTTT 3′) which incorporated a *Kpn*I site and was complementary to a region 100 bp downstream of the stop codon in pα2(I)Δ1. Overlap extension produced a fragment of 1100 bp which was digested with *Xho*I and *Kpn*I and ligated into pα1(III)Δ1 from which an 1860 bp fragment had been removed. Recombinant pα2(I):(III)BGR was constructed using the same 5′ amplification primer used to synthesize the proα2(I)Δ1:(III) series of chimeras and a 3′ oligonucleotide which was identical to that used to generate the proα1(III-)Δ1:(I)CP,C constructs except that it contained a *Bam*HI site instead of *Kpn*I (both complementary to pα2(I)Δ1). Primary amplification products were generated from pα2(I)Δ1:(III)B$^{s–c}$ and pα2(I)Δ1 with internal

oligonucleotides determining the junction. Overlap extension produced a fragment which was digested with *Sfi*I and *Bam*HI and ligated into pα2(I)Δ1. Site-directed mutagenesis was performed essentially as described by Kunkel *et al.* (1987), except that extension reactions were performed in the presence of 1 unit T4 DNA polymerase and 1 μg T4 gene 32 protein (Boehringer, Lewes, UK).

Transcription in vitro
Transcription reactions were carried out as described by Gurevich *et al.* (1987). Recombinant plasmids pα1(III)Δ1, pα1(III)Δ1:(I)CP,C and pα2(I)Δ1, pα2(I)Δ1:(III)CP, A, F, F$^{s–c}$, B$^{s–c}$, C$^{s–c}$ (10 μg) were linearized and transcribed using T3 RNA polymerase, or T7 RNA polymerase (Promega, Southampton, UK) respectively. Reactions (100 μl) were incubated at 37°C for 4 h. Following purification over RNeasy columns (Qiagen, Dorking, UK), RNA was resuspended in 100 μl RNase-free water containing 1 mM DTT and 40 units RNasin (Promega, Southampton, UK).

Translation in vitro
RNA was translated using a rabbit reticulocyte lysate (FlexiLysate, Promega, Southampton) for 2 h at 30°C in the absence of exogenous DTT. The translation reaction (25 μl) contained 17 μl reticulocyte lysate, 1 μl 1 mM amino acids (minus methionine), 0.45 μl 100 mM KCl, 0.25 μl ascorbic acid (5 mg/ml), 15 μCi [L-³⁵S]methionine (Amersham International, Bucks, UK), 1 μl transcribed RNA and 1 μl (~2×10⁵) semi-permeabilized cells (SP-cells) prepared as described by Wilson *et al.* (1995). After translation, *N*-ethylmaleimide was added to a final concentration of 20 mM. SP-cells were isolated by centrifugation in a microfuge at 10000 *g* for 5 min and the pellet resuspended in an appropriate buffer for subsequent enzymic digestion or gel electrophoresis.

Bacterial collagenase digestion
SP-cells were resuspended in 50 mM Tris–HCl pH 7.4, containing 5 mM CaCl₂, 1 mM phenylmethanesulfonyl fluoride (PMSF), 5 mM *N*-ethylmaleimide and 1% (v/v) Triton X-100 and incubated with 3 units collagenase form III (Advance Biofacture, Lynbrook, NJ) and incubated at 37°C for 1 h. The reaction was terminated by the addition of SDS–PAGE sample buffer.

Proteolytic digestion
Isolated SP-cells were resuspended in 0.5% (v/v) acetic acid, 1% (v/v) Triton X-100 and incubated with pepsin (100 μg/ml) for 2 h at 20°C or 16 h at 4°C. The reactions were stopped by neutralization with Tris–base (100 mM). Samples were then digested with a combination of chymotrypsin (250 μg/ml) and trypsin (100 μg/ml) (Sigma, Poole, Dorset, UK) for 2 min at room temperature in the presence of 50 mM Tris–HCl pH 7.4, containing 0.15 M NaCl, 10 mM EDTA. The reactions were stopped by the addition of soy bean trypsin inhibitor (Sigma, Poole, Dorset, UK) to a final concentration of 500 μg/ml and boiling SDS–PAGE loading buffer. Samples were then boiled for 5 min.

Thermal denaturation
Pepsin-treated samples were resuspended in 50 mM Tris–HCl pH 7.4, containing 0.15 M NaCl, 10 mM EDTA, and aliquots placed in a thermal cycler. A stepwise temperature gradient was set up from 31°C to 40°C with the temperature being held for 2 min at 1°C intervals. At the end of each time period the sample was treated with a combination of chymotrypsin and trypsin, as described above.

SDS–PAGE
Samples resuspended in SDS–PAGE loading buffer [0.0625 M Tris–HCl pH 6.8, SDS (2% w/v), glycerol (10% v/v) and Bromophenol Blue] in the presence or absence of 50 mM DTT and boiled for 5 min. SDS–PAGE was performed using the method of Laemmli (1970). After electrophoresis, gels were processed for autoradiography and exposed to Kodak X-Omat AR film, or images quantified by phosphoimage analysis.

J.F.Lees, M.Tasab and **N.J.Bulleid**

References

Bachinger,H.P., Bruckner,P., Timpl,R., Prockop,D.J. and Engel,J. (1980) Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen: role of disulphide bridges and peptide bond isomerization. *Eur. J. Biochem.*, **106**, 619–632.

Bachinger,H.P., Fessler,L.I., Timpl,R. and Fessler,J.H. (1981) Chain assembly intermediate in the biosynthesis of type III procollagen in chick embryo blood vessels. *J. Biol. Chem.*, **256**, 13193–13199.

Brass,A., Kadler,K.E., Thomas,T., Grant,M.E. and Boot-Handford,R.P. (1992) The fibrillar collagens, collagenVII, collagenX and the C1q complement proteins share a similar domain in their C-terminal non-collagenous regions. *FEBS Lett.*, **303**, 126–128.

Bruckner,P. and Prockop,D.J. (1981) Proteolytic enzymes as probes for the triple-helical conformation of procollagen. *Anal. Biochem.*, **110**, 360–368.

Bulleid,N.J., Wilson,R. and Lees,J.F. (1996) Type III procollagen assembly in semi-intact cells: chain association, nucleation and triple helix folding do not require formation of inter-chain disulfide bonds but triple helix nucleation does require hydroxylation. *Biochem. J.*, **317**, 195–202.

Cheah,K.S.E., Grant,M.E. and Jackson,D.S. (1979) Translation of type II procollagen mRNA and hydroxylation of the cell-free products. *Biochem. Biophys. Res. Commun.*, **91**, 1025–1031.

Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, **47**, 45–148.

Dion,A.S. and Myers,J.C. (1987) COOH-terminal propeptides of the major human procollagens: Structural, functional and genetic comparisons. *J. Mol. Biol.*, **193**, 127–143.

Doege,K.J., and Fessler,J.H. (1986) Folding of carboxyl domain and assembly of procollagen I. *J. Biol. Chem.*, **261**, 8924–8935.

Gurevich,V.V., Pokrovskaya,I.D., Obukhova,T.A. and Zozulya,S.A. (1991) Preparative *in vitro* mRNA synthesis using SP6 and T7 polymerases. *Anal. Biochem.*, **195**, 207–213.

Hojima,Y., van der Rest,M. and Prockop,D.J. (1985) Type I procollagen carboxy-terminal proteinase from chick embryo tendons. *J. Biol. Chem.*, **260**, 15996–16003.

Horton,R.M. (1993) PCR protocols: current methods and applications. In White,B.A. (ed.), *Methods in Molecular Biology*. Humana Press Inc., Totowa, NJ, Vol. 15, Chapter 25.

Jiminez,S.A., Bashey,R.I., Benditt,M. and Yanowski,R. (1977) Identification of $\alpha1(I)$ collagen trimer in embryonic chick tendons and calvaria. *Biochem. Biophys. Res. Commun.*, **78**, 1354–1361.

Kadler,K. (1994) Fibril forming collagens. In Sheterline,P. (ed.), *Protein profiles: Extracellular matrix 1*. Academic Press, London, UK.

Kessler,E., Kazuhiko,T., Biniaminov,L., Brusel,M. and Greenspan,D.S. (1996) Bone morphogenic protein-1: The type I procollagen proteinase. *Science*, **271**, 360–362.

Kielty,C.M., Hopkinson,I. and Grant,M.E. (1993) The collagen family: structure, assembly and organization in the extracellular matrix. In Royce,P.M. and Steinmann,B. (eds), *Connective Tissues and its Heritable Disorders: Molecular, Genetic and Medical Aspects*. Wiley-Liss Inc, pp. 103–147.

Kunkel,T.A., Roberts,J.D. and Zakour,R.A. (1987) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol.*, **154**, 367–382.

Laemmli,U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685.

Lees,J.F. and Bulleid,N.J. (1994) The role of cysteine residues in the folding and association of the COOH-terminal propeptide of types I and III procollagen. *J. Biol. Chem.*, **269**, 24354–24360.

Moro,L. and Smith,B.D. (1977) Identification of $\alpha1(I)$ collagen trimer and normal type I collagen in a polyoma virus induced mouse tumour. *Arch. Biochem. Biophys.*, **182**, 33–41.

Olsen,B.R., Hoffman,H.-P. and Prockop,D.J. (1976) Interchain disulfide bonds at the COOH-terminal end of procollagen synthesized by matrix-free cells from chick embryonic tendon and cartilage. *Arch. Biochem. Biophys.*, **175**, 341–350.

Prockop,D.J. and Kivirikko,K.I. (1995) Collagens: Molecular biology, diseases and potential for therapy. *Annu. Rev. Biochem.*, **64**, 403–434.

Prockop,D.J., Olsen,A., Kontusaari,S., Hyland,J., Ala-Kokko,L., Vasan,N.S., Barton,E., Buck,S., Harrison,K. and Brent,R.L. (1990) Mutations in human procollagen genes. *Ann. N Y Acad. Sci.*, **580**, 330–339.

Schofield,D.J., Uitto,J. and Prockop,D.J. (1974) Formation of interchain disulfide bonds and helical structure during biosynthesis of procollagen by embryonic tendon cells. *Biochemistry*, **13**, 1801–1806.

Van der Rest,M. and Garrone,R., (1991) Collagen family of proteins. *FASEB J.*, **5**, 2814–2823.

Wilson,R., Oliver,J., Brookman,J.L., High,S. and Bulleid,N.J. (1995) Development of a semi-permeabilized cell system to study the translocation, folding, assembly and transport of secretory proteins. *Biochem J.*, **307**, 679–687.