Novel Amino-terminal Propeptide Configuration in a Fibrillar Procollagen Undergoing Alternative Splicing*

Jean-Yves Exposito‡, Marina D’Alessio§, and Francesco Ramirez
From the Brookdale Center for Molecular Biology, Mt. Sinai School of Medicine, New York, New York 10029

We isolated overlapping cDNAs from embryonic libraries of the sea urchin *Strongylocentrotus purpuratus* coding for a fibrillar procollegen (2α chain) with a predicted molecular mass of about 320 kDa. The deduced primary structure of the echinoid chain consists of a 265-amino acid carboxyl-propeptide, a triple helical domain made of 337 uninterrupted Gly-X-Y repeats, and an unusually long amino-propeptide. Aside from a 10-cysteine globular region, a collagenous sequence, and a nonhelical segment, this protein domain includes a novel 4-cysteine motif repeated several times. Interestingly, preliminary evidence indicates that different combinations of the 4-cysteine repeats are encoded by alternatively spliced transcripts. Irrespective of this, the sea urchin 2α procollagen chain represents the longest fibrillar molecule identified to date by cDNA cloning experiments in both vertebrate and invertebrate organisms.

In higher vertebrates, five distinct collagen trimers (types I, II, III, V, and XI) participate in the formation of morphologically similar supermolecular aggregates, the quarter-staggered fibrils (for a recent review, see Ref. 1). The precursor procollagen subunits of fibrillar collagens exhibit the same overall structure consisting of a central triple helical domain flanked by carboxyl-terminal and amino-terminal propeptides. Unlike the first two domains, amino-terminal propeptides of various procollagen chains differ greatly in length and composition. Accordingly, three distinct amino-terminal propeptide architectures are recognized (2). The first consists of a globular region that harbors 10 similarly spaced cysteinyl residues, a collagogenous sequence that may or may not be discontinuous, and a nonhelical segment that contains the N-proteinase cleavage site (Structure I). All but the first subdomain are present in the second amino-terminal propeptide configuration (Structure II). In the third type of configuration (Structure III), the 10-cysteine subdomain is replaced by a long globular region divided by a 3-cysteine cluster into an upstream slightly basic segment and a downstream highly acidic sequence. Furthermore, one of the fibrillar procollagen chains, pro-α1(II), exists in both the first and second configuration because of alternative splicing of the sequence coding for the 10-cysteine globular region (3).

In contrast to vertebrates and despite substantial morphological data, very little is known about the primary structure of fibrillar procollagen molecules in invertebrates. The sole exceptions are some partial sequences from cDNAs of the fresh water sponge *Ephydatia mulleri* and the Mediterranean sea urchin *Paracentrotus lividus* (4-6). All or incomplete, these data revealed a close evolutionary kinship between the vertebrate and invertebrate proteins. They also documented the functional contribution of several phylogenetically retained structures of the proteins to metazoan fibrillogenesis.

In addition to serving as supportive elements, collagens are intimately involved in a variety of physiological processes and cellular activities (7). Such a dual role is reflected in the complexity and diversification of the molecular circuitries that modulate collagen production during development and in the adult organism (8). The sea urchin represents an instructive and simple organism for studying collagen function and regulation during early animal embryogenesis. Sea urchin collagens have been implicated in different morphogenetic programs of the developing embryo, such as gastrulation and spiculogenesis (9-14). For example, expression of a nonfibrillar collagen gene (Spcoil) by the differentiating mesenchyme cells of the sea urchin *Strongylocentrotus purpuratus* has recently been shown to govern the point of differentiation at which these cells initiate biomineralization (14).

Because of our long-standing interest in collagen evolution, we have recently completed the determination of the sequences of two sea urchin fibrillar collagens, termed 1α and 2α chain (5, 6). Here, we present the data pertaining to COLP2α, the gene coding for the 2α procollagen chain of *S. purpuratus*.

MATERIALS AND METHODS

Embryo Cultures and Nucleic Acids Purification and Analysis—Collection of *S. purpuratus* gametes, *in vitro* fertilization, and embryo cultures were performed according to the standard protocol (15). Genomic DNA was purified from sperm as previously described (16). Total RNA, prepared according to a published protocol (17), was eluted twice through an oligo(dT)-cellulose column (Boehringer Mannheim). For Northern blot analysis, 1 μg of poly(A)+ RNA was fractionated through a 1% agarose gel containing 2.2 M formaldehyde, transferred onto a nitrocellulose filter (Millipore), and hybridized to a 500-bp EcoRI/KpnI probe corresponding to the 5′-foremost segment of COLP2a (Fig. 1).

cDNA Cloning and Sequencing—Approximately 5 μg of late gastrula stage poly(A)+ RNA was utilized as a template to generate two embryonic cDNA libraries in the λgt10 and λgt11 vector using oligo(dT) and random primers and following the recommendations...
Cloning of 2a Procollagen cDNAs—A 2.5-kb cDNA (2) was initially isolated from an *S. purpuratus* embryonic library by cross-hybridization with a probe (Uni 13) previously shown to code for the carboxyl-terminal propeptide domain of the *P. lividus* 2a procollagen (6) (Fig. 1). Comparison of the deduced amino acid sequences of 12 and Uni 13 showed a substantial degree of homology extending to the highly divergent carboxy-telopeptide (Fig. 2). In addition, the two carboxyl-terminal propeptides display in identical positions a potential Asn-linked glycosylation site and the 7 cysteines which, in vertebrate chains, are involved in intracellular assembly of the procollagen trimer (Fig. 2). Based on these data, we concluded that the *S. purpuratus* clone represents the counterpart of the *P. lividus* 2a collagen gene product. Additional library screenings yielded eight overlapping clones covering a 9594-bp-long open reading frame (Fig. 1). As discussed more extensively below, some of the cDNAs provided evidence that COLP2a transcripts undergo alternative splicing.

Structure of 2a Procollagen—The conceptual amino acid translation of the sea urchin cDNAs revealed that this polypeptide is unique among vertebrate and invertebrate fibrillar procollagens, for its predicted molecular mass is ~320 kDa. Moreover, nearly 60% of the chain is accounted for by an unusually long amino-terminal propeptide domain characterized by a novel subdomain positioned between the 10-cysteine globular region and the collagenous sequence (Fig. 1).

The first subdomain of the sea urchin amino-terminal propeptide, the 10-cysteine globular region, begins 38 residues after the start site of translation immediately following a characteristic signal peptide sequence (Fig. 3). A remarkable similarity can be observed when the cysteine cluster of the sea urchin chain is aligned to the same region of the human procollagen chains (Fig. 4) (3). First, the spacing between the cysteinyl residues is maintained across collagens, with the exception of the interval between cysteines 6 and 7. This particular spacing is also different in the pro-al(II1) collagen chain (Fig. 4) (22). Second, and like the human chains, the vertebrate subdomain retains several invariant residues that are also conserved in the analogous cysteine-rich motif of thrombospondin (3, 23). Third, the echinoid and human sequences are nearly identical around the two consecutive cysteines likely to be engaged in interchain bonding.

Following the 10th cysteine of the globular domain is the new subdomain, which is made of a 4-cysteine motif repeated 12 times (Fig. 3). With the exception of the first repeat, whose estimated pI is 8.3, the other 11 repeats are substantially included in a commercial kit (Amersham Corp.). One of the libraries was initially screened under relaxed conditions of hybridization (18) with a 660-bp BglII/EcoRI fragment of clone Uni 13, which codes for the carboxyl-terminal propeptide domain of the *P. lividus* 2a procollagen (6). Additional upstream screenings of both libraries were performed under stringent hybridization conditions utilizing appropriate S. purpuratus cDNA segments (19). Multiple overlapping sequencing of both DNA strands was achieved by sequencing small cDNA subclones, by generating progressive deletions with the exonuclease III/mungbean nuclease (U. S. Biochemical Corp.) method (20), and by employing synthetic oligonucleotides. Nucleotide sequence was determined with a modified protocol (4) of Sanger et al. (21) using the Sequenase enzyme (U. S. Biochemical Corp.). Sequence analysis was aided by the computer program MULTALIN. Oligonucleotides were synthesized by an Applied Biosystems model 380 synthesizer.

**RESULTS**

**Fig. 1.** Partial restriction map of 2a procollagen cDNAs. Overlapping clones are depicted below a diagrammatic representation of the protein. In the amino-terminal propeptide domain (N-propeptide), the 12 repeats (R1-R12) are illustrated relative to the position of the encoding cDNAs. The sequences of repeats absent in clones F6 and F9 are indicated by the dotted lines. B, BamH1; E, EcoRI; K, KpnI; F, PstI; S, SacI; X, XbaI. The wavy line signifies the signal peptide. On the right side is a Northern blot of COLP2a showing multiple hybridizing bands whose approximate sizes are indicated to the left of the autoradiogram. C-propeptide, carboxy-terminal propeptide.
The multiple amino-terminal propeptide subdomains are designated by the open triangles. In the cysteine-rich globular region, cysteinyl residues are boxed (see also Fig. 6). In the repeated subdomain, the boundaries and identity of each repeat are indicated by the horizontal arrows and numbers, respectively (see also Fig. 6). In the collagenous sequence, Gly-X-Y triplets are underlined. Amino acids are italicized with theoretical PI values ranging from 3.8 to 4.2. A putative cell attachment sequence (Arg-Gly-Asp) and a potential glycosylation site are noted in repeats respectively. The homology between the triple-helical domain by the horizontal arrows and numbers, respectively (see also Fig. 6). In the collagenous sequence of the amino-terminal propeptide (Fig. 3). From this alignment, the following consensus sequence and known peptide motifs.}

**Fig. 3.** Amino-terminal propeptide sequence. Amino acids are numbered on the right from the start site of translation; they extend to the region immediately before the beginning of the triple-helical domain (see Fig. 6). The boundaries of the four amino-terminal propeptide subdomains are designated by the open triangles. In the cysteine-rich globular region, cysteinyl residues are boxed (see also Fig. 4). In the repeated subdomain, the boundaries and identity of each repeat are indicated by the horizontal arrows and numbers, respectively (see also Fig. 6). In the collagenous sequence, Gly-X-Y triplets are underlined. Amino acids are italicized with theoretical PI values ranging from 3.8 to 4.2. A putative cell attachment sequence (Arg-Gly-Asp) and a potential glycosylation site are noted in repeats respectively. The homology between the triple-helical domain by the horizontal arrows and numbers, respectively (see also Fig. 6). In the collagenous sequence of the amino-terminal propeptide (Fig. 3). From this alignment, the following consensus sequence and known peptide motifs.

**Fig. 4.** Comparison of vertebrate and invertebrate 10-cysteine globular subdomains. Sequence alignment around the 10-cysteine subdomain (see Fig. 6) and human a1(I), a1(II), a1(III), and a2(V) procollagen is shown. Cysteines (numbered in the amino to carboxyl direction) and invariant residues cited in the text are boxed. Only invariant residues were identified in this comparative analysis. Note that the ordering of the chain in the cross-alignment is arbitrary.

**Fig. 3.** Amino-terminal propeptide sequence. Amino acids are numbered on the right from the start site of translation; they extend to the region immediately before the beginning of the triple-helical domain (see Fig. 6). The boundaries of the four amino-terminal propeptide subdomains are designated by the open triangles. In the cysteine-rich globular region, cysteinyl residues are boxed (see also Fig. 4). In the repeated subdomain, the boundaries and identity of each repeat are indicated by the horizontal arrows and numbers, respectively (see also Fig. 6). In the collagenous sequence, Gly-X-Y triplets are underlined. Amino acids are italicized with theoretical PI values ranging from 3.8 to 4.2. A putative cell attachment sequence (Arg-Gly-Asp) (24) and a potential glycosylation site are noted in repeats 7 and 5, respectively. The homology between the 12 repeats can be readily appreciated when portions of their sequences are aligned (Fig. 5). From this alignment, the following consensus sequence can be derived: X_{260}GX_{260}LWX_{375}GXX_{400}CX_{400}CX_{400}L/FX_{375}CX_{400}CX_{400}L/FX_{375}CX_{400}CX_{400}L (where numbers in parentheses signify an average number of residues). A computer-aided search failed to identify appreciable homology between this consensus sequence and known peptide motifs.

Twenty-four Gly-X-Y triplets constitute the collagenous sequence of the amino-terminal propeptide (Fig. 3). They contain a small interruption likely to render the upstream set of four triplets too short to participate in triple helical assembly. This subdomain is separated from the main tripeptide domain by 21 residues containing a potential prolylthiopeptide signal (Ala-Gln) (13). Should the amino-terminal propeptide be cleaved, then a Lys-cross-linking site would be located in the amino-teleopeptide of the mature a-chain (Fig. 3).

The triple helical domain of the sea urchin 2a procollagen is made of 337 uninterrupted Gly-X-Y repeats. As previously
shown for other invertebrate and lower vertebrate chains and unlike triple helical domains of higher vertebrates (4–6, 18), numerous Gly-X-Y triplets display a glycine residue either at position X or Y (Fig. 6). In addition, the 2α chain contains several putative cell attachment sequences but lacks the cross-linking sites normally found at both ends of the triple-helical domain (Fig. 6) (26).

The two major structural features of the carboxyl-terminal propeptide domain were already discussed in the previous section. In completing this description, a potential C-proteinase cleavage site (Arg-Asp) is predicted to reside 25 residues from the end of the triple helical domain (Fig. 2). If this assumption were correct, the mature 2α chain would contain a second Lys-mediated cross-linking site in the carboxy-terminal domain (Fig. 2).

Evidence for Alternative Splicing of COLP2α—Sequence analysis of several amino-terminal propeptide-encoding cDNAs suggested that COLP2α transcripts undergo alternative splicing. To be precise, clone F6 spans the 5′-untranslated region to the end of repeat 11 and precisely lacks the sequence of repeats 6–8. Clone F9 begins at about the same position and ends in the second third of repeat 8. The sequence of repeats 2–5 is absent in this cDNA. Clone f4 harbors only sequence coding for the repeated subdomain, from the end of repeat 3 to the very beginning of repeat 10. Finally, clone f3 begins within repeat 9, includes repeats 10–12, and ends in the first third of the triple helical domain (Fig. 1). Collectively, the four overlapping clones cover 5769 bp of contiguous sequence coding for the 2α amino-terminal propeptide (Fig. 3).

Independent evidence for COLP2α alternative splicing was obtained by Northern blot hybridization to a probe specific for the region immediately upstream of the repeated subdomain of the amino-terminal propeptide. This identified at least three major hybridizing bands that range in size from approximately 8 to 10 kb (Fig. 1). Furthermore, preliminary analysis of cDNA amplified by the polymerase chain reaction technique strongly suggests the existence of additional alternatively spliced transcripts coding for distinct combinations of the amino-terminal propeptide repeats (data not shown). We believe that some of the background seen in the Northern blot of Fig. 1 might be caused by the hybridization of these alternative and plausibly minor transcripts. Experiments in progress are elucidating the exact number, relative representation, and composition of all COLP2α transcripts, along with the maximum length of the repeated subdomain.

DISCUSSION

The complete primary structure of two invertebrate fibrillar procollagens, 1α and 2α chains, have been deduced from sequences of overlapping cDNA clones. The sea urchin 1α procollagen comprises 1414 amino acids and is evolutionarily related to the vertebrate pro-α2(I) collagen (27). In contrast, the hypothetical structure of the 2α chain is unique among invertebrate and vertebrate fibrillar procollagen molecules. The primary differences are the length and composition of the amino-terminal propeptide domain, which exhibits a novel configuration, Structure IV. In addition to the three characteristic subdomains of Structure I, Structure IV contains a very large subdomain consisting of a novel peptide module repeated several times. Moreover, the mesenchyme cells of
late gastrula sea urchin embryos appear to produce alternatively spliced COLP2α transcripts theoretically encoding isoforms with amino-terminal propeptides of different length.

This is the second case of a fibrillar collagen gene whose amino-terminal propeptide-coding domain undergoes alternative splicing. In the vertebrate pro-α1(II) collagen, the 10-cysteine globular subdomain coding exon is in fact alternatively spliced in a developmentally regulated manner. Such a phenomenon was first recognized in human chondrocytes by Ryan and Sandell (33) and later confirmed in Xenopus and chicken embryos by other investigators (18, 28). The phylogenetic retention of the developmental pattern of alternative splicing has been interpreted as suggesting functionally distinct roles of the resulting pro-α1(II) products during vertebrate embryogenesis (18). It will be of interest to determine whether alternative splicing of COLP2α is also developmentally regulated and, more importantly, to understand what is the significance, if any, of producing different size amino-terminal propeptides.

The contribution of collagen molecules to sea urchin development is well established. In vivo studies have documented the importance of properly formed collagen aggregates for the progression of the gastrulation process (9, 10). In vitro experiments have emphasized the role of the collagen substrate in promoting the biomineralization program of cultured cells (11–14). Our structural data are consistent with the proteolytical removal of the long amino-terminal propeptide of the 2α chain. It is, therefore, tempting to speculate that the 2α procollagen may yield two polypeptides nearly identical in size and serving distinct morphogenetic functions. Such a situation would be analogous to the postulated morphogenetic functions assigned to cleaved propeptides of vertebrate procollagens in cartilaginous matrices (29, 30). The availability of COLP2α clones provides a means to test this hypothesis, in that specific antibodies could be employed in interference experiments on micromers differentiating in culture (14).

The three S. purpuratus collagen genes hitherto identified by cDNA cloning experiments are all expressed in the same cell lineages, but at different ontological times (14). It could be argued that a common regulatory program governs critical changes in matrix composition during the differentiation of primary and secondary mesenchyme cells. Conversely, each of these genes is likely to contain distinct cis-acting elements responsible for timely onset of transcription. Experiments currently in progress are characterizing the regulatory sequences of the three sea urchin collagen genes. These studies promise to elucidate some of the mechanisms and factors controlling collagen gene expression in the developing animal embryo.

Acknowledgments—We thank James Andriotakis for excellent technical assistance, Roseann Lingeza for typing the manuscript, and Dr. Leslie Pick for many helpful suggestions.

REFERENCES

1. van der Rest, M. and Garrouste, H. (1991) J. Biol. Chem. 266, 1214-1223
2. Lelievre, H. and D’Alessio, M. (1991) J. Biol. Chem. 266, 1214-1223
3. Ryan, M. and Sandell, L. (1990) J. Biol. Chem. 265, 10334-10339
4. Exposito, J. V., and Garrouste, H. (1990) J. Biol. Chem. 265, 10334-10339
5. Suzuki, H. R., Ryan, M., Bieker, J. J., and Gambino, R. (1991) J. Biol. Chem. 266, 2646-2650
6. H. R. Suzuki, M. D’Alessio, J. Y. Exposito, R. Gambino, F. Ramirez, and M. Solursh, manuscript in preparation.

H. R. Suzuki, M. D’Alessio, J. Y. Exposito, R. Gambino, F. Ramirez, and M. Solursh, manuscript in preparation.