Structure and developmental expression of a sea urchin fibrillar collagen gene

(gene evolution/mesenchyme lineage/invertebrate collagens/skeletogenesis)

MARINA D'ALESSIO*, FRANCESCO RAMIREZ†, HIROAKI R. SUZUKI‡, MICHAEL SOLURSH‡, AND ROBERTO GAMBINÒ§

*Department of Microbiology and Immunology, Morse Institute of Molecular Genetics, State University of New York-Health Science Center at Brooklyn, Brooklyn, NY 11203; †Department of Biology, University of Iowa, Iowa City, IA 52242; and ‡Istituto di Biologia dello Sviluppo, Consiglio Nazionale delle Ricerche, Palermo, 90123, Italy

Communicated by Elizabeth D. Hay, August 31, 1989

Abstract We have isolated and characterized cDNA and genomic clones that specify a Paracentrotus lividus procollagen chain. The cDNAs code for 160 uninterrupted Gly-Xaa-Yaa triplets and a 252-amino acid carboxyl propeptide. Analysis of the deduced amino acid sequences indicated that the sea urchin polypeptide exhibits structural features that are characteristic of the fibril-forming class of collagen molecules. Partial characterization of two genomic recombinants revealed that the 3' end of the echinoid gene displays a complex organization that closely resembles that of a prototypical vertebrate fibrillar collagen gene. In situ and Northern (RNA) blot hybridizations established the size, time of appearance, and tissue distribution of the collagen transcripts in the developing sea urchin embryo. Collagen mRNA, ~6 kilobases in size, is first detected in the forming primary mesenchyme cells of late blastulae where it progressively accumulates until the free-swimming/feeding pluteus larval stage. Interestingly, collagen transcripts are also detected in the forming secondary mesenchyme cells of late gastrulae, and by the prism stage, their derivatives appear to be the most intensively labeled cells.

Collagen molecules composed of subunits that display the characteristic Gly-Xaa-Yaa repeated sequence are found in the intercellular spaces of all Metazoa (1). In vertebrates, different collagen types are segregated into distinct groups according to the structural features of the proteins and the organizational motifs of the genes (2, 3). The largest of these groups includes the fibril-forming collagens (types I–III, V, and XI) (2). These trimeric molecules similarly arise from procollagen precursors which, after assembly and secretion, are converted to collagen by proteolytic removal of the terminal propeptides (2). As in all vertebrate collagens, the organization of the fibrillar genes is extremely complex in that the number of introns ranges between 49 and 53 (3). Moreover, their triple helical domain-coding exons appear to relate to a basic 54-base pair (bp) unit encoding six Gly-Xaa-Yaa repeats (3). Comparable multi-exon structures have not been observed in the most thoroughly studied invertebrate collagen genes—namely, those of Drosophila melanogaster and Caenorhabditis elegans (4, 5).

In addition to serving as supportive elements, the collagens are intimately involved in determining the spatial and temporal diversity of extracellular matrices and, consequently, in influencing several morphogenetic programs and cellular activities (6). The sea urchin represents an instructive and relatively simple experimental model for studying control of collagen expression during early embryogenesis (7). Collagen biosynthesis, which begins at blastula and increases several-fold thereafter, and disposition of crosslinked collagen mol-

cules in the extracellular matrix of the embryo are essential for both processes of gastrulation and spiculogenesis (8–10). At least three immunologically distinct collagen types have been identified in sea urchin embryos by using antibodies to vertebrate fibrillar (types I and III) and basement membrane (type IV) collagens (9, 11). The subunit of one of these collagen types is thought to be encoded in Strongylocentrotus purpuratus by the gene that Venkatesan et al. (12) isolated by cross-hybridization to a mouse type IV collagen probe. They based this conclusion on the identification of a bona fide 212-bp exon coding for a discontinuous series of Gly-Xaa-Yaa repeats. More recently, using a C. elegans collagen probe Saïta et al. (13) have similarly isolated a Paracentrotus lividus genomic fragment that contains five exons potentially encoding 97 uninterrupted Gly-Xaa-Yaa triplets. Although the S. purpuratus and P. lividus collagen-coding exons hybridize to developmentally regulated mRNAs, no conclusive evidence has yet been presented that the genes are actually transcribed because their identity was exclusively based on limited genomic analyses (12, 13).

To establish a solid experimental ground for studying collagen expression during sea urchin embryogenesis, as well as to further define the evolution of the collagen gene family, in the present studies we isolated and characterized cDNA and genomic clones that specify a developmentally regulated P. lividus fibrillar collagen.

MATERIALS AND METHODS

Embryo Cultures and Nucleic Acid Purification. P. lividus embryos were cultured in sterile seawater with antibiotics (13). Sperm DNA and poly(A)⁺ RNA were prepared as described (13). Similarly, conditions for Northern (RNA) blot hybridization were essentially as detailed (13).

Cloning Experiments and DNA Sequencing. Prism-stage embryonic RNA was used to generate a Agt10 cDNA library (14), which was screened with either a 2.6-kilobase (kb) genomic probe, previously shown to contain five putative collagen-coding exons, or with two synthetic 21-mers, corresponding to the sequences of the second and fifth exons of the 2.6-kb fragment (13). The genomic library was constructed by inserting into the BamH1 site of the ADASH phage vector (Stratagene) size-fractionated P. lividus DNA partially digested with Sau3A (13). High-stringency hybridization conditions for genomic and cDNA probes, as well as for synthetic oligonucleotides, have been detailed (13, 15). Nucleotide sequencing was performed using the dideoxynu-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

1To whom reprint requests should be addressed: Brookdale Center for Molecular Biology, The Mount Sinai Medical Center, One Gustave Levy Place, New York, NY 10029.

2The sequence reported in this paper has been deposited in the GenBank data base (accession no. M25282).
cloning the cDNA encoding the 3'-untranslated region of the 1a collagen gene (COLL1a). Results of these investigations constitute the subject of this report.

**Structural Features of the Collagen Molecule.** The 1a collagen cDNAs, Uni 11 and Uni 54, were isolated using the 2.6-kb probe and the exon-specific oligonucleotides, respectively. They share 2.5 kb of identical sequences corresponding to the 3'-untranslated region of the message and include a short poly(A) tail (Fig. 1). The 5'-overlapping chain-termination procedure on double-stranded DNA (16). Sequencing of both DNA strands was achieved for cDNAs by generating progressively overlapping deletions with the exonuclease III/mung bean nuclease method (17) and for genomic clones by using synthetic oligonucleotides that correspond to appropriate cDNA sequences. Sequences were analyzed using the computer program MULTALIN developed by Corpet (18).

**In Situ Hybridization.** Embryos were fixed with 1% glutaraldehyde in saline, dehydrated, and embedded in paraffin according to Angerer et al. (19). A modification of this protocol was used for preparation, treatment, hybridization, and washing of the embryo sections (20). For in vitro transcription, the 750-bp Bgl II–EcoRI fragment of clone Uni 54 (Fig. 1) was subcloned into the BamHI/EcoRI sites of the transcription vector pT7/T3-18 (Bethesda Research Laboratories). Labeled single-stranded RNA probes were synthesized on 2 μg of linear template as described (20).

**RESULTS**

**Cloning of P. lividus Collagen cDNAs.** Previous investigations have established that in *P. lividus* a 2.6-kb genomic fragment, which contains several putative collagen-coding exons, hybridizes to an embryonic 6-kb transcript that reaches maximal accumulation at prism stage (13). Hence, a prism-stage embryonic cDNA library was screened with the aforementioned 2.6-kb probe and, as a result, several positive clones were identified. Upon structural analysis, the recombinants were found to encode two distinct collagen-like polypeptides, arbitrarily referred to as 1a and 2a chains. Despite the stringency of the hybridization conditions used, neither of the two sets of clones was found to correspond to the genomic probe. An identical result was obtained after screening the cDNA library with the exon-specific oligonucleotides. Notwithstanding this finding, the implications of which will be discussed later, we proceeded to extensively

![Diagram](image1.png)

**Fig. 1.** (A) Partial restriction map of the 1a chain encoding cDNAs showing the relationship to the major domains of the protein, namely helical domain (α) and carboxyl propeptide. The arrow indicates the relative position of the putative carboxyl cleavage site for the proteinase, whereas the dotted line signifies the mRNA 3'-untranslated region with the poly(A) tail (An). (B) Northern blot of poly(A)+ RNA from gonads (lane 1), unfertilized eggs (lane 2), 32-cell (lane 3), morula (lane 4), blastula (lane 5), gastrula (lane 6), prism (lane 7), and pluteus (lane 8) stages. The filter was hybridized to the 750-bp Bgl II–EcoRI subclone of Uni 54 (6 kb) and, as an internal control, to pep-1, a membrane protein encoding cDNA (1.4 kb), which displays a different pattern of developmental expression (31). Sizes of the transcripts were determined by comparison to the relative migration of rRNA markers.

![Diagram](image2.png)

**Fig. 2.** Nucleotide and deduced amino acid (in one-letter code) sequences of COLL1a. Numbering is from the first nucleotide of Uni 54. The vertical bars signify the end of the helical domain, and the asterisk signifies the termination codon. The putative carboxyl cleavage site for proteinase (R-D) is highlighted by the arrow. The amino acid residues discussed in the text are boxed.
sequences code for two distinct domains of the polypeptide—namely, a 252-amino acid, cysteine-rich carboxy propeptide and an uninterrupted 478-amino acid helical domain (Fig. 2). Examination of the deduced amino acid sequences revealed a variety of structural features characteristic of fibril-forming collagens (2). In the helical domain, they include the tetrapeptide Lys-Gly-His-Thr, likely to represent a substrate for lysine-mediated crosslinking of the fibrils, and at its carboxy terminus several Gly-Pro-Pro triplets conceivably involved in the stabilization of the trimer (Fig. 2) (2). Assuming that our choice of the putative carboxy cleavage site for proteinase is correct, the 26-amino acid carboxy telopeptide contains a third structural feature—namely, the crosslinking lysine residue (Fig. 2) (2). The analogy with vertebrate fibrillar collagens is also apparent when a selected portion of the sea urchin carboxy propeptide is aligned with those of three representative human chains (Fig. 3) (22). In this region there is, in fact, a remarkable conservation in the spatial architecture of and, to some extent, the sequence context around the cysteinyl residues involved in the association and alignment of individual procollagen chains before triple helix formation (Fig. 3) (2). It has been suggested that the variable number of cysteines in different carboxy propeptides, 7 or 8, correlates with the formation by individual chains of either heterotrimers or both homo- and heterotrimers, respectively (23). Because the echinoid chain lacks one of the variable cysteinyl residues, notably the second cysteine (Fig. 3), it could be argued that this procollagen molecule represents the subunit of a heterotrimetric collagen type. Finally, although an additional highly conserved vertebrate element, the N-linked carbohydrate-attachment site (Asn-Xaa-Ser-Thr), is missing at the comparable position of the echinoid carboxy propeptide, an alternative potential glycosylation site (Asn-Ser-Thr) is noted nine residues to the right of the fifth cysteinyl residue (Fig. 3).

Gene Organization. To elucidate the structure of COL1A we isolated and partially characterized two overlapping genomic clones, Nap 111 and Nap 131, which cover the 3' end of the gene; this led to the identification of 19 exons the organization of which was found to closely resemble that of a prototypical vertebrate fibrillar collagen gene (Fig. 4). This evolutionary kinship is more evident in the triple helical coding domain the exons of which clearly adhere to the vertebrate 54-bp motif (3) (Fig. 4).

In vertebrates the carboxy propeptide-coding sequences are consistently distributed among four exons (exons 49–52), two of which (exons 51 and 52) are invariably in size (3). Exon 49 represents the 'junction' exon because, albeit variable in size, it always codes for the carboxy terminus of the triple helical domain, the carboxy telopeptide, and the amino terminus of the carboxy propeptide (3). Such an arrangement of sequences results in the same relative distribution of the cysteine codons among the four carboxy propeptide-coding exons of different collagen genes (3). The sea urchin gene appears to share the latter feature, as well as the variable length of the coding sequences in the last exon. An important difference was, however, noted for the echinoid junction exon that contains mostly triple helical domain-coding sequences and a very short noncollagenous component made by most, but not all, of the carboxy telopeptide sequences (Fig. 4). As a consequence, the carboxy propeptide-coding domain of the echinoid gene displays five rather than four exons, and the putative carboxy proteinase cleavage site is not encoded by the junction exon. From the aforementioned distribution of the cysteine codons in vertebrates, we propose that the first two carboxy propeptide exons of the sea urchin gene are evolutionarily related to the vertebrate junction exon (Fig. 4).

Developmental Expression. To characterize the size and time of appearance of COL1A transcripts, the cDNA 3'-
DISCUSSION

In this report we have described the isolation and characterization of a P. lividus collagen gene and provided information pertaining to both collagen evolution in Deuterostomia and collagen expression during sea urchin embryogenesis.

The P. lividus polypeptide is structurally related to vertebrate collagens, notably to the fibril-forming class of molecules (2). Aside from the continuous stretch of repeated Gly-Xaa-Yaa triplets, this conclusion is supported by the presence of several functional/structural features the phylogenetic retention of which illustrates their unreplaceable contribution to the molecular architecture and metabolism of metazoan fibrillar collagens. In contrast to the nearly intronless insect and nematode genes (5, 6), the echinoid collagen gene exhibits an organization almost identical to that of a vertebrate gene. This finding, in turn, demonstrates that the establishment of the collagen multi-exon structure predates vertebrate radiation and suggests fundamental evolutionary differences between protostomial and deuterostomial genomes.

The mRNA encoded by COLLa1 is developmentally regulated and accumulates only in the mesenchyme cells of the...
embryo. Collagen gene expression in primary mesenchyme cells is in agreement with the recognized developmental program of this cell lineage, which distinctively results in a major identifiable product, the larval skeleton (7). Collagen accumulation in secondary mesenchyme cells might conceivably indicate additional ontogenic roles for this protein. For example, collagen may provide extracellular signals to the interacting primary and secondary mesenchyme cells during the invagination of the vegetal pole and the formation of the archenteron (24). In this respect, a parallel can be drawn between COLLa expression in both forming primary and secondary mesenchyme cells and type I collagen production during epithelial-mesenchymal transitions in vertebrate embryos (25). Moreover, these and unpublished data on the 2α collagen gene indicate that distinct developmental programs, such as gastrulation and skeletogenesis, are associated with the biosynthesis of different collagens. Such a finding is consonant with previous studies that strongly suggested a genetic heterogeneity of functionally distinct collagen types in the developing sea urchin embryo (8–11, 30).

A final point concerns some conflicting results regarding collagen expression in the developing sea urchin embryo (12, 26, 27). Briefly, Vankatesan et al. (12) originally reported that the 212-bp exon probe recognizes a specific 9-kb transcript that first appears at morula stage, rises significantly in blastulae, and later decreases in proportion to total RNA. Subsequently, the same group established that maximal collagen expression is in the mesenchyme cells of gastrulae (27). Moreover, an antibody raised against a synthetic peptide, which corresponds to the poorly antigenic Gly-Xaa-Yaa discontinuity encoded by the 212-bp exon, was found to weakly interact with an epitope present on the spicule sheet of the whole embryo (27). At about the same time, Nemer and Harlow (26) reported that the 212-bp exon probe hybridizes to two distinct transcripts: (i) 9 kb in length, first seen in blastulae and displaying maximal accumulation in gastrulae, and (ii) 7 kb in size, first detected in gastrulae and reaching maximal concentration in plutei. Our finding that the P. lividus 2.6-kb probe did very effectively cross-hybridize to cDNAs from two distinct collagen transcripts strongly supports the latter observation. Both studies therefore suggest a reconsideration of previous work (12, 27) in that they document the importance of probe specificity when analyzing closely related gene products, such as the collagens.

Irrespective of these considerations, the predominant, if not exclusive, expression of COLLa in the micromere-primary mesenchyme cell lineage provides an attractive model for understanding regulation of transcription in a system highly amenable to experimental manipulations (7). Such a goal has been reported for two other well-characterized cell lineage-specific genes—namely, the cytoskeletal actin CyIIa and the spicule glycoprotein SM50 (28, 29).

This work is dedicated to the memory of Professor Alberto Monroy. We are indebted to Drs. G. Giudice, M. Kurkinen, K. Joho, and G. Spinelli for many helpful suggestions and Dr. M. Di Carlo for providing us with the clone bep-1. We are thankful to Mr. A. Oliva and J. Prince for their excellent technical assistance and Ms. R. Lingez for the preparation and typing of the manuscript. This work was supported by grants from the National Institutes of Health (AR-38648 and HD-18577) and Progetto Finalizzato Regione Siciliana. M.D. is on leave of absence from the International Institute of Genetics and Biophysics, Centro Nazionale delle Ricerche, Naples.