Complete Primary Structure of a Sea Urchin Type IV Collagen \( \alpha \) Chain and Analysis of the 5' End of Its Gene*

(Received for publication, July 31, 1992)

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

We isolated several overlapping cDNAs from Strongylocentrotus purpuratus coding for a nonfibrillar collagen chain structurally homologous to the vertebrate type IV collagen chains and arbitrarily termed 3\( \alpha \) chain. The deduced amino acid sequence of the sea urchin polypeptide includes a 28-residue signal peptide, a 14-residue amino-terminal non-collagenous segment, a triple-helical domain of 1390 residues containing 23 imperfections, and a 226-residue carboxyl-terminal non-collagenous region. Comparison of the sea urchin amino- and carboxyl-terminal non-collagenous domains with those of the vertebrate type IV collagen chains indicated a high level of sequence identity to the \( \alpha I(IV) \) and \( \alpha 5(IV) \) chains. This evolutionary relationship was further strengthened by the analysis of the genomic organization of the 5' portion of the sea urchin gene, which also provided the composition of some of the upstream sequences. In addition, this work demonstrated that our gene product is identical to that encoded by the partial cDNA clone recently isolated by others (Wessel, G. M., Etkin, M., and Benson, S. (1991) Dev. Biol. 148, 261–272) who demonstrated its involvement in the biomineralization process of cultured mesenchyme cells.

Collagen molecules participate in the formation of a wide array of supramolecular aggregates that confer critical biomechanical and physiological properties to the tissues and organs of all metazoan organisms (1). Based on morphological criteria, these extracellular structures are broadly divided into fibrillar (type I–III, V, and XI collagens) and nonfibrillar (type IV, VI–X, and XII–XIV collagens) networks (1). Contrasting the homogeneity of the former group is the substantial diversity of architectural forms of the nonfibrillar aggregates, which include sheet-like structures, hexagonal lattices, beaded filaments, and anchoring fibrils (1). In addition to merely serving as supportive elements, collagen molecules are also recognized to play a dynamic role in several cellular activities, morphogenetic processes, and developmental programs (2). Such a dual function is reflected in the complexity and diversification of the regulatory mechanisms that modulate the tissue- and stage-specific patterns of collagen gene expression (3).

Our current understanding of collagen structure, function, and regulation relies for the most part on studies performed in higher vertebrate systems. This despite the large body of descriptive information that exists for the collagens of simpler and experimentally more amenable organisms. A case in point is the sea urchin embryo in which collagen biosynthesis has been implicated in the progression of gastrulation and in the differentiation of skeletogenic mesenchyme cell cultures (4–6).

Biochemical and immunological data have indicated that the sea urchin contains several genetically distinct collagens believed to represent the counterparts of vertebrate fibrillar and nonfibrillar molecules (6–11). This postulate was recently confirmed by the cloning and characterization of three collagen genes from two distinct sea urchin species, *Paracentrotus lividus* and *Strongylocentrotus purpuratus* (6, 12–17).

Using a nematode collagen probe, Saitta *et al.* (13) isolated from the former organism a genomic fragment subsequently shown to encode a collagen evolutionarily related to the fibrillar group of molecules (14). The primary structure of this gene product, arbitrarily termed 1\( \alpha \) chain, was recently elucidated in its entirety in *S. purpuratus* and shown to be homologous to that of the vertebrate pro-\( \alpha 2(1) \) collagen gene (16). Likewise, the protein sequence of an unusually long fibrillar procollagen, named 2\( \alpha \) chain, was deduced from overlapping cDNAs isolated from embryonic libraries of *P. lividus* and *S. purpuratus* (15, 17). In both sea urchin species, these fibrillar collagen genes are co-expressed in the mesenchyme cell lineage of late-gastrula stage embryos (14, 15).

The third echinoid collagen was originally identified in the *S. purpuratus* genome by Venkatesan *et al.* (12) who isolated a 212-bp exon coding for an interrupted collagenous sequence using a mouse type IV collagen probe. These investigators showed that the collagen-coding probe hybridizes to a 9-kb transcript first detected during blastula formation (12), whereas others demonstrated a more complex pattern of hybridization by the same exon probe (18). To be precise, they reported that the genomic fragment cross-hybridizes with a 7-kb RNA which is first detected in the gastrulae, as well as with a second 9-kb species apparently co-expressed at blastula stage together with the homologous transcript of similar size (18). The nonfibrillar nature of this collagen gene, termed Spcoll1, was recently confirmed by partial cDNA cloning experiments which, however, left unresolved the issue

---

* This work was supported by National Institutes of Health Grant GM-41849. This is article 117 from the Brookdale Center for Molecular Biology at the Mt. Sinai School of Medicine in New York. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Present address: Inst. de Biologie et Chimie des Proteines, Centre National de la Recherche Scientifique, Lyons, France.

‡ On leave of absence from the International Institute of Genetics and Biophysics, Consiglio Nazionale delle Ricerche, Naples, Italy.

The abbreviations used are: bp, base pair(s); COLP3a, the gene coding for the *S. purpuratus* 3\( \alpha \) collagen; kb, kilobase(s); NC1, carboxyl-terminal non-collagenous domain.
pertaining to the identity of this collagen chain (6). Here we demonstrate that this nonfibrillar molecule of *S. purpuratus* is a type IV collagen α chain.

MATERIALS AND METHODS

Embryo Cultures, RNA Isolation, and Analysis—Gametes collection, *in vitro* fertilization, and embryo cultures were carried out according to the standard protocol (19). Poly(A)+ RNA was isolated by oligo(dT)-cellulose chromatography of total RNA purified from gastrulae and plutei using the procedure described by Cathala et al. (20). For Northern blot hybridizations approximately 1 μg of poly(A)+ RNA was analyzed as described previously (14, 15).

Isolation and Sequencing of cDNA and Genomic Clones—For cDNA isolation, 5 × 10⁶ recombinant phages from a late-gastrula stage *S. purpuratus* library (16, 17) were initially screened under cross-hybridizing conditions using *P. lividus* cDNA fragments encoding triple-helical sequences of the 1α and 2α procollagens (14, 15). Additional cDNA and genomic screenings were performed under stringent conditions of hybridization using appropriate *S. purpuratus* cDNA probes. Clones were sequenced as described previously (16, 17), and resulting data were analyzed with the aid of the computer program MULTALIN (21).

Determination of the Start Site of Transcription—For primer extension, 100 pg of a 32P-nucleotide oligomer complementary to the 5′-coding portion of the 3α mRNA was 32P-labeled by T4 kinase, incubated with 1 μg of late-gastrula stage poly(A)+ RNA, and extended by reverse transcriptase as described (22). For nuclease protection, a 576-bp genomic fragment, whose 3′ end lies 168 bp upstream of the ATG codon, was first subcloned into the transcription vector pT7/T3-19 (Life Technologies Inc., Gaithersburg MD). Uniformly 32P-labeled antisense RNA was then synthesized by transcribing 1 μg of linearized template using the T7 RNA polymerase. The product of the reaction was annealed to 1 μg of late-gastrula stage poly(A)+ RNA and subjected to RNase protection according to the protocol of Krieg and Melton (23). Products of the primer extension and nuclease protection reactions were analyzed by autoradiography after electrophoresis in a 5% polyacrylamide, 50% urea (w/v) gel (23).

FIG. 1. Overlapping 3α collagen cDNAs. Clones are depicted below a schematic representation of the 3α collagen chain. Signal peptide (SP) is signified by the gray rectangle, whereas amino and carboxyl non-collagenous regions (NC), as well as non-collagenous interruptions of the triple-helix (black), are in white. Relative position of cysteinyl residues (C) are indicated along with those of EcoRI sites. On the right is a Northern blot hybridization to clone I28 of RNA from gastrulae (G) and plutei (P). Size of hybridizing bands, estimated using DNA size markers, is indicated on the right side of the autoradiogram.

RESULTS

Cloning of 3α Collagen cDNAs—While this work was already in progress, Wessel et al. (6) reported partial sequence information of the transcript encoded by the genomic clone initially isolated by Venkatesan et al. (12) using a murine type IV collagen probe. To be precise, this 2721-bp-long cDNA (SpColl1) codes for a series of repeated Gly-X-Y triplets interrupted 13 times by short non-collagenous sequences (6). Consistent with previous analyses (12, 18), SpColl1 was found to hybridize to a 9-kb transcript which is first detected at the swimming blastula stage where it accumulates specifically in the mesenchyme cells (6). Concomitantly to and independently from this study, we also identified a nonfibrillar collagen gene product during our cross-species cloning of the *S. purpuratus* 1α and 2α collagen genes (16, 17). As discussed more extensively below, this 1-kb cDNA, I28 (Fig. 1), codes for a triple-helical sequence containing four interruptions, a short amino-terminal non-collagenous region, and a potential signal peptide (Fig. 1). In addition, the 168 bp which lie upstream of the start site of translation of I28 contain several stop codons. Thus, the data indicated that the sequences of I28 correspond to the 5′ end of a nonfibrillar collagen mRNA. Based on the available structural evidence, we predicted that this mRNA encodes a type IV-like α chain. For consistency with our nomenclature of the other sea urchin collagens (14–17), this product was named 3α chain and the corresponding gene COLP3α.

Upon Northern blot analysis, we found that the I28 cDNA hybridizes to a transcript of about 8–9 kb in gastrulae and to an additional 6–7-kb transcript in plutei (Fig. 1). Moreover, parallel *in situ* hybridizations revealed that the developmental pattern of COLP3α expression closely resembles that independently reported for SpColl1 (6). Hence, we hypothe-
putative 28-residue signal peptide, a 14-residue amino-terminal region is highlighted by the dotted line, whereas the ends of the cDNA clone Spcoll (6) are indicated by the stars.

sized that I28 either corresponds to the 5' portion of the Spcoll transcript or to the other 9-kb cross-hybridizing species identified by Nemer and Harlow (18) using the Spcoll genomic probe. To clarify this point, we decided to isolate the full-length sequence of COLP3α. Accordingly, an embryonic cDNA library was subjected to several rounds of screening utilizing 3' fragments of appropriate cDNA clones. This eventually led to the isolation of eight overlapping clones covering the entire COLP3α message in addition to 400 and 260 bp of 3'- and 5' non-coding sequences, respectively (Fig. 1). The 5256-bp coding sequences of the COLP3α cDNAs were also found to include the 2.7 kb of Spcoll, thus ruling out the possibility that our clones represent the cross-hybridizing domains of a type IV collagen chain. These are as follows: a putative 28-residue signal peptide, a 14-residue amino-terminal non-collagenous segment, an interrupted triple-helical domain of 1484 residues, and a 226-residue carboxyl-terminal non-collagenous (NC1) domain (Fig. 2).

The type IV collagen network represents the central core around which numerous macromolecules self-assemble to originate the sheet-like structure of basement membranes (1). At least five genetically distinct type IV collagen chains have been identified in mammals (24-26); type IV-like collagen genes have also been cloned in Drosophila melanogaster, Ascaris suum, and Caenorhabditis elegans (27-30). We utilized the available amino acid sequences of these vertebrate and invertebrate chains to more firmly establish the identity of the S. purpuratus 3α collagen. To this end, we concentrated on the analysis of the type IV collagen domains which are known to be highly conserved across different chains and species, notably the carboxyl-terminal NC1 domain and the amino-terminal 7S region (24-30). In the initial steps leading to the assembly of the basement membrane network, the former domain of the type IV trimer participates in dimers

Fig. 2. Single letter amino acid translation of 3α nucleotide sequence. Amino acid residues are numbered from the initiation site, and triangles mark the boundaries of the triple-helical domain in which imperfections (I) are numbered in the amino to carboxyl direction. The duplicated peptide of the 7S region is highlighted by the dotted line, whereas the ends of the cDNA clone Spcoll (6) are indicated by the stars.

Fig. 3. Alignment of vertebrate and invertebrate NC1 domains. Dashes represent gaps introduced to maximize sequence alignment; cysteines, numbered in the amino to carboxyl direction, are boxed and invariant residues are identified by stars. The notations clb-1 and clb-2 refer to the C. elegans gene products (30). Note that the ordering of the chains in the cross-alignment is arbitrary.
association, whereas the latter is the site whereupon a tetrameric complex is formed and stabilized by disulfide bonds and lysine-derived cross-links (1).

Alignment of the NC1 sequences of the S. purpuratus chain with those of various vertebrate and invertebrate chains (24-31) confirmed the remarkable phylogenetic preservation of this domain configuration in the sea urchin type IV collagen, as well (Fig. 3). This is particularly evident for the number of cysteines implicated in lysine-derived cross-links (1). The comparison revealed the highest levels of sequence identity between the NC1 domain of the S. purpuratus chain and those of the human a1(IV) (71.2%) and a5(IV) (70.8%). The close identity to both mammalian chains is not surprising in view of the fact that these two collagens share the same domain configuration in the sea urchin type IV collagen.

Analysis of the 5' Portion of the COLP3a Gene—In the last set of experiments we isolated the 5' portion of the COLP3a gene and thus determined the boundaries of the first exon. To this end, we isolated a phage clone by screening a genomic library with the cDNA clone L4 (Fig. 1). The resultant positive recombinant, GC18, was then analyzed by separate Southern blot hybridizations to the two EcoRI segments of L4 in order to identify the contiguous EcoRI genomic fragments. This placed the sequences of the 5' and 3' EcoRI probe of L4 in a 3.5- and 5-kb EcoRI genomic fragments of CG18, respectively (data not shown).

To determine the 5' end of the 3a mRNA, a primer exten-
the entire noncoding segment, the signal peptide, and the end of exon 1. This and the previous experiments documented EcoRI site of the 5-kb genomic subclone established the 3' extension product, 156 nucleotides in length (Fig. 5A), which is complementary to the 30 bp positioned 104 nucleotides upstream of a putative TATA box (TATTAT), beginning at position -45, and a CCAAT motif on the opposite strand between -45 and -40, and a TATA-less common promoter (35-37). In contrast, the C. elegans homologs of these genes are located on separate chromosomes (30). A similar situation may exist in the sea urchin, since preliminary evidence excludes the presence of a closely linked α2(IV)-like gene in the genomic clone we isolated.

**DISCUSSION**

We completed the primary structure of an embryonic collagen, 3α chain, from the sea urchin *S. purpuratus* and determined the promoter sequence of the corresponding gene, COLP3α. This gene product is identical to that encoded by the partial clone that Wessel *et al.* (6) recently isolated and showed to play a critical role in the *in vitro* differentiation of mesenchyme cells. We demonstrated that COLP3α encodes a type IV collagen α chain which we believe represents the α1(IV) chain of *S. purpuratus*. Although the structural data could not establish a firm relationship with either this chain or the closely related α5(IV) collagen, we rest this hypothesis on three lines of indirect evidence.

First, COLP3α is expressed at the time in which the primary basement membrane, the basal lamina lining the blastocoel of the developing embryo, is formed (6). Such a pattern of expression parallels that seen for the major type IV chains during vertebrate embryogenesis (2). Second, the 3α transcript is co-expressed with a gene product of similar size that displays an estimated 30% sequence divergence (18), a value that closely resembles the difference between the mammalian α1(IV) and α2(IV) collagens (27). Third, antibodies raised against the triple helical domain of the 3α chain have been shown to recognize a collagenase-sensitive heterotrimer in the culture media of mesenchyme cells (6).

Hence, it is reasonable to argue that the two COLP3α hybridizing bands which appear at distinct stages of sea urchin development correspond to three distinct type IV collagen gene products. To be precise, the 9-kb band might correspond to the two transcripts from the major type IV genes (α1(IV) and α2(IV)), whereas the 7-kb band might represent a minor α1(IV)-like gene product. Such an hypothesis is consistent with early immunological indications of three type IV-like α chains in the sea urchin (4), as well as with the recently recognized heterogeneity of this group of nonfibrillar collagens in mammals (1). A corollary to our hypothesis is that the differential expression of the sea urchin type IV collagen genes may be responsible for distinct morphogenetic properties of basement membranes in the developing organism. Cloning of additional type IV-like transcripts and characterization of the function of these gene products will eventually test the validity of this hypothesis.

Another interesting point emerging from our studies is that the structural identity of individual collagens and the overall diversity of this family of molecules in the sea urchin replicate to some extent the vertebrate situation. It is also evident that, regardless of their identity and function, the sea urchin collagen genes display a pattern of expression specifically restricted to the cellular derivatives of the vegetal plate and the micromeres (6, 14, 15). In this respect, it will be of interest to characterize the molecular mechanisms that orchestrate collagen gene expression in this experimental model widely used for studying cell lineage specification during animal embryogenesis (38).

**Acknowledgments**—We thank James Andriotakis for excellent technical assistance and Roseann Linezea for typing the manuscript.

**REFERENCES**

1. van der Rest, M., and Garrone, R. (1991) *FASEB J.* 5, 2814-2823
2. Hay, R. D. (1991) in *Cell Biology of Extracellular Matrix* (Hay, R. D., ed), pp. 419-462, Plenum Press, New York
3. Ramirez, F., and Di Liberto, M. (1990) *FASEB J.* 4, 1616-1623

**Fig. 6.** Nucleotide sequence of the promoter and first exon of COLP3α. The start site of translation, the TATA and CCAAT motifs, are underlined continuously, whereas the location of the EcoRI restriction site and the oligonucleotide used in the primer extension reaction are underlined by dotted lines. The minor start site of transcription is identified by the star.
5254

Invertebrate Basement Membrane Collagen

4. Wessel, G. M., and McClay, D. R. (1987) Dev. Biol. 121, 149–165
5. Blankeshp, J., and Benson, S. (1986) Exp. Cell Res. 162, 98–104
6. Wessel, G. M. Etkin, M., and Benson, S. (1991) Dev. Biol. 148, 261–272
7. Minna, S., Pucci-Minafra, I., Casano, C., and Giangugzz, F. (1975) Biol. Zool. 42, 290–298
8. Wessel, G. M., Marchese, R. B., and McClay, D. R. (1985) Dev. Biol. 103, 101–103
9. Burke, R. D., Boulang, C., and Sanderson, A. L. (1989) Comp. Biochem. Physiol. 94B, 41–44
10. Benson, S., Smith, L., Wilt, F., and Shaw, R. (1990) Exp. Cell Res. 188, 141–149
11. Shimizu, K., Amemiya, S., and Yoshizato, K. (1990) Biochim. Biophys. Acta 1038, 39–46
12. Venkataresan, M., De Pablo, F., Varghese, G., and Simpson, R. T. (1986) Proc. Natl. Acad. Sci. U. S. A. 83, 3351–3355
13. Saitta, B., Butticé, G., and Gambino, R. (1989) Biochim. Biophys. Res. Commun. 188, 633–639
14. D’Alessio, M., Ramirez, F., Suzuki, H. R., Schurz, M., and Gambino, R. (1989) Proc. Natl. Acad. Sci. U. S. A. 86, 3503–3507
15. D’Alessio, M., Ramirez, F., Suzuki, H. R., Schurz, M., and Gambino, R. (1990) J. Biol. Chem. 265, 7000–7004
16. Exposito, J. V., D’Alessio, M., Solursh, M., and Ramirez, F. (1992) J. Biol. Chem. 267, 15559–15562
17. Exposito, J. V., D’Alessio, M., and Ramirez, F. (1992) J. Biol. Chem. 267, 17404–17409
18. Nemer, N., and Harlow, P. (1988) Biochim. Biophys. Acta 950, 445–449
19. Leahy, P. S. (1988) Methods Cell Biol. 27, 1–13
20. Cathala, G., Savouret, J. F., Mendez, B., West, B. L., Karim, M., Martra, J. A., and Baxter, J. D. (1983) DNA (N. Y.) 2, 329–335
21. Corpet, F. (1988) Nucleic Acids Res. 16, 10881–10890
22. Parr, B. A., Parks, A. L., and Raff, R. A. (1990) J. Biol. Chem. 265, 1408–1413
23. Krieg, P. A., and Melton, D. A. (1985) EMBO J. 4, 3463–3471
24. Hostikka, S. L., and Tryggvason, K. (1988) J. Biol. Chem. 263, 19488–19493
25. Zhou, J., Hertz, J. M., Leinonen, A., and Tryggvason, K. (1992) J. Biol. Chem. 267, 12475–12481
26. Morrison, K. R., Germain, G. G., and Reeder, S. T. (1991) J. Biol. Chem. 266, 34–39
27. Mardones, M., Kalluri, R., Hudson, B. G., and Reeder, T. (1992) J. Biol. Chem. 267, 1253–1258
28. Blumberg, B., MacKrell, A. J., and Fessler, J. H. (1988) J. Biol. Chem. 263, 18529–18537
29. Pettijohn, J., and Kinston, B. I. (1991) J. Biol. Chem. 266, 16149–16156
30. Guo, X., and Kramer, J. M. (1989) J. Biol. Chem. 264, 17574–17582
31. Kurkijärvi, M., Cordon, M. B., Blumberg, B., Bazlow, D., Quinnones, S., Saus, J., and Pihlajaniemi, T. (1987) J. Biol. Chem. 262, 8496–8498
32. Timpl, R., Wiedemann, H., Van Den Dool, V., Puth effect, H., and Kuhni, K. (1981) Eur. J. Biochem. 126, 203–211
33. Hostikka, S. L., Eddy, R. L., Byers, M. G., Højby, M., Shows, T. B., and Tryggvason, K. (1990) Proc. Natl. Acad. Sci. U. S. A. 87, 1606–1610
34. Ruoslahti, E., and Pierschbacher, M. D. (1987) Science 238, 491–497
35. Poiesz, E., Pollner, R., and Kuhn, K. (1988) EMBO J. 7, 2887–2895
36. Burkel, P., Martin, G., and Yamada, Y. (1988) Proc. Natl. Acad. Sci. U. S. A. 85, 9674–9682
37. Soisson, R., Hutoari, M., Hostikka, S. L., Prokop, D. J., and Tryggvason, K. (1988) J. Biol. Chem. 263, 17217–17220
38. Davidson, E. H. (1989) Development (Camb.) 105, 421–445