Eur. J. Biochem. 234, 59–65 (1995)


# Characterization of two genes coding for a similar four-cysteine motif of the amino-terminal propeptide of a sea urchin fibrillar collagen

Jean-Yves EXPOSITO, Nicolas BOUTE, Gilbert DELEAGE and Robert GARRONE

Institut de Biologie et Chimie des Protéines, Centre National de la Recherche Scientifique, Unité Propre de Recherches 412, Université Claude Bernard, Lyon, France

(Received 22 May/2 August 1995) – EJB 95 0815/2

We report the characterization of the 5′ region of the gene coding for the 2α fibrillar collagen chain of the sea urchin *Paracentrotus lividus*. This sequence analysis identified the intron/exon organization of the region of the gene coding for the signal peptide, the cysteine-rich domain and the 12 repeats of the four-cysteine module of the unusually long amino-propeptide. This still unknown four-cysteine motif is generally encoded by one exon, which confirms that the distinct amino-propeptide structures of the fibrillar collagens arise from the shuffling of several exon-encoding modules. Moreover, Southern-blot analysis of the sea urchin genome and sequencing of selected genomic clones allowed us to demonstrate that several sea urchin genes could potentially code for the four-cysteine module. Curiously, one of these genes lacks the exons coding for four repeats of this motif while, in another gene, the same exons are submitted to an alternative splicing event.

*Keywords:* invertebrate; sea urchin; collagen; gene evolution; amino-propeptide.

Collagens constitute a large family of structural proteins of the extracellular matrix present in metazoan organisms (van der Rest and Garrone, 1991; Mayne and Brewton, 1993). These proteins form a large spectre of supramolecular aggregates, i.e. the cross-striated fibrils for the fibrillar collagens (type I, II, III, V and XI), sheet-like structures, hexagonal lattices, beaded filaments and anchoring fibers for some of the non-fibrillar collagens (van der Rest and Garrone, 1991; van der Rest and Bruckner, 1993). The most well known and homogeneous group is the so-called fibrillar collagens. All of the fibrillar collagen molecules are composed of three identical or similar α chains, each being made of an uninterrupted series of Gly-Xaa-Yaa triplets (approximately 338) forming the collagenous domain, flanked by two non-collagenous extensions, the amino-propeptide and the carboxyl-propeptide (van der Rest and Garrone, 1991; Vuorio and de Crombrugghe, 1990). The assembly of these three α chains allows the formation of the precursor procollagen molecule. During the maturation of collagens, the amino-propeptide and the carboxyl-propeptide parts are generally removed (Vuorio and de Crombrugghe, 1990).

Among the vertebrate fibrillar collagen chains, the most conserved domain is the non-collagenous C-propeptide, whereas the central triple-helical domain is well conserved in size, although its sequence is variable (Vuorio and de Crombrugghe, 1990). The last domain, the N-propeptide, represents the most variable part of the fibrillar collagen chains and three distinct N-propeptide configurations have been characterized (Lee et al., 1991). They differ from each other in size and organization. All of the N-propeptides contain a short triple-helical segment and a non-collagenous amino-telopeptide. Between the signal peptide and the short triple-helical region, a cysteine-rich globular domain, also called the thrombospondin 2 (tsp-2) motif (Bork, 1992) is present in structure I; it is absent in structure II and replaced in structure III by a long globular region with an acidic subdomain and a basic subdomain. The N-propeptide looks like a mosaic of peptide motifs which can be submitted to alternative splicing at the messenger RNA level and/or exists in other extracellular components. The alternatively spliced form of the pro-α1(II) chain entirely lacks the cysteine-rich globular domain (Ryan and Sandell, 1990; Ryan et al., 1990). This cysteine-rich region is present in two thrombospondin molecules (Lawler and Hynes, 1986; Adams and Lawler, 1993). *Drosophila* sog and *Xenopus* chordin proteins contain four repeats of a cysteine-rich motif which is distantly related to the cysteine-rich globular region of the fibrillar procollagen chains and the thrombospondin molecules (Francois et al., 1995; Sasai et al., 1995). The basic part of the long globular region present in the N-propeptide of structure III is also called the thrombospondin 1 (tsp-1) motif (Bork, 1992) or the PARP domain (proline/arginine-rich protein; Zhidkova et al., 1993). The PARP domain of the fibrillar pro-α1(V), pro-α1(XI) and pro-α2(XI) chains (Zhidkova et al., 1993; Greenspan et al., 1991; Takahara et al., 1991; Yoshioka and Ramirez, 1990) is also present in the non-fibrillar collagen types IX, XII, XIV and XVIII (Rehn and Pihlajaniemi, 1994) and in thrombospondin molecules (Lawler and Hynes, 1986; Adams and Lawler, 1993). More recently, different isoforms of the chicken and rat pro-α1(XI) collagen chains (Zhidkova et al., 1995; Thom Oxford et al., 1995) and human and mouse pro-α2(XI) collagen chains (Zhidkova et al., 1995; Tsumaki and Kimura, 1995) have been characterized. These isoforms result from

*Correspondence to* J. Y. Exposito, IBCP, UPR-CNRS 412, 7 passage du vercors, F-69367 Lyon cedex 07, France

*Fax:* +33 72 72 26 02.

*Abbreviations.* PARP, proline-arginine-rich protein; *COLL2α*, the gene coding for the *Paracentrotus lividus* 2α chain; *COLP2α*, the gene coding for the *Strongylocentrotus purpuratus* 2α chain; *COLP5α*, the gene coding for the *Strongylocentrotus purpuratus* 5α chain; SURF, sea urchin fibrillar peptide motif.

*Note.* The nucleotide sequences reported in this paper have been submitted to the GenBank/EMBL Data Bank and are available under accession numbers X89800-X89806.

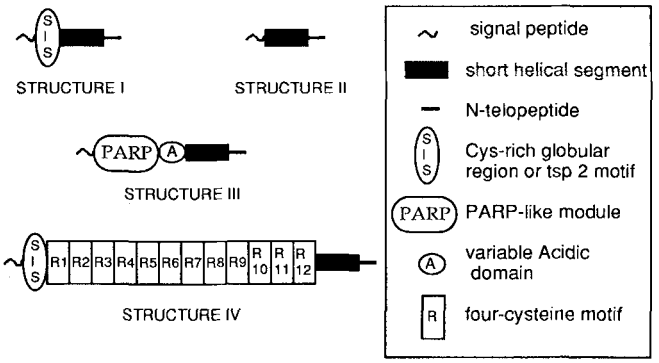



**Fig. 1. Schematic representation of the N-propeptide configurations of fibrillar procollagens.** Note that the various subdomains are not represented in scale. In vertebrate, the N-propeptide of structure I included the pro-$\alpha$1(I), pro-$\alpha$1(II), pro-$\alpha$1(III) and pro-$\alpha$2(V) chains (Vuorio and de Crombrugghe, 1990; Woodbury et al., 1989). The pro-$\alpha$2(I) chains and the alternatively spliced form of the pro-$\alpha$1(II) chains represented the N-propeptide configuration of structure II (Vuorio and de Crombrugghe, 1990; Ryan and Sandell, 1990). The N-propeptide of structure III included the pro-$\alpha$1(V), pro-$\alpha$1(XI) and pro-$\alpha$2(XI) chains (Greenspan et al., 1991; Takahara et al., 1991; Yoshioka and Ramirez, 1990; Zhidkova et al., 1993).

a complex alternative splicing system which involves the sequence coding for the acidic subdomain of the N-propeptide.

The fibrillar collagens are also well represented in invertebrate species. Information concerning the primary structure of fibrillar collagens was obtained in a fresh water sponge (Exposito and Garrone, 1990; Exposito et al., 1993), in a deep sea worm (Mann et al., 1992) and in sea urchin (D'Alessio et al., 1989; Exposito et al., 1992a,b). These data clearly indicated that the fibrillar collagen gene family had evolved relatively little. Moreover, the data concerning the sea urchin fibrillar collagen chain emphasized the variability of the N-propeptide structure and the mosaic organization of this domain (Exposito et al., 1992a,b). A new N-propeptide configuration was characterized (Exposito et al., 1992b). This new N-propeptide configuration, or structure IV, consists of a structure I, in which a new domain is located between the cysteine-rich region and the short triple-helical segment. This new domain contains 12 repeats (R1–R12, from the amino to the carboxyl part) of a still unknown four-cysteine motif. The consensus sequence of this 140–145-amino-acids motif is $X_{(40)}GX_2LWX_{11}GXGX_{39}CX_6CX_2L/FX_{(23)}CX_{(4)}CX_1$ (where numbers in parentheses represents an average number of residue). Several alternatively spliced isoforms of these sea urchin fibrillar collagen chains differ in distinct combinations of this four-cysteine motif. One of the isoform lacks the four-cysteine repeats R2–R5, and another lacks the four-cysteine motifs R6–R8.

In our study, we present data concerning the gene organization of the sequence coding for this new N-propeptide configuration. Genomic analysis confirmed the mosaic structure of the N-propeptide domain and allowed us to show that a family of genes could potentially code for this new protein module.

## MATERIALS AND METHODS

**Materials.** Restriction enzymes and modified enzymes were purchased from Promega. [$\alpha$-32P]dCTP at 3000 Ci/mmol, [35S]dATP[$\alpha$S] at 1000 Ci/mmol and [$\gamma$-32P]ATP at 3000 Ci/mmol were obtained from New England Nuclear.

**Cloning and DNA sequencing.** For genomic isolation, $2\times10^5$ recombinant phages from a *Strongylocentrotus purpuratus* (Exposito et al., 1992a) or *Paracentrotus lividus* (kindly provided by Dr Christian Gache, station marine, Villefranche sur mer, France) genomic library were screened under cross-hybridizing conditions using *S. purpuratus* cDNA coding for the N-terminal part of the 2$\alpha$ fibrillar collagen chain (Exposito et al., 1992b). Purification and analysis of genomic recombinants were carried out according to standard protocols (Sambrook et al., 1989). Restriction fragments of the genomic clones were subcloned into the pBluescript SK vector (Stratagene) and sequenced using the Sequenase TM enzyme (U.S. Biochemical Corp.) with the dideoxynucleotide chain termination procedure on double-stranded DNA (Zagursky et al., 1986). In some cases, synthetic oligonucleotides primers purchased from Isoprim were used. All the coding sequences were determined from both strands.

**Computer analysis.** The sequences were analyzed by the computer programs DNAid (Dardel and Bensoussan, 1988) and Antheprot (Geourjon and Deléage, 1993). Evolutionary trees were calculated with the programs PHYLIP and the maximum likehood method (Golding and Felsenstein, 1990).

## RESULTS

Previous studies have led us to characterize two sea urchin fibrillar collagen chains by cDNA cloning. The first, 1$\alpha$, represented a vertebrate homolog of the pro-$\alpha$2(I) collagen chain (Exposito et al., 1992a), whereas the second, the 2$\alpha$ chain, consisted of a new fibrillar collagen class, with a unique N-propeptide configuration, which we previously called structure IV (Exposito et al., 1992b; Fig. 1).

**Genomic organization of the gene coding for *P. lividus* 2$\alpha$ chain *COLL2$\alpha$* gene.** To elucidate the gene organization of this new class of fibrillar collagen genes, but also to obtain data about the promoter region, a genomic library from the sea urchin *P. lividus* was screened, using as a probe the F6 cDNA insert that contains part of the 5′ untranslated region and the sequence coding for the signal peptide, the cysteine-rich domain and the repeats R1–R5 and R9–R11 of the 2$\alpha$ chain of the sea urchin *S. purpuratus*. This switch was decided since *P. lividus*, a Mediterranean sea urchin species, is more easily available in France. Among the 10 genomic clones isolated with the cross-species screening, four of them were more extensively analyzed after preliminary blot studies using selected part of the F6 cDNA insert (Figs 2A and 3). The overall identity between the amino-terminal part of the 2$\alpha$ chain from the two sea urchin species was about 85%. The same inter-species identity had been previously shown for the carboxyl-terminal region of this chain (Exposito et al., 1992b). The 35-kb sequence overlapped by the four genomic clones included the 18 5′ exons of *COLL2$\alpha$*. The most 5′ genomic clone contained the 5′ exon or exon 1 of this gene, with the 5′ untranslated region and the 97-bp sequence coding for the signal peptide. Exon 2, of 219 bp, coded for the cysteine-rich region. The same genomic organization had been established for the gene coding for the vertebrate pro-$\alpha$1(I), pro-$\alpha$1(II), pro-$\alpha$1(III) and pro-$\alpha$2(V) chains (D'Alessio et al., 1988; Ryan et al., 1990; Benson-Chanda et al., 1989; Truter et al., 1993). All the repeats, except R1, R3 and R5, were encoded by one exon, which began with the two last bases of a codon. The repeats R1, R3 and R5 were coded by two exons, with the intronic sequences at the same location as for repeats R3 and R5 (Fig. 4A). Surprisingly, two exons could potentially code for repeat R4. These two exons shared nearly 100% identity, whereas the flanking intronic sequences are more divergent (Fig. 4B). Since two overlapping genomic clones showed the same genomic or-

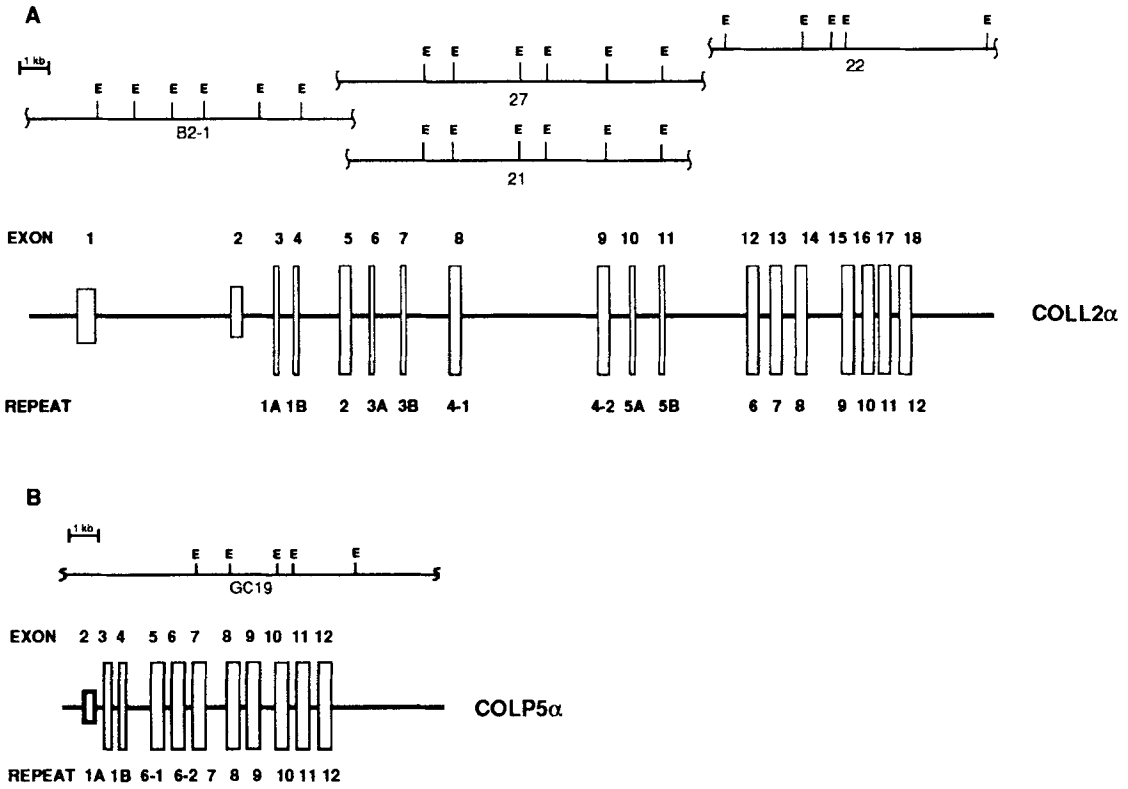



**Fig. 2. Partial restriction map of the sea urchin 2α (A) and 5α (B) procollagen genomic clones together with the relative position of the exons.** (A and B) Exons are represented with empty boxes below the genomic clones. For each exon, the top and the bottom numbers indicate the numbering, 5′ to 3′, of the genes and the corresponding name of the coded repeat, respectively. For repeats R1, R3 and R5 of the 2α chain and r1 of the 5α chain, which are coded by two exons, A and B indicated the amino-terminal and the carboxyl-terminal part, respectively. E, *Eco*RI.

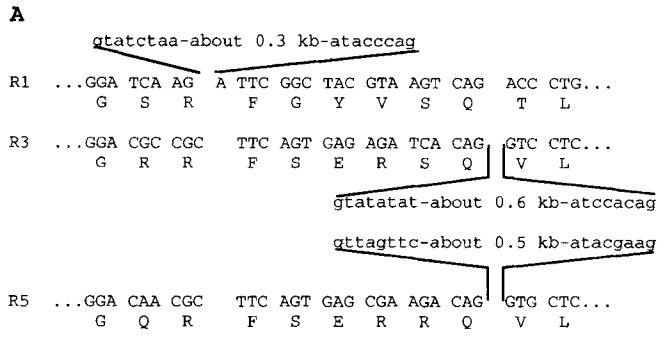
**Table 1. Sequence identities between the different repeats of the 2α chain and repeats r1 and r8 of the 5α chain.**

| Identity for | Identity for | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R1 | r1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | r8 | R9 | R10 | R11 | R12 |
| R1 | * | 58.5 | 29.3 | 34.8 | 25.7 | 28.8 | 22.1 | 24.1 | 21.4 | 26.6 | 30.1 | 20.7 | 21.8 | 24.3 |
| r1 | | * | 29.5 | 29.0 | 25.9 | 23.2 | 24.3 | 24.3 | 23.0 | 23.0 | 26.6 | 20.9 | 24.1 | 22.5 |
| R2 | | | * | 22.5 | 70.9 | 20.7 | 39.0 | 27.9 | 31.1 | 32.4 | 27.1 | 29.2 | 22.9 | 29.3 |
| R3 | | | | * | 21.6 | 64.0 | 27.3 | 23.0 | 23.7 | 25.9 | 33.1 | 26.6 | 30.9 | 26.1 |
| R4 | | | | | * | 20.1 | 39.0 | 28.6 | 35.0 | 33.8 | 31.2 | 30.3 | 27.0 | 27.1 |
| R5 | | | | | | * | 28.8 | 25.4 | 20.9 | 22.3 | 28.1 | 27.3 | 32.4 | 26.1 |
| R6 | | | | | | | * | 23.7 | 35.7 | 33.1 | 22.9 | 34.8 | 27.0 | 28.6 |
| R7 | | | | | | | | * | 26.1 | 26.1 | 34.5 | 27.1 | 26.4 | 25.2 |
| R8 | | | | | | | | | * | 69.8 | 27.9 | 47.9 | 24.5 | 26.6 |
| r8 | | | | | | | | | | * | 26.6 | 44.6 | 27.5 | 26.8 |
| R9 | | | | | | | | | | | * | 30.5 | 33.8 | 27.0 |
| R10 | | | | | | | | | | | | * | 29.1 | 28.6 |
| R11 | | | | | | | | | | | | | * | 28.1 |

ganization, we suggest that duplication of repeat R4 is not a cloning artefact. At this time, we do not know if the two repeats are used, and in which configuration. Nevertheless, sequencing of one reverse-transcribed RNA, amplified by PCR, revealed that the R4-2 repeat is expressed at the pluteus stage.

Alignment studies of the 12 repeats (Table 1) revealed that the identity is 20–48%, except for repeats R2 and R4 (71% identity) and R3 and R5 (64% identity). One of the alternatively spliced isoforms previously characterized lacked these four repeats (Exposito et al., 1992b). With this study, we can separate the 12 repeats into two more homologous classes, i.e. the uneven and the even repeats (Fig. 5). From these different approaches, we could elaborate the putative steps leading to the formation of the genomic region coding for these 12 repeats (Fig. 6). The first step was the duplication of a primordial exon (genetic unit 1) coding for a four-cysteine motif flanked by these intronic sequences. This led to the formation of a region, or genetic unit 2, coding for an uneven repeat and an even repeat. After four sets of duplication of this genetic unit, a first intron insertion in repeat R1 occurred. A second intron insertion led to the formation of a new genetic unit, genetic unit 3. Duplication of genetic unit 3 led to the formation of the exons coding for repeats R2–R5. As discussed later, the order of the different steps cannot be clearly established.


**Fig. 3. Comparison of the amino-terminal part of the 2α and 5α chains.** Residues are numbered from the initiation site of translation and from the cysteine-rich region for the 2α and the 5α chains, respectively. Black lozenge symbols indicate the relative positions of the exon boundaries. When these boundaries are included in a codon, two numbers on the left of the lozenge symbol indicate the exact location of the intervening sequence. On the right of the lozenge symbol, the name of the repeat of the four-cysteine motif is indicated. Vertical bars and horizontal doted lines indicate amino acids identities and gap inserted to obtain the best alignment, respectively. Only one of the R4 repeats was included.



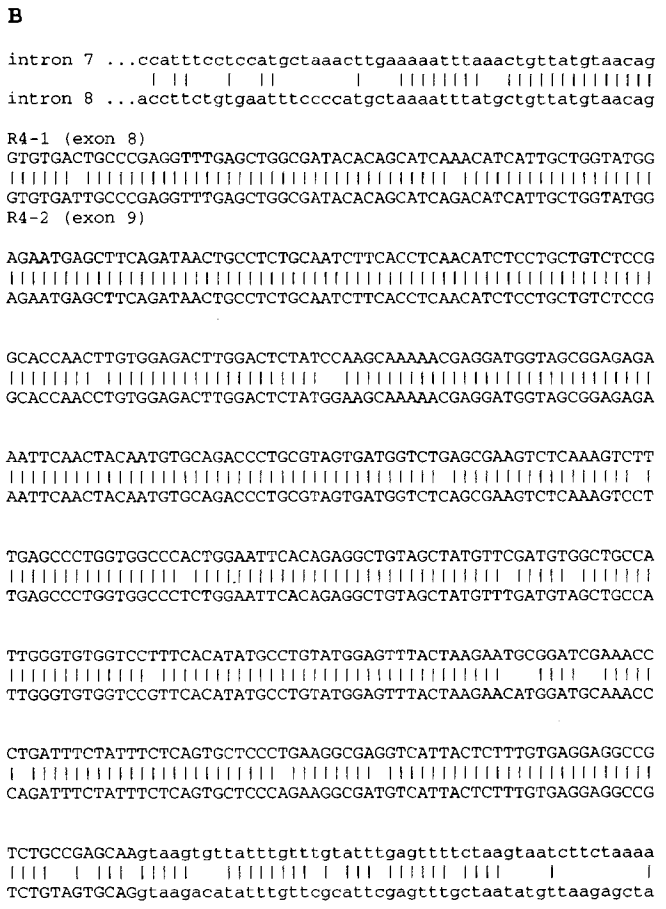




**Fig. 4. Intronic boundaries of the exon coding for repeats R1, R3 and R5 (A) and for the two repeats R4 (B) of the 2α chain.** Upper-case letters illustrate exon sequences, lower-case letters indicate the intronic sequences.

**A new family of genes coding for this four-cysteine motif.** During screening of genomic libraries from two sea urchin species (*P. lividus* and *S. purpuratus*), several slightly positive genomic clones were detected and purified. Moreover, Southern-blot analysis of sea urchin genomic DNA under moderate stringency revealed that at least two distinct parts of the sea urchin genome contained sequences related to this new peptide motif (data not shown). To confirm these results, one of these slightly positive genomic clones, GC19, from the *S. purpuratus* library, was further characterized. This genomic clone was selected since it is the only one that hybridized to probes specific for repeats R1 and R12. Selected parts of GC19 were sequenced (Figs 2B and 3). We characterized 11 exons of a new gene, which we called *COLP5α* for the gene coding for the 5α collagen-like chain of

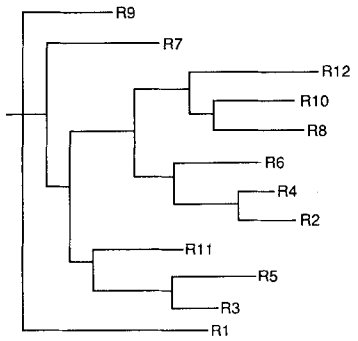



**Fig. 5. Evolutionary tree calculated from the nucleic acids sequences of the 12 repeats of the four-cysteine motif of the 2α chain.** The length of the horizontal lines are proportional to the similarity between the sequences, whereas the length of the vertical lines has been chosen to obtain a clearer representation.

the sea urchin *S. purpuratus*, with respect to our previous nomenclature (Exposito et al., 1992a). The most 5′ exon codes for a cysteine-rich region or tsp-2 module, and the others code for nine repeats of the four-cysteine motif. Alignment studies revealed that these nine repeats present in the 5α chain have a 2α counterpart, with an interchain identity of about 60%. To distinguish a similar repeat of the four-cysteine motif between the two genes, lower-case letters were used for the 5α chain instead the upper-case letters for the 2α chain. Following the cysteine-rich domain, the 5α chain contained a r1 repeat, two r6 repeats (r6-1 and r6-2) and repeats r7−r12. The two r6 repeats shared 58% sequence identity. Moreover, the whole sequence between the exons coding for repeat r1 and r6-2 was fully sequenced. In comparison to the *COLL2α* gene, the *COLP5α* gene showed the same organization. Repeat r1 was coded by two exons and the other repeats were encoded by one exon. Reverse-transcription PCR analysis failed to detect any *COLP5α* expression during the first stages of the sea urchin embryogenesis.

## DISCUSSION

We characterized the genomic organization of the sequence coding for the unusual long N-propeptide of the sea urchin fibrillar 2α procollagen chain and confirmed the mosaic structure with module evolution of the different N-propeptide configuration. This allowed us also to precise the limits of the new four-cysteine peptide motif that we call SURF (sea urchin fibrillar peptide motif). Genomic analysis led us to characterize a new family of genes which could potentially code for this SURF peptide motif.

The exon/intron structure of the two genes characterized in this study confirms numerous studies which showed that functional or structural domains are encoded by one exon (Traut, 1988). The exon shuffling model could explain the presence of the different exon modules in different genes. In the two sea urchin genes, two types of exon modules were characterized. The first encoded the cysteine-rich region or the so-called tsp-2 peptide motif. This domain was characterized in fibrillar collagen chains harboring an N-propeptide configuration of type I, but also in the thrombospondins 1 and 2 (Bork, 1992). The second exon module encoded the SURF motif. 75% of the SURF motif characterized was encoded by one exon. Two intronic insertions explained that three of the SURF motifs were coded by two exons. This is in favour of the model of Gilbert et al. (1986) who argued that the progenitor gene contained all the primordial introns and that intron insertion or deletion might occur during evolution.

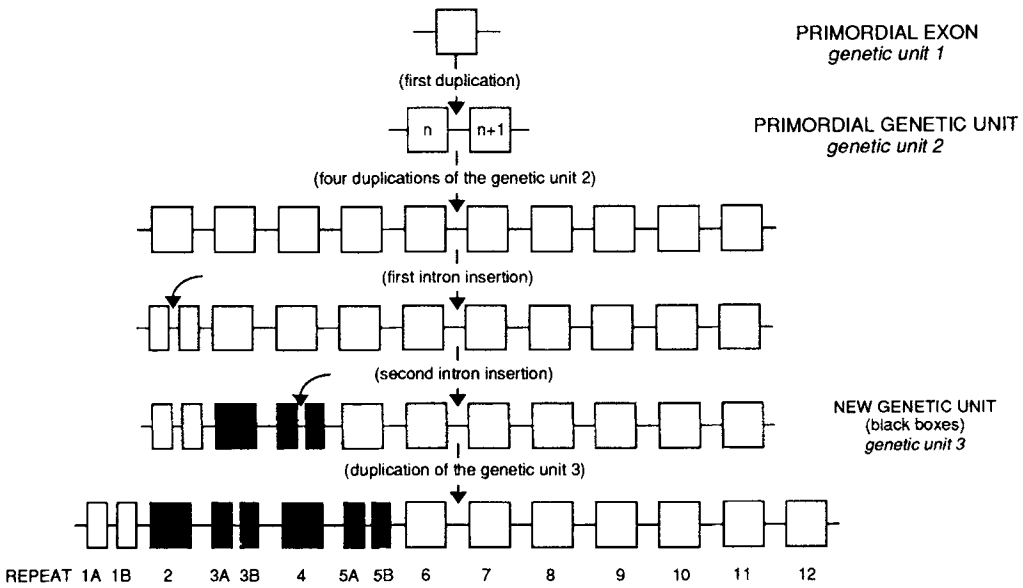



**Fig. 6. Putative steps leading to the formation of the genomic region coding for the 12 repeats of the four-cysteine motif of the 2$\alpha$ chain.**

From our study, we propose a model of evolution of *COLL2α* which invokes that three different genetic units are submitted to duplications (Fig. 6). The third genetic unit, which was only observed in *COLL2α*, was formed by the intronic insertion in an uneven coding exon and allowed the formation of the exons coding for the SURF motif R2−R5 after duplication. Although our model could explain the formation of the genomic region coding for these SURF motif, it was difficult to define exactly the order of the different steps. According to our data, the second intron insertion, which led to the formation of the third genetic unit, was a more recent event since (a) the SURF motifs R2−5 were absent in *COLP5α*, (b) the two SURF pairs R2-R4 and R3-R5, showed the greatest identity (70% compared to 19−30%), whereas the same SURF motif in the two genes showed approximately 60% identity. We suggest that the duplication event that leads to the formation of these two genes is prior to the formation of the third genetic unit. Of course, we cannot exclude that the two genes evolved at the same speed and that a deletion event occurred in *COLP5α*.

Beside these suggestions, the most interesting feature is that the SURF motifs R2−R5 are involved in one of the alternatively spliced isoforms (Exposito et al., 1992b). We isolated a new gene, *COLP5α*, that lacked the genetic information coding for these four repeats whereas, in the other gene, an alternative splicing complex led to the presence or the absence of these SURF motifs in the procollagen chain. Unfortunately, we do not know at the present time if *COLP5α* is a functional gene, at which stage of development it is expressed and if it codes for a fibrillar collagen chain. We believe that it is really the case since, as already shown for the fibrillar collagen chains, the cysteine-rich region or tsp-2 domain is encoded by a single exon, whereas the same domain is encoded by two exons in the thrombospondin genes (Lawler and Hynes, 1986). The putative proteins encoded by *COLL2α* and *COLP5α* show the same module arrangement, whereas a different organization has been shown for the thrombospondin molecules (Adams and Lawler, 1993) and the *Drosophila* sog and *Xenopus* chordin proteins (Francois et al., 1995; Sasai et al., 1995). Moreover, *COLL2α* and *COLP5α* show the same genomic organizations.

The N-propeptide, which is the most variable part of the fibrillar procollagen chains, increases this diversity by a complex alternative splicing event. This is shown for the vertebrate pro-α1(II) chain cysteine-rich procollagen globular domain (Ryan and Sandell, 1990), for the vertebrate pro-α1(XI) and pro-α2(XI) variable acidic procollagen domain (Zhidkova et al., 1995; Thom Oxford et al., 1995; Tsumaki and Kimura, 1995), and for the sea urchin 2α chain SURF modules. The function of these SURF motifs remains to be understood, as does the functional significance of these different alternatively spliced isoforms and probably different gene products.

This work was supported partly by a grant from the *Association pour la Recherche sur le Cancer*. Thanks are due to Lauraine Panaye for improvements to the English.

## REFERENCES

Adams, J. & Lawler, J. (1993) The thrombospondin family, *Curr. Biol. 3*, 188−190.

Benson-Chanda, V., Su, M. W., Weil, D., Chu, M. L. & Ramirez, F. (1989) Cloning and analysis of the 5′ portion of the human type-III procollagen gene (*COL3A1*), *Gene (Amst.) 78*, 255−265.

Bork, P. (1992) The modular architecture of vertebrate collagens, *FEBS Lett. 307*, 49−54.

D'Alessio, M., Bernard, M., Pretorious, P. J., de Wet, W. & Ramirez, F. (1988) Complete nucleotide sequence of the region encompassing the twenty-five exons of the human proα1(I) collagen gene (*COL1A1*), *Gene (Amst.) 67*, 105−115.

D'Alessio, M., Ramirez, F., Suzuki, H. R., Solursh, M. & Gambino, R. (1989) Structure and developmental expression of a sea urchin fibrillar collagen gene, *Proc. Natl Acad. Sci. USA 86*, 9303−9307.

Dardel, F. & Bensoussan, P. (1988) DNAid: aMacintosh full screen editor featuring a built-in regular expression interpreter for the search of specific patterns in biological sequences using finite state automata, *Comput. Appl. Biosci. 4*, 483−486.

Exposito, J. Y. & Garrone, R. (1990) Characterization of a fibrillar collagen gene in sponges reveals the early evolutionary appearance of two collagene gene families, *Proc. Natl Acad. Sci. USA 87*, 6669−6673.

Exposito, J. Y., D'Alessio, M., Solursh, M. & Ramirez, F. (1992a) Sea urchin collagen evolutionarily homologous to vertebrate pro-α2(I) collagen, *J. Biol. Chem. 267*, 15559−15562.

Exposito, J. Y., D'Alessio, M. & Ramirez, F. (1992b) Novel aminoterminal propeptide configuration in a fibrillar procollagen undergoing alternative splicing, *J. Biol. Chem. 267*, 17404−17408.

Exposito, J. Y., van der Rest, M. & Garrone, R. (1993) The complete intron/exon structure of *Ephydatia müller*i fibrillar collagen gene

suggests a mechanism for the evolution of an ancestral gene module, *J. Mol. Evol. 37*, 254–259.

Francois, V., Solloway, M., O'Neill, J. W., Emery, J. & Bier, E. (1994) Dorsal-ventral patterning of the Drosophila embryo depends on a putative negative growth factor encoded by the short gastrulation gene, *Genes & Dev. 8*, 2602–2616.

Geourjon, C. & Deléage, G. (1993) Interactive and graphic coupling between multiple alignments, secondary structure predictions and motif/pattern scanning into proteins, *Comput. Appl. Biosci. 9*, 87–91.

Gilbert, W., Marchionni, M. & McKnight, G. (1986) On the antiquity of introns, *Cell 46*, 151–154.

Golding, B. & Felsenstein, J. (1990) A maximum likehood approach to the detection of selection from a phylogeny, *J. Mol. Evol. 31*, 511–523.

Greenspan, D. S., Cheng, W. & Hoffman, G. G. (1991) The pro-$\alpha$1(V) collagen chain, *J. Biol. Chem. 266*, 24727–24733.

Lawler, J. & Hynes, R. O. (1986) The structure of human thrombospondin, an adhesive glycoprotein with multiple calcium-binding sites and homologies with several different proteins, *J. Cell Biol. 103*, 1635–1648.

Lee, B., D'Alessio, M. & Ramirez, F. (1991) Modifications in the organization and expression of collagen genes associated with skeletal disorders, *Crit. Rev. Eukaryotic Gene Expr. 1*, 172–187.

Mann, K., Gaill, F. & Timpl, R. (1992) Amino-acid sequence and cell-adhesion activity of a fibril-forming collagen from the tube worn *Riftia pachyptila* living at deep sea hydrothermal vents, *Eur. J. Biochem. 210*, 839–847.

Mayne, R. & Brewton, R. (1993) New members of the collagen superfamily, *Curr. Opin. Cell Biol. 5*, 883–890.

Rehn, M. & Pihlajaniemi, T. (1994) $\alpha$1(XVIII), a collagen chain with frequent interruptions in the collagenous sequence, a distinct tissue distribution, and homology with type XV collagen, *Proc. Natl Acad. Sci. USA 91*, 4234–4238.

Ryan, M. C. & Sandell, L. J. (1990) Differential expression of a cysteine-rich domain in the amino-terminal propeptide of type II (cartilage) procollagen by alternative splicing of mRNA, *J. Biol. Chem. 265*, 10334–10339.

Ryan, M. C., Sieraski, M. & Sandell, L. J. (1990) The human type II procollagen gene: identification of an additional protein-coding domain and location of potential regulatory sequences in the promoter and first intron, *Genomics 8*, 41–48.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular cloning: a laboratory manual*, 2nd edn, Cold Spring Harbor Laboratory, Cold Spring Harbor NY.

Sasai, Y., Steinbeisser, H., Geissert, D., Gont, L. K. & Robertis, E. M. (1994) Xenopus chordin: a novel dorsalizing factor activated by organizer-specific homeobox genes, *Cell 79*, 779–790.

Takahara, K., Sato, Y., Okazawa, K., Okamoto, N., Noda, A., Yaoi, Y. & Kato, I. (1991) Complete primary structure of human collagen $\alpha$1(V) chain, *J. Biol. Chem. 266*, 13124–13129.

Thom Oxford, J., Doege, K. J. & Morris, N. P. (1995) Alternative exon splicing within the amino-terminal nontriple-helical domain of the rat pro-$\alpha$1(XI) collagen chain generates multiple forms of the mRNA transcript which exhibit tissue-dependant variation, *J. Biol. Chem. 270*, 9478–9485.

Traut, T. (1988) Do exons code for structural or functional units in proteins? *Proc. Natl Acad. Sci. USA 85*, 2944–2948.

Truter, S., Andrikopoulos, K., di Liberto, M., Womack, L. & Ramirez, F. (1993) Pro-$\alpha$2(V) collagen gene; pairwise analysis of the amino-propeptide coding domain, and cross-species comparison of the promoter sequence, *Connect. Tissue Res. 29*, 51–59.

Tsumaki, N. & Kimura, T. (1995) Differential expression of an acidic domain in the amino-terminal propeptide of mouse pro-$\alpha$2(XI) collagen by complex alternative splicing, *J. Biol. Chem. 270*, 2372–2378.

van der Rest, M. & Bruckner, P. (1993) Collagens: diversity at the molecular and supramolecular levels, *Curr. Opin. Struct. Biol. 3*, 430–436.

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

Vuorio, E. & de Crombrugghe, B. (1990) The family of collagen genes, *Annu. Rev. Biochem. 59*, 837–872.

Woodbury, D., Benson-Chanda, V. & Ramirez, F. (1989) Amino-terminal propeptide of human pro-$\alpha$2(V) collagen conforms to the structural criteria of a fibrillar procollagen molecule, *J. Biol. Chem. 264*, 2735–2738.

Yoshioka, H. & Ramirez, F. (1990) Pro-alpha 1(XI) collagen. Structure of the amino-terminal propeptide and expression of the gene in tumor cell lines, *J. Biol. Chem. 265*, 6423–6426.

Zagursky, R. J., Berman, M. L., Baumeister, K. & Lomax, N. (1986) Rapid and easy sequencing of large linear double stranded DNA and supercoiled plasmid DNA, *Gene Anal. Tech. 2*, 232–238.

Zhidkova, N. I., Brewton, R. G. & Mayne, R. (1993) Molecular cloning of PARP (proline/arginine-rich protein) from human cartilage and subsequent demonstration that PARP is a fragment of the NH2-terminal domain of the collagen $\alpha$2(XI) chain, *FEBS Lett. 326*, 25–28.

Zhidkova, N. I., Justice, S. K. & Mayne, R. (1995) Alternative mRNA processing occurs in the variable region of the pro-$\alpha$1(XI) and pro-$\alpha$2(XI) collagen chains, *J. Biol. Chem. 270*, 9486–9493.