Short Chain Collagens in Sponges Are Encoded by a Family of Closely Related Genes*

Jean-Yves Exposito, Dominique Le Guellec, Qin Lu, and Robert Garrone†

From the Institut de Biologie et Chimie des Protéines, Centre National de la Recherche Scientifique Unité Propre de Recherche 412, Associée à l’Université Claude Bernard, Bâtiment 403, 43 Boulevard 11 Novembre 1918, 69622 Villeurbanne, France

Two previously described sponge cDNAs, EmC4 and C23, respectively encoding a short chain collagen and a fibrillar collagen, were used to characterize collagen gene families in a freshwater sponge. EmC4 detected several clones when used to screen a cDNA library. Two overlapping clones, EmC13 1 and 2, were sequenced and appeared highly homologous to EmC4. Contrarily to C23, EmC4 hybridized with 10–12 fragments of genomic DNA digested with restriction endonucleases and detected 10 times more positive clones than C23 when used to screen a genomic library. The genomic clone G41 contained two closely related genes, COLNF13, corresponding to EmC13 and COLNF6. Partial characterization of COLNF13 revealed two partial exons and four complete exons of 153, 219, 207, and 144 base pairs, with split glycine codons at their boundaries. The deduced encoded protein is a short chain collagen containing two uninterrupted collagenous domains of 66 and 171 amino acids and non-collagenous domains. A characterized 207-base pair exon of COLNF6 is 77% identical with the comparable COLNF13 exon. In situ hybridization using EmC4 cDNA and electron microscopy suggested that the cells expressing these genes were secreting spongins, a non-fibrillar, surface collagen of these sponges.

Collagens are extracellular, structural proteins present in all multicellular animals (1). In spite of this wide occurrence, their structure has been mainly studied in a few vertebrate species such as chicken, rat, mouse, bovine, and man. Fourteen collagen types have been described so far (2, 3). Several collagen types (I–III, V, and XI) belong to a closely related group of molecules and are referred to as the fibrillar collagens. They are all derived from a single ancestral gene and are expressed in mesodermal derivatives (20, 21) and are developmentally regulated (22). In addition to these vertebrate gene counterparts, invertebrates contain other structural classes of collagen. The best evidence comes from the study of cuticular collagens in nematodes (23, 24). The number of genes coding for the collagenous proteins of the cuticle has been estimated to range between 40 and 150 (25). These genes are small in size (about 1 kb) when compared with the vertebrate collagen genes described so far (7). They encode four classes of proteins, all containing two short triple helical domains (24).

We have recently characterized by cDNA analysis a nonfibrillar, short chain collagen in sponges (26). The deduced structure reveals a unique organization with features reminiscent of type IV collagen (NC1 domain) and of nematode cuticular collagens. The data presented here demonstrate that the sponge genome contains a family of about 10 closely related nonfibrillar collagen genes. In situ hybridization data indicate that the products of these genes may correspond to the microfibrillamentous sponge collagen called spongins.

EXPERIMENTAL PROCEDURES

Sponge Culture—The freshwater sponges Ephydatia müllerii were obtained in the laboratory from assexual buds (gemmules) collected in small water courses near Lyons and stored at 4 °C. 3–6-day-old sponge cultures were used in this study.

Materials—Restriction endonucleases and other nucleic acid-modifying enzymes were purchased from Boehringer Mannheim. Nitrocellulose filters (0.45 μm, pore diameter) and [α-32P]dCTP at 3000 Ci/mmol were obtained from Amersham Corp. [35S]dATP(δS) at 1338 Ci/mmol and GeneScreen Plus® hybridization transfer membrane were purchased from Du Pont-New England Nuclear. All of these materials were used according to the supplier’s recommendations.

Screening of the cDNA and Genomic Libraries—The insert cDNA of the EmC4 clone (26), labeled with [α-32P]dCTP by the random primed oligolabeling technique (27), was used to screen cDNA and genomic libraries previously synthesized (15, 26). Using the plaque hybridization method (28), positive clones were isolated under high stringencies for the washing of the nitrocellulose filters (15 mM NaCl, 1.5 mM sodium citrate, pH 7, 0.1% sodium dodecyl sulfate at 65 °C). Southern Blotting—Genomic DNA isolated from sponge cultures,

* 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.

† To whom correspondence should be sent: Cytologie Moleculaire, Bâtiment 403, Université Claude Bernard, 43 Bd. 11 Novembre 1918, 69622 Villeurbanne Cedex, France. Tel.: 33-72-44-82-80; Fax. 33-78-84-33-84.

‡ The abbreviations used are: kb, kilobase(s); bp, base pair(s).
according to the method of Maniatis et al. (29), and genomic clones were digested with several restriction endonucleases, and the fragments were separated by agarose gel electrophoresis. The DNA transfer on GeneScreen Plus® membranes, the hybridization and washing of these filters under high stringencies were performed according to the protocols of the New England Nuclear recommendations.

**DNA Sequencing and Sequence Analysis**—The insert cDNAs and restriction fragments of the genomic clones were sequenced by the deoxyxynucleotide chain termination method (30) using modified T7 polymerase (Sequenase kit, United States Biochemical Corp.) and [32P]dATP. Sequencing of both strands was performed in M13 mp18 and/or M13 mp19 phages. The orientation of the cloning fragments was tested by the S1 nuclease digestion method (31). Computer analysis was performed using the DNAid program (32).

**In Situ Hybridization**—3-day-old sponges were fixed for 30 min at 4°C with 1% paraformaldehyde in 0.04 M phosphate buffer, pH 7.2, and embedded in paraplast. 6-μm sections of the developed sponges were collected onto gelatin-coated slides and stored at 4°C until use. They were then deparaffinized and incubated in 5 μg/ml proteinase K in 20 mM Tris-HCl, pH 7.4, 2 mM CaCl2 for 15 min at room temperature, rinsed in phosphate-buffered saline and dehydrated in absolute ethanol. The EmC4 cDNA was labeled with [32P]dCTP to a specific activity of 4 × 10⁸ cpm/μg and purified through a Sephadex G-50 column followed by ethanol precipitation. Sections were incubated for 16 h at 37°C in 50% formamide, 4 × SSC (1 × SSC: 0.16 M NaCl, 0.016 M sodium citrate), 1% Denhardt’s solution, 10% dextran, 500 μg/ml DNA from herring sperm (Boehringer Mannheim), and 1 mg/ml tRNA with 0.25 μg/ml of denatured probe. They were then dehydrated in 4 × SSC and washed at room temperature during 30 min in each of the following solutions: 2 × SSC (twice), 1 × SSC (twice), 0.5 × SSC. The sections were then dipped in K5 photographic emulsion (Ilford, France), exposed 5–15 days at 4°C, processed in D19 developer (Kodak, France), and stained with toluidine blue. Observations were performed with a Zeiss Universal microscope.

**Electron Microscopy**—3-day-old sponge cultures were fixed in 0.4% glutaraldehyde in 0.02 M sodium cacodylate, hydrochloric acid buffer, pH 7, for 45 min at 4°C. They were then washed in the same buffer, postfixed for 30 min at 4°C in 1% osmium tetroxide buffered with 0.02 M sodium cacodylate, hydrochloric acid, pH 7. After a progressive dehydration with ethanol, the samples were embedded in Epon. 1-μm sections were observed in phase contrast microscopy in order to select areas for ultrastructural studies. The ultrathin sections were stained with uranyl acetate and lead citrate. The observations were carried out on a Philips 300 electron microscope.

**RESULTS**

**Characterization of Nonfibrillar Collagen cDNAs**—A cDNA library has been constructed in Agt10 from poly(A)-rich RNA purified from 3-day-old freshwater sponge cultures. This led to the isolation of a clone, EmC4, coding for a short chain nonfibrillar collagen (26). The 1.2-kb cDNA insert of this clone was used to screen the same cDNA library at high stringency. Several additional positive clones have been detected. Two of the analyzed clones, EmC13-1 and EmC13-2, were found to be distinct from EmC4 (Fig. 1A). These overlapping clones contained at their 3′ end a poly(A) stretch of 48 units. The longest clone contained an open reading frame coding for 102 amino acids and was terminated by a stop codon TGA with 188 nucleotides downstream. The open reading frame extended to the 5′ end of the insert, suggesting that this cDNA is partial. The nucleotide sequences of the translated region and the amino acid sequence of the presumptive translation product were 89% identical to EmC4. A 75% identity was obtained for the untranslated regions.

**The Corresponding Nonfibrillar Collagen Genes**—In order to estimate the number of distinct homologous nonfibrillar collagen genes in the genome of the freshwater sponge E. mulleri, the genomic DNA of this sponge was digested with several restriction endonucleases and subsequently analyzed by Southern blotting (Fig. 2A). Two identical sets of digests were tested with 32P-labeled EmC4 or C23 cDNA probes, the
latter coding for a sponge fibrillar collagen (15). Using the EmC4 probe, 10–12 bands of variable intensity were revealed for each enzyme used. With the C23 probe, only one positive band was obtained for every digest, except for the PvuII. However, we have shown previously that several PvuII sites are present in the C23 insert cDNA (18). This result suggests that the fibrillar collagen gene is only present at one copy per haploid genome.

When the genomic library constructed in XGEM 11 (15) was screened with the same two cDNA probes, the EmC4 probe revealed 10 times more positive clones than the C23 probe. Among the 80,000 clones screened with the EmC4 cDNA probe, 82 were highly positive. Two of them, G41 and G54, were digested with the restriction endonuclease SacI and subsequently analyzed by Southern blotting using as a probe the 32P-labeled EmC4 insert cDNA (Fig. 2B). A single band corresponding to a 3-kb fragment was observed for the clone G54, whereas the clone G41 showed several fragments of 0.1 (not visible on the blot presented), 0.4, 0.7, 0.85, and 7 kb. The latter two fragments did hybridize less intensely than the others. Since EmC4 cDNA contained several SacI endonuclease restriction sites, the clone G41 was selected for further studies. Sequences from two distinct nonfibrillar collagen genes were found in this clone. They appeared separated by less than 1 kb and they were oriented in opposite directions (Fig. 1B). The gene COLNF12 corresponds to the EmC13 cDNA and yielded the strongly hybridizing fragments detected on the Southern blot (Fig. 2B). The extremity of the insert corresponding to the 3' end of COLNF13 was characterized (Fig. 3A). A sequence (AUUAAA) 14 nucleotides from the 3' end (20 nucleotides from the poly(A) tail of the cDNA sequence) is reminiscent of the canonical polyadenylation signal AUUAAA (33). A noteworthy difference with the EmC4- deduced protein was that the EmC13 protein contained a putative 3'-splicing site. Cysteine residues are underlined, and a potential N-glycosylation site is underlined with dashes. The termination codon is indicated by a black dot, and the putative polyadenylation signal is underlined. Nucleotide numbers are indicated on the right of the figure. The transcribed sequence that is similar between COLNF13 and the previously described EmC4 cDNA clone (26) is comprised of nucleotides 225–705. Amino acid numbers are listed under nucleotide numbers; their position only the completely characterized domains COL2 (C2), NC2, COL1 (C1), and NC1, which are numbered separately. B, COLNF6, asterisks point to identical amino acids in the two protein chains derived from the genomic sequences.
COL2, the 13 amino acids of the noncollagenous domain NC2, and the first 51% amino acids of the amino-terminal part of the COL1 domain. The 5% partially characterized exon coded for at least the last 16 amino acids of a noncollagenous domain NC3 and the first 6% amino acids of the collagenous domain COL2. The second gene, COLNF6, was contained in part in the 0.85-kb SacI fragment. An exon of 207 bp was completely sequenced (Fig. 3B). It begins and ends by a split codon glycine and has 77% sequence identity with the comparable exon of the COLNFI3 gene. At the amino acid level, the translation products are 70% identical. No apparent homology was observed between the comparable intronic sequences (Fig. 3).

Detection of the Nonscllular Collagen mRNAs by in Situ Hybridization—The development of E. mülleri from asexual gemmules is complete after 3 days in culture, and the general outlines of its organization are shown on a diagram (Fig. 5). Autoradiograms of paraffin-embedded sponge sections, hybridized with the labeled EmC4 probe, showed clusters of silver grains on aligned basal cells (Fig. 6A) and on individual cells, always located near the siliceous skeletal units of the animal (Fig. 6B). No significant labeling was observed over other cell types. Control sections were unlabeled (Fig. 6C).

Electron Microscopy—An electron microscope study of developing sponges was undertaken to characterize the cell types labeled by in situ hybridization. The results were consistent with previous ultrastructural descriptions (34). The basal cell layer is composed of flattened cells lying on a microfibrillar pad of spongin (a collagenous protein of sponges morphologically different from intercellular collagen fibrils). These basal cells are connected with internal cells associated with the siliceous skeleton (Fig. 7A). Individual elements of the skeleton are also visible, embedded within the basal spongin (Fig. 7A). Bundles of 20-nm-wide collagen fibrils are present inside the mesenchyme of the sponge. In some instances, they appeared close to the basal cells (Fig. 7A, A and inset). The cells that are close to the skeleton contain numerous profiles of rough endoplasmic reticulum. They are separated from the skeleton by a meshwork of 10-nm-wide spongin microfilaments containing an amorphous component (Fig. 7B).

**DISCUSSION**

The data reported here establish the existence of a new family of collagen genes present in sponges. COLNFI3 and COLNF6 genes have homologous intron-exon structures, and their putative translation products are very similar. The prototype of these collagen genes was previously described as EmC4 (26). It is composed of two short collagenous domains, COL1 and COL2, of 171 and 66 amino acids, respectively, with perfect Gly-X-Y repeats and three noncollagenous domains, a long (156 amino acids) C-terminal NC1 domain, a short NC2 domain separating the triple helices, and a partially characterized N-terminal NC3 domain. Results from in situ hybridization showed that these genes are expressed mainly in cells located at the base of the animal and in cells neighboring the siliceous skeleton. Electron microscope examination of developing sponges revealed that this localization corresponds to cells of the basal epithelium and of the spongin-secreting tissues (34). In freshwater sponges, it is known that these two cell types are similar. They often join to form a continuous epithelium delineating the sponge basal surface and ramifying inside the animal body, around the skeleton (Fig. 5). They secrete the collagenous microfibrils (34) constituting the spong II (35). Spong II is composed of a mixture of these 10-nm microfibrils and amorphous ma-
The hasal epithelial layer.

...animal sticks to its substratum and which link together the individual siliceous elements of the skeleton (36). In freshwater sponges this collagenous material can be considered as an exoskeleton (37). Although sponge collagen fibrils can be deposited by almost all cells, the secretion of spongulin is restricted to a few cell types (36), making the results of in situ hybridization highly significant.

The Southern blot analysis of the genomic sponge DNA (Fig. 2A) demonstrates that approximately 10 highly homologous genes belong to the same family as the gene coding for the EmC4 chain. Even if an amplified genomic library doesn't fairly represent the genome, results obtained by screening the sponge genomic library with the EmC4 cDNA probe indicate that there could be 10 times more genes for nonfibrillar collagen than for fibrillar collagen. In the nonfibrillar collagen gene family, the gene COLNF13, coding for the chain EmC13, has been partially characterized at the cDNA and genomic DNA level, and the gene COLNF6 has been investigated at the genomic level only. Some of the detected genomic sequences could, of course, represent pseudogenes. The high sequence homology between these genes could correspond to (i) a recent event of duplication, (ii) a conversion event, and/or (iii) a high selection pressure.

This gene family is reminiscent of the nematode cuticular collagen gene family, since (i) the gene products could have a comparable superficial location, (ii) similarly located cysteines are found in the sponge protein domains NC2 and NC3 and in the nematode protein domains C and A (24), (iii) in both animals, the corresponding mRNAs have short sizes, and (iv) there is a high sequence homology within each family. However, the putative gene products differ in the NC1 domain, in the size of the triple helical domains and by the presence or absence of triple helix imperfections. Moreover, no homology in intron-exon structures can be found between the nonfibrillar collagen genes of sponges and nematodes. In sponges, the sequences coding for the triple helical domains of the EmC6 and EmC13 chains correspond to several exons. All of them can be related to a 54-bp unit (i.e., 153 = 54 + 54 + 45, 144 = 54 + 45 + 45, 207 = 54 + 54 + 54 + 45). These data are in agreement with the hypothesis of Buttice et al. (38), who argued that the sequence encoding the triple helical domain of nonfibrillar collagen genes for vertebrates and invertebrates is related to the 54-bp unit, as for fibrillar collagen genes. However, the main difference between exons coding for sponge nonfibrillar or fibrillar collagen lies in their extremities. All the exons coding for the triple helical domain of fibrillar collagen begin by an intact glycine codon and terminate by an intact Y codon (15), as is the case for the exons of vertebrate fibrillar collagen genes. As shown in this paper, all the introns occurring in collagenous domains of nonfibrillar collagen genes occur between the first and the second base of a split glycine codon. Similar exon/intron splice junctions have been observed in vertebrate nonfibrillar collagen genes, although in this case intact glycine codons have been also detected.

The fact that split glycine codons are present in a collagen gene family belonging to the most primitive multicellular animals suggests two hypotheses. (i) The two collagen gene organizations, corresponding grossly to the fibrillar and the nonfibrillar collagen families, arose from two distinct primordial sequences, one with an intact glycine codon at its 5' end, the other with the last two bases of a glycine codon in this region. The reshuffling of exons with such structures could explain the existence of mosaic genes (i.e., genes coding for vertebrate collagen types IV and IX) containing the two kinds of extremities. (ii) The two collagen gene organizations arose from the same ancestral sequence, but have diverged very early into two distinct sequences. Considering the nonrandomness of the distribution of the split codon glycine, Vuorio and de Crombrugge (7) suggested that these split codons arose as secondary events occurring in initially nontruncated exons. Our data suggest that such split codons represent a very early event in evolution.

The sponge nonfibrillar collagens are composed of domains presenting similarities with other collagens. As with the EmC4 sponge collagen (26), the NC1 domain of the EmC13 chain can be divided into two subdomains, like the NC1 domain of basement membrane collagens. In each subdomain, the sequence following the first cysteine is highly conserved between the sponge collagen chains and the type IV collagen chains. The cysteines in the EmC13 noncollagenous domains are conserved in the similar domains (C and A) of nematode cuticular collagens (24), and the two collagenous domains of EmC13 have sizes almost identical with the COL1 (66 residues) and COL2 (172 residues) domains in the E3 transcript (11) of the human type XIII collagen (Fig. 4).

Our data demonstrate that at least the two genes COLNF13 and COLNF6 are organized into a cluster, since their sequences have been detected in the same G41 genomic clone. They are separated by less than 1 kb and are orientated in opposite directions. These genes belong to a family of closely related genes expressed mainly in epithelial cells that elabo-

---

2 J.-Y. Exposito, R. Garrone, and M. van der Rest, manuscript in preparation.
rate stacked layers of a nonfibrillar collagen. By these characteristics, these collagens bear some resemblance to nematode cuticular collagen and basement membrane collagen.

Acknowledgments—We are greatly indebted to M. van der Rest for his constant interest in this work and his critical reading of the manuscript. We thank R. Willems and A. Bosch for their expert technical assistance.

REFERENCES


Fig. 7. Transmission electron microscopy. A, sponge basis. An epithelial layer (E) is lying against basal spongin (BS) and extends around the inorganic skeleton (SK). The boxed area enlarged in the inset shows a cross-section of a bundle of internal collagen fibrils. ×4800. Bar, 1 μm. Inset, ×19,700. Bar, 100 nm. B, cell secreting spongin around the inorganic skeleton. MF, 10-nm-wide spongin microfilaments; A, amorphous deposits. ×21,900. Bar, 500 nm.