Invertebrate Data Predict an Early Emergence of Vertebrate Fibrillar Collagen Clades and an Anti-incest Model

Abdel Aouacheria, Caroline Cluzel, Claire Lethias, Manolo Gouy, Robert Garrone, Jean-Yves Exposito

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 279, No. 46, Issue of November 12, pp. 47711–47719, 2004

Printed in U.S.A.

Invertebrate Data Predict an Early Emergence of Vertebrate Fibrillar Collagen Clades and an Anti-incest Model*[S]

Received for publication, August 5, 2004, and in revised form, August 31, 2004
Published, JBC Papers in Press, September 8, 2004, DOI 10.1074/jbc.M408950200

Abdel Aouacheria‡§, Caroline Cluzel¶, Claire Lethias¶, Manolo Gouy‡, Robert Garrone¶, and Jean-Yves Exposito¶‖

From the ¶Institut de Biologie et Chimie des Protéines, CNRS, Unité Mixte de Recherche 5086, Institut Fédératif de Recherche 128 BioSciences Lyon-Gerland, Université Claude Bernard-Lyon 1, 7 Passage du Vercors, 69367 Lyon Cedex 07, France and ‡Laboratoire de Biométrie et Biologie Evolutive, Unité Mixte de Recherche 5558, CNRS, Université Claude Bernard-Lyon 1, 69622 Villeurbanne Cedex, France

Fibrillar collagens are involved in the formation of striated fibrils and are present from the first multicellular animals, sponges, to humans. Recently, a new evolutionary model for fibrillar collagens has been suggested (Boot-Handford, R. P., Tuckwell, D. S., Plumb, D. A., Farrington Rock, C., and Poulsom, R. (2003) *J. Biol. Chem*. 278, 31067–31077). In this model, a rare genomic event leads to the formation of the founder vertebrate fibrillar collagen gene prior to the early vertebrate genome duplications and the radiation of the vertebrate fibrillar collagen clades (A, B, and C). Here, we present the modular structure of the fibrillar collagen chains present in different invertebrates from the protostome *Anopheles gambiae* to the chordate *Ciona intestinalis*. From their modular structure and the use of a triple helix instead of C-propeptide sequences in phylogenetic analyses, we were able to show that the divergence of A and B clades arose early during evolution because α chains related to these clades are present in protostomes. Moreover, the event leading to the divergence of B and C clades from a founder gene arose before the appearance of vertebrates; altogether these data contradict the Boot-Handford model. Moreover, they indicate that all the key steps required for the formation of fibrils of variable structure and functionality arose step by step during invertebrate evolution.

Fibrillar collagens are present from sponges to humans and are the primary component of striated fibrils (1, 2). A fibrillar procollagen molecule is made of three identical or different pro-α chains. Each pro-α chain contains a central triple helix made of ~338 Gly-Xaa-Yaa triplets flanked by non-collagenous telopeptide regions, which in turn are flanked by the N- and C-propeptides. The C-propeptide is known to contain the most conserved regions of the fibrillar α chains. This domain is involved in α chain recognition and in the registration of the major triple helical domain, an important step preceding elongation of the triple helix from the carboxyl to the amino terminus. In vertebrates, a short region of the C-propeptide appears to be involved in chain recognition (3). Recently, Boot-Handford and Tuckwell (4) indicated that most of this recognition sequence is absent in all invertebrate fibrillar chains characterized to date. During the extracellular maturation of procollagen molecules, the propeptides are generally removed by specific proteinases. The resultant collagen molecules participate in fibril formation.

In humans, fibrillar collagens are subdivided quantitatively into major (types I-III) and minor (types V/XI) collagens. According to the collagen types incorporated into the fibrils and their ratio, the partial processing of the N-propeptide, and interactions with other extracellular matrix components, the shape and functional properties of the fibrils can vary. The importance of quantitatively minor collagens in the regulation of fibril diameter has been pointed out in several studies (5–7). From the model of Linsenmayer *et al.* (6), the retention of the type V N-propeptide at the surface of types I/V heterotypic fibrils is one of the key elements regulating the diameter of these fibrils.

From phylogenetic studies and the exon/intron organization, it has been shown that vertebrate fibrillar collagens can be divided into two clades (8–10): the A clade, including types I-III and the pro-α2(V) chain, and the B clade, including the pro-α1(V), pro-α3(V), and type XI chains. Moreover, the α chains of the A and B clades possess a vWc[1] and a TSPN module in their N-propeptide, respectively, in addition to a minor triple helix. It should be noted that the A clade pro-α2(I) chain presents a short N-propeptide reduced to the minor triple helix.

Recently, *COL24A1* and *COL27A1* have been characterized (11–13). These two genes encode collagen chains belonging to a new fibrillar collagen group, the C clade. They contain a C-propeptide, a major triple helix, and an N-propeptide including a TSPN module but not a minor triple helix. Moreover, the major triple helix is shorter than that of classical fibrillar collagens and presents several imperfections in the Gly-Xaa-Yaa triplet repeat. It has been suggested that vertebrate fibrillar collagens share a single common ancestor that arose at the very dawn of the vertebrate world and prior to the genome duplication events (4, 12). From invertebrate data, these authors have argued that this ancestor possesses a vWc module in its N-propeptide, a major triple helix, and a C-propeptide lacking most of the elongated chain selection sequence. The founder vertebrate fibrillar would have acquired this sequence and the characteristics of A clade members. Duplication of this gene

* 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.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental figures.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AJ786364 and AJ786365

§ Supported by the Association pour la Recherche contre le Cancer.

‖ To whom correspondence should be addressed. Tel.: 33-4-72-72-26-77; Fax: 33-4-72-72-26-04; E-mail: jy.exposito@ibcp.fr.

[1] The abbreviations used are: vWc, von Willebrand factor-type C; TSPN, thrombospondin amino-terminal-like domain; contig, group of overlapping clones; EST, expressed sequence tag.

and the swapping of the exon encoding the vWc module for that encoding a TSPN domain would permit the formation of the B clade. The most recent clade, the C, would then have arisen from the B clade after the deletion of the elongated chain selection sequence. This model is clearly distinct from our previous studies indicating that some invertebrate collagens are more closely related to B clade than A clade collagens (14–16). Moreover, it is difficult to understand how a sequence can be gained from invertebrates to vertebrates and then be lost during the divergence of the C clade from the B clade. One explanation for this conundrum is that almost all the invertebrate fibrillar chains described to date are quantitatively major ones and that, like their vertebrate counterparts, they possess a vWc module in their N-propeptide. However, we have recently shown that sea urchin also possesses a quantitatively minor fibrillar collagen chain (17).

In this study, we present evidence arguing against the Boot-Handford *et al.* model (4, 12) and indicating that the A and B/C clades arose early during evolution. These data suggest that the diversity of the vertebrate fibrillar collagens and subsequently fibril diversity are due not to a rare genomic event but instead to a step-by-step evolutionary process from the most primitive animal to humans, which corresponds to the maximum parsimony hypothesis.

EXPERIMENTAL PROCEDURES

Genomic Cloning and Reverse Transcriptase-PCR of Paracentrotus lividus Sequences Encoding Fibrillar Collagens—A search of sea urchin genes encoding fibrillar collagen α chains was conducted at the Human Genome Sequencing Center Web site (Baylor College of Medicine, Houston, TX) using the C-propeptide sequence from the *Strongylocentrotus purpuratus* 1α chain (18). The URL of this center is www.hgsc.bcm.tmc.edu. From this analysis, we were able to discover two regions of the *S. purpuratus* genome encoding the C terminus of a classical C-propeptide distinct from the three previously characterized sea urchin fibrillar collagen chains (17–19). The two sequence files corresponded to Contig18666 and Contig84214 and encoded part of the fibrillar collagens termed 6α and 7α, respectively, in this study. To characterize the ortholog genes in the sea urchin *P. lividus*, we first amplified by PCR (30 cycles) the most 3′-exon of these two genes using 500 ng of *S. purpuratus* genomic DNA and the Taq Expand polymerase kit (Roche Applied Science). The oligonucleotides used for this amplification (6α sense primer GCATCGTCCAACTGACGTTC, 6α antisense primer GACTCTTAGGTTGATAGAGG, 7α sense primer GTCCATTTTCACTCTGAGGTCC, and 7α antisense primer GGTCCATAGTGACCTTGCTC) were synthesized by Sigma Genosys. The two amplified fragments were used to screen 80,000 genomic clones from *P. lividus* at moderate stringency as described previously (20). Shotgun sequencing of positive genomic clones for each PCR probe used permitted us to obtain sequences encoding the last Gly-Xaa-Yaa triplets and the C terminus of the *P. lividus* 6α and 7α chains. At this point, reverse transcriptase-PCR was conducted using total RNA extracted from peristome tissues. For the 6α chain, a 1206-bp fragment product was amplified with sense primer CTGAAGGACCACGTGGTGTAATG and antisense primer CGCGACTGATTCGACCTTACTG. For the 7α chain, sense primer GGTTCAGCTGGTGCAAAAGGACAAAG and antisense primer GAGTCGTTCATATTCTTCTAGCACG permitted a 1034-bp fragment to be generated. PCR conditions, purification, cloning, and sequencing of PCR fragments were carried out as described previously (20). Searches of *S. purpuratus* genomic sequences were also conducted using the TSPN sequence from the human pro-α1(V) chain. This led to the identification of two contigs (Contig3949 and Contig50826) that might encode two TSPN modules related to fibrillar collagens of the B/C clade as suggested from Blast analysis. EST analysis using the human pro-α1(V) TSPN module was conducted using another sea urchin Web server (goblet.molgen.mpg.de/cgi-bin/blast-seaurchin.cgi?db=urchibase) at the Max Planck Institute for Molecular Genetics (Berlin, Germany). This analysis led to the identification of an EST (StrPu691.007800) encoding a TSPN module, which gives the best score with human α1(IX) collagen during Blast analysis.

Accession Numbers—Protein sequences of fibrillar collagens were obtained from the European Bioinformatics Institute (www.ebi.ac.uk/). Their accession numbers are P02452, P08123, Q14047, P02461, P20908, P05997, P25940, P12107, P13942, Q7Z5L5, and Q8IZC6 for the human α1(I), α2(I), α1(II), α1(III), α1(V), α2(V), α3(V), α1(XI), α2(XI), α1(XXIV), and α1(XXVII) chains, respectively. For invertebrates, the accession codes are: sea urchin *S. purpuratus* 1α, Q26634; *S. purpuratus* 2α, Q26639; sea urchin *P. lividus* 5α, CAE53096; abalone *Haliotis discus* Hdcol1α, O97405; *H. discus* Hdcol2α, O97406; lugworm *Arenicola marina* Fam1α, P90679; freshwater sponge *Ephydatia mülleri* Emf1α, P18856 and Q06452; hydra *Hydra attenuata* HcolI, Q8MUF5. Other accession numbers are P30849 and P39059 for human α1(IX) and α1(XV) collagens.

Data Base Searches—Searches in genomic databases were done using TBLASTN (21) implemented on the *Ciona intestinalis* (genome.jgi-psf.org/ciona4/ciona4.home.html), *Apis mellifera* (www.ncbi.nlm.nih.gov/BLAST/Genome/Insects.html), and *Anopheles gambiae* (www.ensembl.org) genome project Web sites. C-propeptide sequences from human pro-α1(I), pro-α1(V), and pro-α1(XXVII) chains were used to investigate these invertebrate genomes. Four *C. intestinalis* genes encoding fibrillar α chains were obtained. These genes are *ci0100150759*, *ci0100154301*, *ci0100131606*, and *ci0100144916*. Overlapping ESTs covering the complete coding sequence were available for *ci0100150759* (cilv051j16, citb048m08, citb058a23, citb059g23, citb066c17, citb073c06, citb075p17, citb077j05, citb100l05, and citb41d02) and *ci0100154301* (ciad097j13, ciad101m17, ciad19b03, ciad20c22, cicl062k24, ciht038b02, cilv063m20, citb046p23, citb087a11, and rcitb081p22). Several ESTs covering part of the coding region were obtained for *ci0100131606* (rcilv051d20, rciad06f06, citb10e10, citb066j05, citb048f05, and cigd010e18) and *ci0100144916* (rcitb31h13, rcitb064m18, ciad68o20, and ciht012m05). The coding regions of *ci0100131606* and *ci0100144916* that were not confirmed by ESTs were deduced by their similarity to comparable regions of fibrillar collagens characterized to date and from similarities to *Ciona savignyi* ortholog genes. For this purpose, sequences from these two *C. intestinalis* genes were used to investigate the *C. savignyi* genome at NCBI. For *A. gambiae*, two fibrillar collagen genes were identified, namely *ENSANGT00000019179* and *ENSANGP00000021001*. For *A. mellifera*, two fibrillar collagen genes were also found by Blast searches at NCBI and named in this study: Api-1 and Api-2. Accession numbers of the genomic contigs including Api-1 and Api-2 are AADG02012353 and AADG030117591 (for Api-1) and AADG02005865 (for Api-2). The common name, the species name, and the abbreviations used in the text and in the figures are given in Table I.

Exon/Intron Organization of Human Fibrillar Collagen Genes—The exon/intron organization of human fibrillar collagen genes was obtained at NCBI (www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html).

Sequence Analyses—Multiple alignments were performed using ClustalW (22) and were manually corrected using the Seaview alignment editor when necessary (23). Ungapped alignments were computed to derive trees according to neighbor joining and maximum parsimony methods using the Phylo win program (23); 1,000 replicates were generated for bootstrapping analysis.

RESULTS

Diversity of Fibrillar Collagens in Sea Urchin

Three sea urchin fibrillar collagen chains (1α, 2α, and 5α) have previously been characterized (17–19). They can be divided quantitatively into major (1α and 2α) and minor (5α) collagen chains (17). Blast searching of the *S. purpuratus* genome resources using a sequence encoding the 1α C-propeptide allowed us to identify two genomic sequences encoding the C-terminal part of the C-propeptides unrelated to those of the three known sea urchin fibrillar collagens. Using these sequences, we have been able to determine the last Gly-Xaa-Yaa triplets and the complete C-propeptide sequences of two fibrillar-like α chains (6α and 7α) in the sea urchin *P. lividus* (see "Experimental Procedures"). The unique feature of these two C-propeptides is their unusual length (344 and 308 residues for 6α and 7α, respectively), with the additional sequence located between the end of the triple helix and the most amino-terminal cysteine residue (number 1) of the C-propeptide (data not shown).

B Clade Fibrillar Collagen Is Present in Protostomes

To investigate invertebrate fibrillar collagen chains, we used C-propeptide sequences from human pro-α1(I), pro-α1(V), and

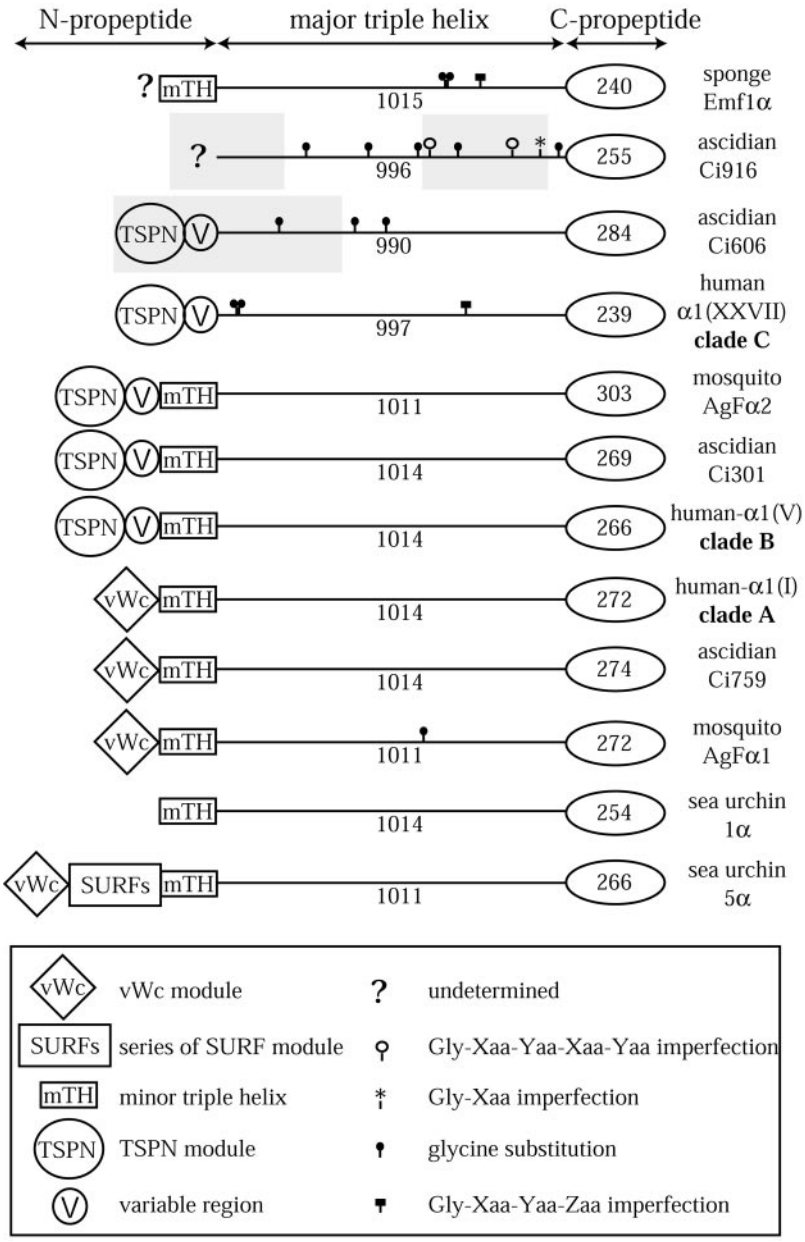

FIG. 1. **Modular structure of fibrillar α chains.** In this illustration, some human and invertebrate fibrillar collagen chains are aligned and assembled with regard to their modular structure. The different modules are not represented to scale. For each α chain, Arabic numbers indicate the length in amino acids of the major triple helix and of the C-terminal non-collagenous region including the C-telopeptide and the C-propeptide. The relative positions of glycine substitutions and imperfections present in the major triple helix are indicated. *Gray boxes* indicate regions of ascidian *C. intestinalis* α chains that are not confirmed by EST sequences. For these two regions, we have compared ortholog genomic regions of *C. intestinalis* with those of *C. savignyi*.

pro-α1(XXVII) chains to analyze the genome of three invertebrates. Two of them are the protostomes *A. gambiae* and *A. mellifera*, whereas the third is an invertebrate chordate, the ascidian *C. intestinalis*. For the two protostomes and the ascidian, two and four fibrillar α chains were deduced respectively from genomic data. The schematic structures of the *A. gambiae* and *C. intestinalis* α chains in addition to other invertebrate α chains and members of the three vertebrate clades are presented in Fig. 1.

In the protostome *A. gambiae*, the two genes encoding the fibrillar α chains have been named *AgFα1* and *AgFα2* (Table I). As shown in Fig. 1, the modular structures of the AgFα1 and AgFα2 chains are similar to those of vertebrate A and B clades, respectively. Hence, *AgFα1* encodes an N-propeptide including a vWc module, whereas *AgFα2* encodes an N-propeptide including a TSPN module and a minor triple helix. Moreover, the most C-terminal cysteine residue (cysteine 8) of the AgFα2 C-propeptide is followed by three amino acids as in the fibrillar α chains of the B clade. As in mosquito, two fibrillar collagen genes are present in the honeybee *A. mellifera* (*Api-1* and *Api-2*). One of them (*Api-2*) encodes an α chain including a TSPN module and a minor triple helix in its N-propeptide (not shown). These results indicated that the formation of the B clade arose early during evolution.

The four *C. intestinalis* genes encoding fibrillar α chains have been named *Ci759*, *Ci301*, *Ci606*, and *Ci916* (Table I). As shown in Fig. 1 and as in protostomes, *C. intestinalis* possesses two fibrillar α chains similar in their modular structure to vertebrate A and B clades, encoded by *Ci759* and *Ci301*, respectively. These similarities were also observed for the exon/intron organization of the genomic region encoding the major triple helix (Fig. 2*A*). Hence, the *Ci759* exon-intron organization is more closely related to the A clade than the B–C clades, whereas that of *Ci301* has closer similarity to the B clade. One special feature of *Ci759* is the presence of two exons of 100 and 62 bp in length corresponding to a 162-bp exon in the A clade. Insertion of an intronic sequence between the two first bases of a glycine codon in a 162-bp exon might explain this special feature (Fig. 2*A*). EST (citb100l05) covering these exons confirms the splice junctions. A special feature of *Ci301* is the

TABLE I
Species names and abbreviations used in the text and figures

Common name	Species name	Sequence identity	Abbreviation
Mosquito	*Anopheles gambiae*	ENSANGG00000016690	AgFα1
		ENSANGG00000018512	AgFα2
Honeybee	*Apis mellifera*	AADG02012353	Api-1
		AADG02005865	Api-2
Ascidian	*Ciona intestinalis*	ci0100150759	Ci759
		ci0100154301	Ci301
		ci0100131606	Ci606
		ci0100144916	Ci916
Sea urchin	*Strongylocentrotus purpuratus*	StrPu691.007800	IX
		Contig3949	3949
		Contig50826	50826

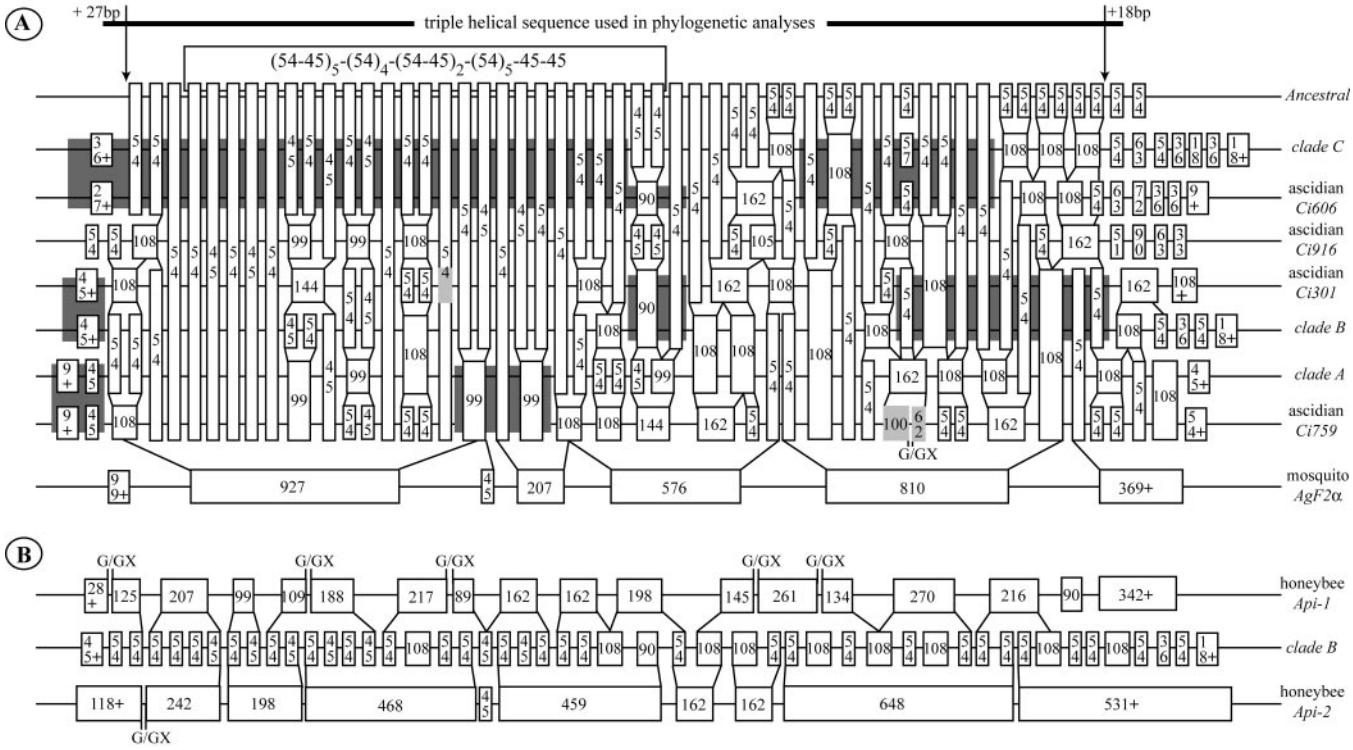

FIG. 2. **Exon/intron organization of collagen genes.** *A*, alignment of the genomic region encoding the major triple helix of *C. intestinalis*, *A. gambiae*, and prototype members of α chains of the A–C vertebrate clades. The putative structure of an ancestral fibrillar collagen gene is also presented. At the *top* the triple helical sequences chosen for the phylogenetic studies are indicated. Our first choice was to use the most conserved and specific pattern of exons between all the aligned genes (*white box*). A multiple alignment of all the complete major triple helix confirmed the relationships between these chains in this region. Moreover, this multiple alignment permits us to extend the sequence in 5′ and 3′ of our first choice (*thick black line*). *B*, the exon/intron organization of *A. mellifera* collagen genes is aligned with a prototype gene of the B clade. The gene names are indicated at the *right* of their exon/intron organization. *Arabic numbers* indicate the length of the exons in bp. For the exon junctions, the length of the sequence coding for the triple helix is followed by a plus. *G/GX* represents exons beginning with the second base of a glycine codon. The exons discussed in the text are *lightly shaded*. *Boxes* that are *shaded darkly* indicate similar exon-intron organization between genes discussed in the text.

presence of two 54-bp exons separated by an intervening sequence of 37 bp. Several ESTs (cilv063m20, citb072k01 and ciad11e06) covering this region indicated a mutual alternative use of these exons leading to the potential production of two α chain isoforms, including a 1014 amino acid triple helix.

The third *C. intestinalis* gene, *Ci606*, encodes an α chain sharing several features of the vertebrate α chains of the C clade. Hence, the N-propeptide of Ci606 contains a TSPN module but not a minor triple helix. Moreover, the major triple helix of Ci606 is composed of 990 residues and presents several imperfections (Fig. 1). Finally, the exon/intron organization of the region encoding the major triple helix of Ci606 is closely related to the C clade (Fig. 2*A*). However, the *Ci606* exon/intron organization also reminds the B clade. For the last *C. intestinalis* α chain, Ci916, we have not been able to identify the primary structure of the N-propeptide. The major triple helix is shorter, 996 residues in length, and contains two Gly-Xaa-Yaa-Xaa-Yaa and one Gly-Xaa imperfections, but also five glycine substitutions (Fig. 1).

Phylogenetic Studies

Human Fibrillar Collagens—Previous studies clearly indicated that human collagens could be divided into three clades (9, 12). The A and B clades are linked to the HOX and Notch paralogons, respectively (2, 24, 25). From their exon/intron organizations, the presence of a TSPN module in their N-propeptide and phylogenetic analyses, the C clade is more closely related to the B clade than the A clade (11, 12). This

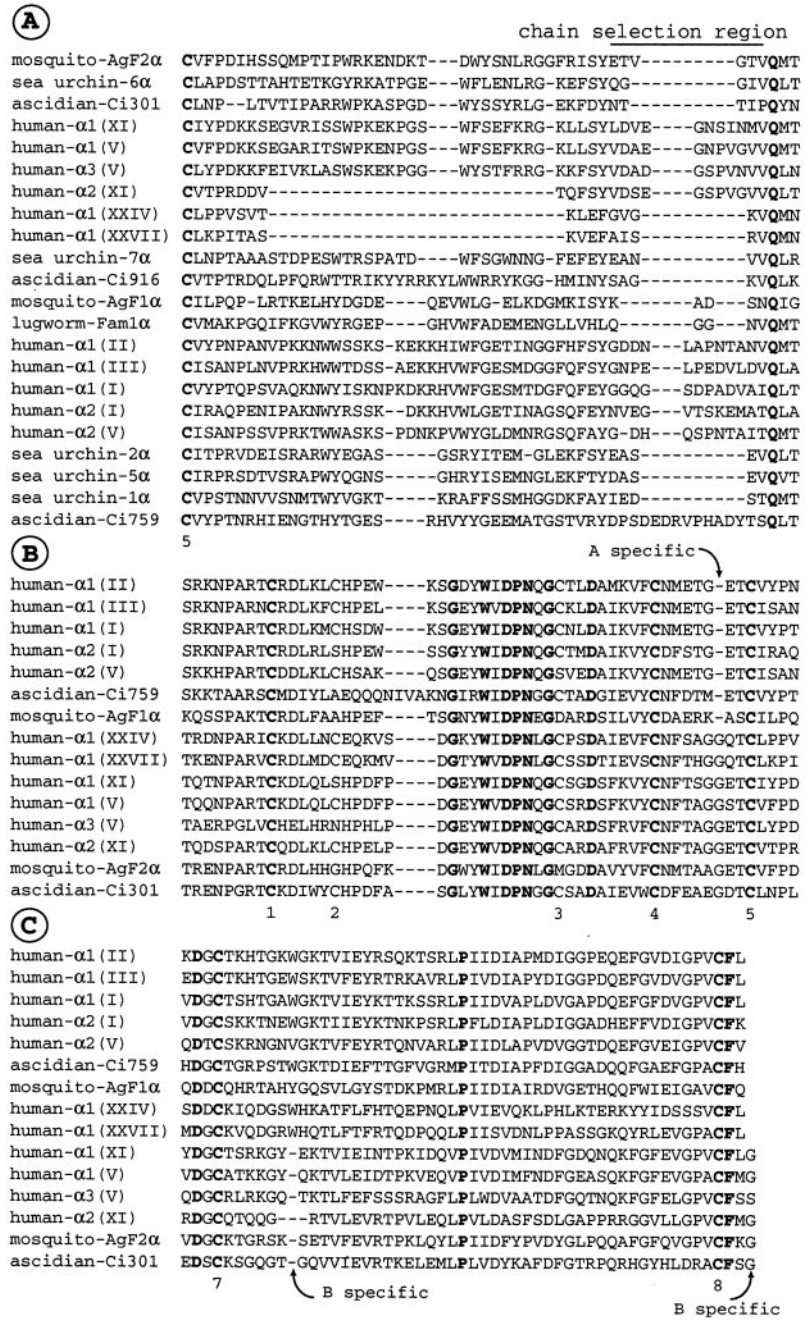

FIG. 3. **Multiple alignment of the C-propeptide.** *A*, part of the C-propeptide alignment obtained using human, sea urchin, ascidian, lugworm, and mosquito sequences. This region contains the cysteine residue number 5 and part of the chain recognition sequence. *B* and *C*, part of the C-propeptide alignment obtained using human, mosquito, and ascidian (Ci301 and Ci759) sequences. The *regions* presented in *B* and *C* include cysteine residues 1 to 5 and 7 to 8, respectively. *Arrows* indicate gaps that appear specific to the A or B clade chains. Identical residues between all the α chains analyzed are in *bold*. Numbering of the cysteine residues is indicated *below* the multiple alignments.

relationship is confirmed at the genomic level because *COL24A1* and *COL27A1* belong to the NOTCH or 1/9/19p paralogon (26).

Use of Triple Helix Instead of C-propeptide in Phylogenetic Analyses—Most of the phylogenic studies done on fibrillar collagens have used the C-propeptide sequences and more often the conserved regions of this domain (9, 12). Until now, the C-propeptide domain has been defined as the most conserved region of fibrillar collagens. However, with the increase in invertebrate data, the robustness of this definition is less convincing. Hence, except for short regions of conserved sequences around the cysteine residues, the C-propeptide contains long stretches of poorly conserved sequences, especially in the region that contains the chain recognition sequence (see Fig. 3*A*). Moreover, the size of this domain is very variable (Fig. 1). For this reason, we decided to use in our phylogenetic analyses the major triple helical sequences in addition to the C-propeptide domain. Fig. 2 illustrates the correspondence between the exons encoding the major triple helix of human and invertebrate fibrillar α chains. With the exception of the 5′- and 3′-genomic regions, a perfect alignment can be made between them. In the case of the honeybee genes (Fig. 2*B*), the discrepancies arise probably from the insertion of new introns, as suggested by the presence of introns between the first two bases of the Gly codons.

For the phylogeny studies, the most conserved regions of the C-propeptide between all the α chains were used. For the triple helix, our first choice of sequence included the exon pattern $[(54-45)_5\text{-}(54)_4\text{-}(54-45)_2\text{-}(54)_5\text{-}45\text{–}45]$. Multiple alignment of the major triple helix of human and invertebrate α chains confirms our choice (data not shown) and permits us to define another set of triple helix sequences including 906 residues (Fig. 2*A*). This larger sequence was used in the phylogenetic analyses.

Phylogenetic analysis made from C-propeptide sequences using maximum parsimony (Fig. 4*A*) and neighbor joining (see supplemental data) methods confirms that human fibrillar α

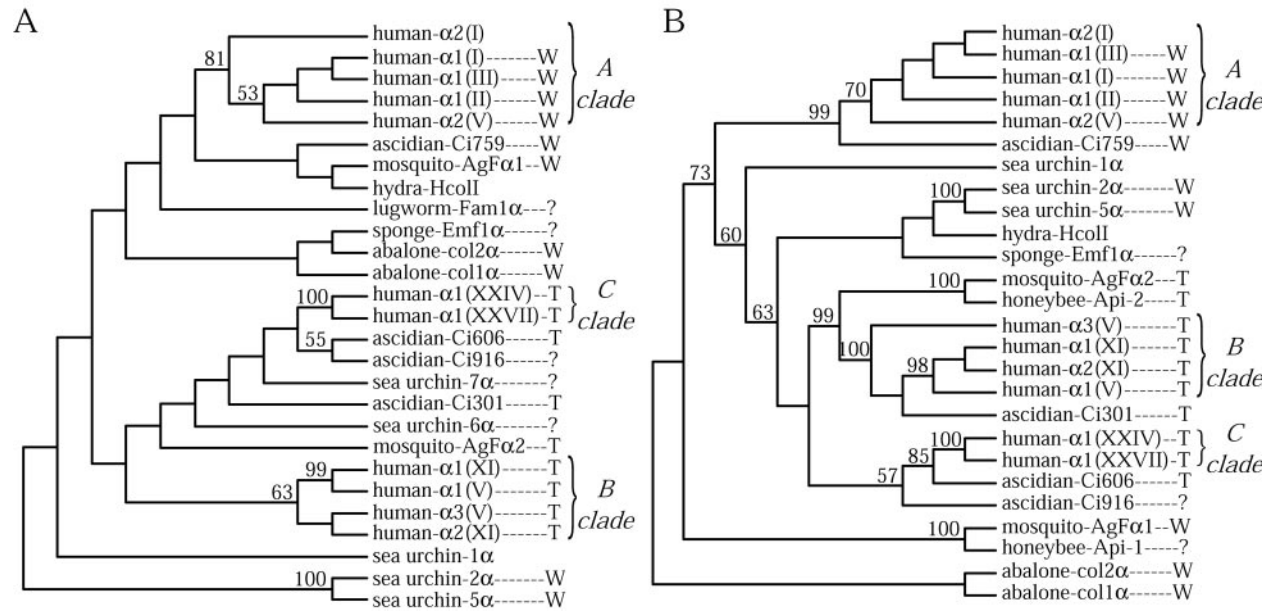

FIG. 4. **Phylogenetic analysis of fibrillar collagens using C-propeptides (*A*) and triple helices (*B*).** C-propeptide and triple helical sequences from human and invertebrates were aligned using the ClustalW program. Ungapped alignments were used to derive trees according to maximum parsimony method using the Phylo win program (23). The bootstrap values at nodes are indicated and represent the percentage of 1000 bootstrap replications. The presence of a vWc (*W*) or TSPN (*T*) module for each α chain is shown at the right of each tree. The *question mark* indicates that the N-propeptide region of the corresponding chain has not been characterized to date. The three vertebrate clades are indicated. See the supplemental data for phylogenetic analyses inferred by the neighbor joining method.

chains are distributed into the three previously defined clades (12). The distribution of human fibrillar α chains into these three clades is corroborated by high bootstrap values at the node of clade separations. However, it is difficult to make any correlation between any one invertebrate α chain and a vertebrate clade. Indeed the α chains that include a vWc or a TSPN module in their N-propeptide seem to be more closely related. The same analysis was done using the triple helical sequences, as shown in Fig. 4*B*, which confirms the assignment of some invertebrate α chains to vertebrate clades as previously indicated from the modular structure of the fibrillar collagen chains (Fig. 1). The robustness of this tree was clearly shown when we indicated which chains included a vWc or a TSPN module in their N-propeptides. Hence, use of the triple helix data permit a clear separation of chains containing a TSPN module from chains including a vWc module. The *C. intestinalis* Ci759, Ci301, and Ci606 α chains are included in groups including the A, B, and C clades, respectively. For Ci916, the maximum parsimony method seems to relate this fibrillar chain to the C clade (Fig. 4*B*) with the C-propeptide analyses (Fig. 4*A*) clustering this α chain with the *C. intestinalis* C clade-like Ci606 fibrillar collagen chain. The *A. gambiae* AgFα1 and AgFα2 chains are related to the A and B clades, respectively. However, the A clade-like property of the AgFα1 chain is only supported by the neighbor joining analysis (see supplemental data) and its modular structure (Fig. 1). To validate the robustness of the triple helix analysis, we also eliminated every third glycine residue from sequences and made the same phylogenetic analyses. The trees made with these sequences were comparable with those presented in Fig. 4*B* (data not shown).

For sea urchin, the data shown in Fig. 4*A* might indicate that 6α and 7α are related to the vertebrate B or C clades. Interestingly, a search of TSPN module in sea urchin genomic servers permitted us to obtain three sequences encoding this module. Multiple alignments of the TSPN modules from invertebrate and human α chains were used to derive trees according to maximum parsimony (Fig. 5) and neighbor joining methods. Phylogenetic analysis made from this alignment confirms that Ci606 is a clade C-like α chain. One of the sea urchin TSPN modules seems to be related to human type IX collagen, whereas the two other TSPN sequences could be part of an N-propeptide.

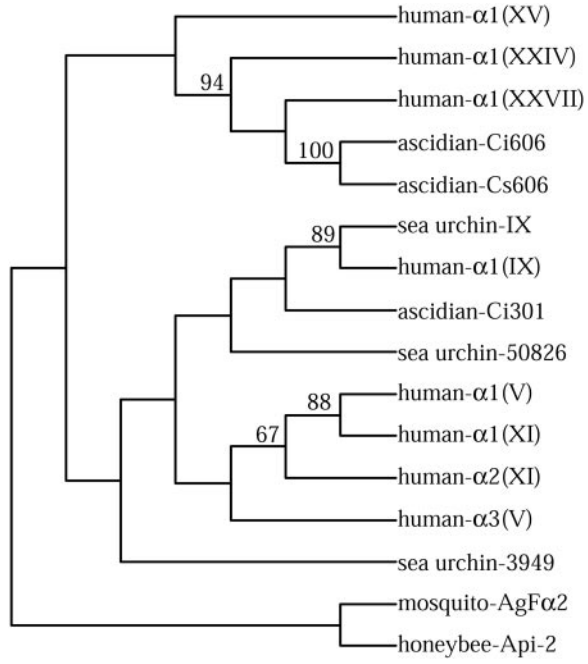

FIG. 5. **Phylogenetic analysis of fibrillar collagens using the TSPN module.** The TSPN module sequences of invertebrate and vertebrate fibrillar α chains were aligned using the ClustalW program. Ungapped alignment was used to construct trees according to the maximum parsimony method. Cs606, the *C. savignyi* ortholog gene of Ci606. The three sea urchin sequences, IX, 3949 ,and 50826 have been deduced from the EST StrPu691.007800, the genomic contig 3949, and the genomic contig 50826, respectively (Table I). See supplemental data for the multiple alignment and for the phylogenetic analysis inferred by the neighbor joining method.

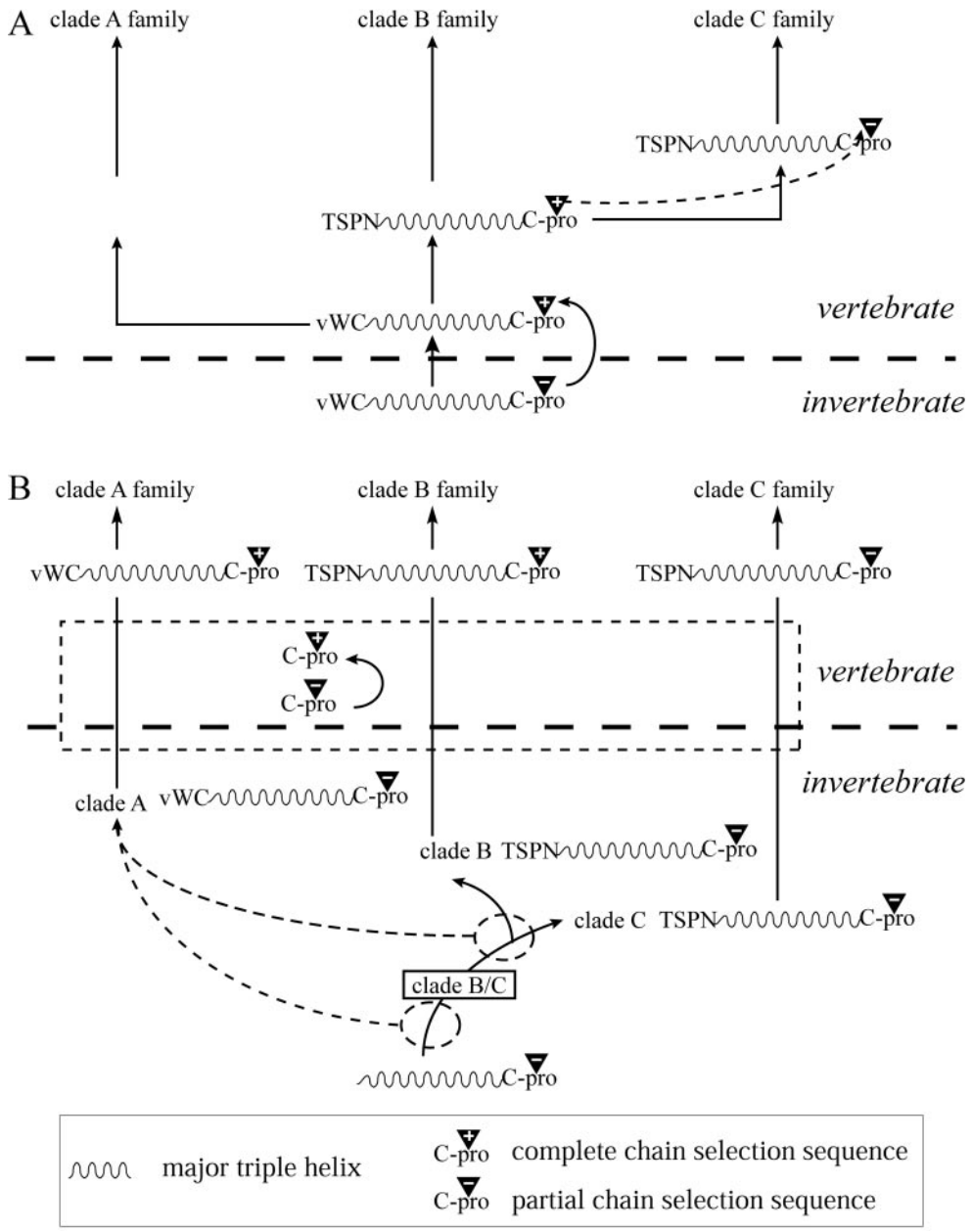

FIG. 6. **Evolution of fibrillar collagens.** The Boot-Handford *et al.* (12) model (*A*) is presented above our evolution model (*B*). Two major differences are found between these models. The first is that the emergence of the three vertebrate clades occurs in invertebrates in our model instead of the divergence of these three clades from one invertebrate α chain in the Boot-Handford model. The second divergence arises from the first. Hence, we consider that the selection chain sequence arose, not from a rare genomic event as in *A*, but instead that the modern selection chain region has been created and improved during evolution (*dashed box*).

The Vertebrate Elongated Chain Selection Sequence

Boot-Handford and Tuckwell (4) have indicated that all invertebrate α chains characterized to date lack a large part of the elongated chain selection sequence. As shown in Fig. 3*A*, with the exception of the ascidian Ci759 chain, all the new invertebrate α chains presented in this work show this deletion and present a short chain selection sequence. Although speculation about the appearance of a long chain selection sequence will be presented later, it is already clear that this sequence is present in a highly divergent region of the C-propeptide (Fig. 3*A*). When we used only human sequences and the *A. gambiae* and *C. intestinalis* α chains related to clades A and B, we could observe some patterns that appeared specific to the A clade (Fig. 3*B*) or B clade (Fig. 3*C*).

DISCUSSION

The data presented in this work clearly indicate that the steps leading to the formation of the fibrillar collagen clades arose early during evolution. Although the presence of α chains related to the A and B clades in insects is enough to justify this statement, evolutionarily studies point out that the divergence of the B and C clades arose before the emergence of vertebrates. Moreover, the data presented in this work argues against the Boot-Handford and Tuckwell model (4) indicating that the long chain selection sequence arose from a rare genomic event. Altogether, these data suggest that the major steps leading to the structural and functional diversity of fibrils are ancient and arose before the emergence of vertebrates.

Exon/Intron Organization of Genes Encoding Fibrillar α Chains—In the genomic structure presented in Fig. 2, it is

interesting to note that only the collagen genes of *A. gambiae* and *A. mellifera* present an original structure in comparison to other species. As for the *Drosophila* type IV collagen gene (27), both *A. gambiae AgF1*α and *AgF2*α are compact (less than 10 kb) and their coding sequences divided into 7 and 9 exons, respectively. These data exclude the sequence encoding the signal peptide. The coding sequences of *A. mellifera* are less compact and include more exons. However, as shown in Fig. 2*B*, several insertions seem to have occurred in these genes. These intronic insertions might explain the less convincing correlation of their exon/intron organization in comparison to other known fibrillar collagen genes. These intronic events might also explain the presence of exons beginning with the second base of a glycine codon instead of an intact glycine codon. With *Ci759*, the two *A. mellifera* fibrillar collagen genes are unique because of the presence of split glycine codons at the 5′ end of some exons. An analogy can be made with the genes encoding the mouse and chick α2(XI) chain, with the presence of split codons in exons 19–24 in chick and intact glycine codons in corresponding exons in mouse (28).

Triple Helix versus C-propeptide in Phylogenetic Analysis—For a long time, the C-propeptide domain has been considered to be the most conserved part of the fibrillar α chains and the region of choice for evolutionary studies. As shown in Fig. 3*A*, the C-propeptide also presents some regions highly variable in size and sequence between all the α chains characterized to date. Hence, for Fig. 4, the average length of the C-propeptide sequence used for each α chain is 168 residues. The maintenance of every third glycine in the major triple helix and the high percentage of proline residues can drive an experimental bias in a phylogenetic analysis. However, the length of the region used for the construction of the phylogenetic tree (906 residues *versus* 168 amino acids for the C-propeptide) gives a more accurate picture of the evolution of this family of proteins. Hence, the bootstrap values are higher than those obtained when C-propeptide sequences are used. The best indication that use of the triple helix in our phylogenetic analysis is justified is that we can predict the modular composition of the N-propeptide of an α chain from the primary structure of its major triple helix.

Emergence of Vertebrate Fibrillar Collagen Clades—During the last decades, several studies have suggested that some invertebrate fibrillar α chains appear to be closely related to vertebrate minor collagens (14, 29, 30). The recently suggested model of Boot-Handford *et al.* (12) was very surprising in our view, especially with regard to the appearance of the B and C clades in vertebrates or just before their emergence. Vertebrates possess quantitatively major (types I-III) and minor (types V/XI) fibrillar collagens. Interestingly, all the α chains that include a TSPN module in their N-propeptide belong to the minor class and are members of the B clade. We previously demonstrated that invertebrates possess quantitatively minor collagens (17), and the availability of genome data from invertebrates clearly shows that they may produce fibrillar α chains including a TSPN domain. The presence of α chains from A and B clades in protostomes clearly argues against the Boot-Handford model (Ref. 12 and Fig. 6*A*). For this reason, we present a new evolutionary model (Fig. 6*B*), which is closely related to our previous model (16). In this model, an ancestral α chain possesses the major triple helix and the C-propeptide. From our evolution study it is difficult to define if the ancestral fibrillar α chain contains a vWc or a TSPN module or either of them in this N-propeptide. The presence of a minor triple helix corresponding to the N-propeptide is possible although uncertain. The divergence of the founder α chains from the A and B clades arose early during evolution as indicated by their presence in protostomes. It is difficult to date the emergence of the C clade during evolution, although the formation of this clade predated vertebrate appearance. The last suggestion of our model is that the formation of the long chain selection sequence is not caused by a rare genomic event but to an improvement during the evolution of this sequence.

Functional and Structural Relevance of Fibrillar Collagen Evolution—From previous invertebrate studies (17, 20, 31) and the present data, it has become apparent that the mechanisms at the origin of vertebrate fibril diversity arose step by step during evolution and are not contemporary with the emergence of vertebrates. Hence, we have shown that invertebrates possess quantitatively major and minor fibrillar collagens (17). Another important step arising in invertebrates is the formation of heterotypic fibrils made of minor and major collagen types displaying distinct maturations of their N-propeptide as demonstrated in sea urchin (17, 20). With the presence of at least two other fibrillar collagen chains (6α and 7α) which seem to be related to the B and C vertebrate clades, the diversity of sea urchin α chains clearly suggests that this invertebrate is able to produce a large variety of fibrils. Indeed, the general mechanisms leading to fibril formation are ancient, as noted in some specializations such as in sea urchin where two α chains contain several repeats of the sea urchin fibrillar module in their N-propeptide (17), a domain that appears to be specific to echinoderms. Moreover, genome duplications have increased the number of fibrillar collagen chains in vertebrates and consequently the diversity of collagen fibrils.

The appearance of vertebrates that arose in the lower Cambrian period and their phenotypic complexity have been attributed to large scale gene or genome duplications ("2R" hypothesis) at the origin of the group (32, 33). The linkage of the A–C fibrillar collagen clades with gene clusters in mammalian genomes is in agreement with the duplication events (2, 24, 25). From this 2R hypothesis and their evolutionary model, Boot-Handford and Tuckwell (4) suggested that the "vertebrate fibrillar collagen family evolved by molecular incest resulting from gene duplications." The formation of the various vertebrate fibrillar collagen types has taken place over several hundred millions of years. During this period, fibrillar collagen genes developed their own pattern of expression and their actual coding sequence. As indicated above, the chain selection sequence is not caused by a rare genomic event but is the result of a long evolutionary process. We would prefer to describe the vertebrate evolution of fibrillar α chains as a non-incest theory or a classical model of the evolution of duplicated genes, *i.e.* from the loss of one of the duplicated genes to the acquisition of new functions that evolved with time. Fibrillar collagens are present from sponges to humans and can be considered as a protein specific to metazoan. Their evolution from the most primitive to the "more evolved" metazoa has followed some general rules from an ancestral gene. This story includes duplications, mutations, and acquisition of new genomic information leading to the formation of the vertebrate clades in invertebrates and introduction of some specialization like the presence of sea urchin fibrillar modules in sea urchins. The availability of genome data from other invertebrates such as the sponge and structural studies of invertebrate fibrils will lead to a better understanding of the evolution and functions of the fibrillar collagens.

REFERENCES

1. Myllyharju, J., and Kivirikko, K. I. (2001) *Ann. Med.* **33,** 7–21
2. Exposito, J. Y., Cluzel, C., Garrone, R., and Lethias, C. (2002) *Anat. Rec.* **268,** 302–316
3. Lees, J. F., Tasab, M., and Bulleid, N. J. (1997) *EMBO J.* **16,** 908–916
4. Boot-Handford, R. P., and Tuckwell, D. S. (2003) *BioEssays* **25,** 142–151
5. Birk, D. E., Fitch, J. M., Babiarz, J. P., Doane, K. J., and Linsenmayer, T. F. (1990) *J. Cell Sci.* **95,** 649–657

6. Linsenmayer, T. F., Gibney, E., Igoe, F., Gordon, M. K., Fitch, J. M., Fessler, L. I., and Birk, D. E. (1993) *J. Cell Biol.* **121,** 1181–1189
7. Blaschke, U. K., Eikenberry, E. F., Hulmes, D. J., Galla, H. J., and Bruckner, P. (2000) *J. Biol. Chem.* **275,** 10370–10378
8. Takahara, K., Hoffman, G. G., and Greenspan, D. S. (1995) *Genomics* **29,** 588–597
9. Sicot, F. X., Exposito, J. Y., Masselot, M., Garrone, R., Deutsch, J., and Gaill F. (1997) *Eur. J. Biochem.* **246,** 50–58
10. Saito, M., Takenouchi, Y., Kunisaki, N., and Kimura, S. (2001) *Eur. J. Biochem.* **268,** 2817–2827
11. Koch, M., Laub, F., Zhou, P., Hahn, R. A., Tanaka, S., Burgeson, R. E., Gerecke, D. R., Ramirez, F., and Gordon, M. K. (2003) *J. Biol. Chem.* **278,** 43236–43244
12. Boot-Handford, R. P., Tuckwell, D. S., Plumb, D. A., Rock, C. F., and Poulsom, R. (2003) *J. Biol. Chem.* **278,** 31067–31077
13. Pace, J. M., Corrado, M., Missero, C., and Byers, P. H. (2003) *Matrix Biol.* **22,** 3–14
14. Exposito, J. Y., and Garrone, R. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87,** 6669–6673
15. Exposito, J. Y., van der Rest, M., and Garrone, R. (1993) *J. Mol. Evol.* **37,** 254–259
16. Exposito, J. Y., Cluzel, C., Lethias, C., and Garrone, R. (2000) *Matrix Biol.* **19,** 275–279
17. Cluzel, C., Lethias, C., Garrone, R., and Exposito, J. Y. (2004) *J. Biol. Chem.* **279,** 9811–9817
18. Exposito, J. Y., D'Alessio, M., Solursh, M., and Ramirez, F. (1992) *J. Biol. Chem.* **267,** 15559–15562
19. Exposito, J. Y., D'Alessio, M., and Ramirez F. (1992) *J. Biol. Chem.* **267,** 17404–17408
20. Cluzel, C., Lethias, C., Humbert, F., Garrone, R., and Exposito, J. Y. (2001) *J. Biol. Chem.* **276,** 18108–18114
21. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410
22. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680
23. Galtier, N., Gouy, M., and Gautier, C. (1996) *Comput. Appl. Biosci.* **12,** 543–548
24. Bailey, W. J., Kim, J., Wagner, G. P., and Ruddle, F. H. (1997) *Mol. Biol. Evol.* **14,** 843–853
25. Smith, N. G., Knight, R., and Hurst, L. D. (1999) *BioEssays* **21,** 697–703
26. Popovici, C., Leveugle, M., Birnbaum, D., and Coulier, F. (2001) *Biochem. Biophys. Res. Commun.* **288,** 362–370
27. Blumberg, B., MacKrell, A. J., and Fessler, J. H. (1988) *J. Biol. Chem.* **263,** 18328–18337
28. Perala, M., Elima, K., Metsaranta, M., Rosati, R., de Crombrugghe, B., and Vuorio, E. (1994) *J. Biol. Chem. 1994* **269,** 5064–5071
29. Miura, S., and Kimura, S. (1985) *J. Biol. Chem.* **260,** 15352–15356
30. Tillet, E., Franc, J. M., Franc, S., and Garrone, R. (1996) *Comp. Biochem. Physiol. B* **113,** 239–246
31. Garrone, R. (1985) in *Biology of Invertebrate and Lower Vertebrate Collagens* (Bairati, A., and Garrone, R., eds) pp. 157–175, Plenum Press, New York
32. Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag, New York
33. Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A., and Sidow, A. (1994) *Development* Suppl., 125–133