RESEARCH ARTICLES

Insights into Early Extracellular Matrix Evolution: Spongin Short Chain Collagen-Related Proteins Are Homologous to Basement Membrane Type IV Collagens and Form a Novel Family Widely Distributed in Invertebrates

Abdel Aouacheria,[1] Christophe Geourjon,† Nushin Aghajari,† Vincent Navratil,* Gilbert Deléage,† Claire Lethias,† and Jean-Yves Exposito†*

*Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard Lyon 1, Villeurbanne, France and †Institut de Biologie et Chimie des Protéines (IBCP), UMR CNRS 5086, Université Claude Bernard Lyon 1, IFR128 BioSciences Lyon-Gerland, Lyon, France

Collagens are thought to represent one of the most important molecular innovations in the metazoan line. Basement membrane type IV collagen is present in all Eumetazoa and was found in Homoscleromorpha, a sponge group with a well-organized epithelium, which may represent the first stage of tissue differentiation during animal evolution. In contrast, spongin seems to be a demosponge-specific collagenous protein, which can totally substitute an inorganic skeleton, such as in the well-known bath sponge. In the freshwater sponge *Ephydatia mülleri*, we previously characterized a family of short-chain collagens that are likely to be main components of spongins. Using a combination of sequence- and structure-based methods, we present evidence of remote homology between the carboxyl-terminal noncollagenous NC1 domain of spongin short-chain collagens and type IV collagen. Unexpectedly, spongin short-chain collagen–related proteins were retrieved in nonsponge animals, suggesting that a family related to spongin constitutes an evolutionary sister to the type IV collagen family. Formation of the ancestral NC1 domain and divergence of the spongin short-chain collagen–related and type IV collagen families may have occurred before the parazoan–eumetazoan split, the earliest divergence among extant animal phyla. Molecular phylogenetics based on NC1 domain sequences suggest distinct evolutionary histories for spongin short-chain collagen–related and type IV collagen families that include spongin short-chain collagen–related gene loss in the ancestors of Ecdyzosoa and of vertebrates. The fact that a majority of invertebrates encodes spongin short-chain collagen–related proteins raises the important question to the possible function of its members. Considering the importance of collagens for animal structure and substratum attachment, both families may have played crucial roles in animal diversification.

Introduction

Basement membranes are sheet-like complexes of extracellular matrix structures underlying epithelial and endothelial tissues and surrounding muscle cells, peripheral nerves, and adipocytes. They play important functions as selective barriers for macromolecules and scaffold support for cells and in cell behavior (Erickson and Couchman 2000). Type IV collagen is one of the major constituents of basement membranes. In humans, a total of 6 type IV collagen chains (α1–α6) have been identified, which are involved in the formation of heterotrimeric molecules with $(\alpha1)_2\alpha2$ being the most abundant and ubiquitous isoform (Hudson et al. 1993). Each type IV chain contains a long triple-helical or "collagenous domain" of approximately 1,400 amino acids flanked by the so-called 7S region and a noncollagenous (NC1) domain at the N- and C-terminus, respectively. The NC1 domain plays crucial roles in the hexameric network assembly of type IV molecules. In particular, NC1 is essential for the selection and association of the 3 type IV α chains and also for the initiation of triple helix formation (Boutaud et al. 2000; Borza et al. 2001; Söder and Pöschl 2004; Khoshnoodi et al. 2006). Triple-helical type IV molecules or "protomers" assemble into a complex network, with NC1 regions from 2 protomers associating to form dimers and 7S domains involved in the formation of tetramers (Timpl et al. 1981). Recent X-ray structures of the NC1 hexamer (Sundaramoorthy et al. 2002; Than et al. 2002) have shed further light on protomer and network assembly. The structure of the NC1 monomer represents a novel 3-dimensional (3D)-fold composed predominantly of β-sheets, which interact through a domain-swapping mechanism. The association of 2 NC1 protomers is favored by extensive hydrophobic and hydrophilic interactions at their interface and is stabilized by a covalent cross-link, termed S-hydroxylysyl-methionine, made by Met and Lys residues contributed by both NC1 trimers (Than et al. 2002, 2005; Vanacore et al. 2005). Much attention has been paid to other biological features of the NC1 monomer, as it is the target of pathogenic antibodies in Goodpasture's syndrome and after transplantation in most patients affected with Alport's syndrome (Hudson et al. 2003). In addition, NC1 proteolytic fragments from type IV collagen chains have potent antiangiogenic and antitumor activities in vivo (Ortega and Werb 2002; Hamano and Kalluri 2005).

Type IV is one of the vertebrate collagens, which shows a wide distribution in invertebrates, from cnidarians to chordates. It has been described in a unique group of sponges, Homoscleromorpha, which presents a basement membrane-like structure (Boute et al. 1996). Homoscleromorpha has been included in the class Demospongiae for a long time. However, from recent phylogenetic analyses, Borchiellini et al. (2004) proposed that Homoscleromorpha may rather form one of the 4 main sponge taxa and should

[1] Present address: Apoptosis and Oncogenesis Laboratory, Institut de Biologie et Chimie des Protéines (IBCP), UMR CNRS 5086, Université Claude Bernard Lyon 1, IFR128 BioSciences Lyon-Gerland, 7, Passage du Vercors, Lyon, France.

Key words: extracellular matrix, basement membrane, spongin, collagen, remote homology, metazoan evolution.

E-mail: jy.exposito@ibcp.fr.

Mol. Biol. Evol. 23(12):2288–2302. 2006
doi:10.1093/molbev/msl100
Advance Access publication August 31, 2006

no more be included in the taxon Demospongiae. Thus, a common morphological character of both Homoscleromorpha and Eumetazoa (nonsponge Metazoa), but not Demospongiae, is the presence of a basal membrane with type IV collagen. Other types of collagens were found in Demospongiae species. A family of collagens including a collagenous domain of approximately 120 Gly-Xaa-Yaa triplets and a carboxy (C)-terminal region sharing some similarities with nematode cuticular collagens and vertebrate fibril-associated collagens with interrupted triple helices has been reported in the sponge *Microciona prolifera* (Aho et al. 1993). In addition, a fibrillar collagen chain and a short-chain collagen family have been described in the freshwater sponge *Ephydatia mülleri* (Exposito and Garrone 1990; Exposito et al. 1991). Genes encoding these 2 collagen families are highly expressed during the early development of sponges from asexual buds (gemmules). In these developing animals, 2 collagen supramolecular structures have been defined, that is, the striated fibrils and the spongins. Like in other animals, fibrillar collagens are involved in the formation of striated fibrils. For spongins, our previous data strongly suggested that they are made, at least in part, by the short-chain collagens (for the sake of simplicity, this sponge short-chain collagen family is termed "spongin short-chain collagens" in this article). Indeed, genes encoding the sponge short-chain collagens are highly expressed in cells located in the epithelial layer and around the inorganic skeleton, these cells being precisely those that secrete spongin (for an ultrastructural analysis, see fig. 7 in Exposito et al. 1991; http://www.jbc.org/cgi/reprint/266/32/21923). In freshwater sponges, these 2 cell types are similar and often join to form a continuous epithelium including the sponge basal surface and ramifying inside the animal body, around the skeleton (fig. 5, ibid). Interestingly, these sponge short-chain collagens also share similarities with nematode cuticular collagens (Exposito et al. 1990, 2002). Spongins, which have been defined as an exoskeleton (Garrone 1984), stick the animal to its substratum, link together the skeletal spicules, and are also present in the coat of gemmules. Although the spongin matrix has been defined as an exoskeleton (Garrone 1984), spongins exhibit different morphological aspects among demosponges and according to the tissues (the term "spongin" initially served to designate sponge structures made of microfibrils of about 10 nm in diameter). To date, it is not known if all spongin assemblies are equivalent (Simpson 1984; Garrone 1985) and whether or not they are entirely made of sponge short-chain collagens. At the molecular level, the spongin short-chain collagens contain 2 collagenous domains encompassing 79 Gly-Xaa-Yaa triplets and 3 noncollagenous domains. Notably, the noncollagenous C-terminal domain has also been observed in 2 proteins of the sponge *Suberites domuncula*, with one of them including a short collagenous domain of 24 Gly-Xaa-Yaa triplets (Krasko et al. 2000; Schröder et al. 2000). At this point, it is important to indicate that from the collagen nomenclature, noncollagenous domains have been named purely on the basis of their position from the C-terminus of the collagen chain, that is, the most C-terminal noncollagenous regions have been defined as NC1 domains although their sequences are often unrelated. In that respect, we previously noticed that spongin short-chain collagen NC1 domain could be divided, like type IV NC1, into 2 similar subdomains sharing ~26% of identity (Exposito et al. 1990). However, except for 2 short regions, similarity between the NC1 domains of these 2 collagen families was not obvious. Now, with the availability of complete genomic sequences and improvements in bioinformatic tools, we examined this resemblance in detail.

Here, we show that a novel protein family related to spongin short-chain collagens is present in invertebrates (except Ecdysozoa), including nonsponge organisms but is undetectable in vertebrates. Evidences from comparison of modular structure, careful examination of primary sequence features, and structural modeling of the NC1 domain of *E. mülleri* spongin short-chain collagen strongly suggest a common origin for spongin short-chain collagen and type IV collagen NC1 domains. Phylogenetic studies show that formation of the bipartite NC1 domain and divergence of the spongin short-chain collagen and type IV collagen families may have occurred early in the evolution of multicellular animals (most probably before the parazoan–eumetazoan split), possibly representing cases of ancient intra- and intergenic duplications in the evolutionary history of Metazoa. We propose that although type IV collagen and spongin short-chain collagen NC1 domains diverged appreciably (across more than 500 Myr of evolutionary time), they are component of modular proteins that most likely subserve related structural (stability of a macromolecular network) and biological (barriers and cellular attachment) functions in Metazoa.

Materials and Methods

Database Searching

Published sequences from sponge and type IV collagen chains were obtained using the Entrez Nucleotide database at National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The NC1 sequences of *E. mülleri* spongin short-chain collagen, spongin short-chain collagen–related proteins, and type IV collagen proteins were used to screen nucleotide databases located at NCBI using TBlastN (Altschul et al. 1997). For the screening of genomes, searches were done using the Ensembl Blast server (http://www.ensembl.org/) and the sea urchin genome server (http://www.ensembl.org/index.html). For the recently completed genome of *Nematostella vectensis* (Sullivan et al. 2006), Blast analysis was carried out using a Nematostella server (http://genome.jgi-psf.org/Nemve1.home.html). Accession numbers, species abbreviations, and sources were compiled in table 1. Hidden Markov Model (HMM) runs were performed with the HMMER package. Various multiple alignments with full-length NC1 domain or subdomain sequences of spongin short-chain collagen (–related) proteins were first constructed using ClustalW, profiles were then built with HMMBuild, and UniProt-SwissProt was searched using the HMMSearch program (at http://pbil.univ-lyon1.fr).

Molecular Modeling

The 3D model of *E. mülleri* spongin short-chain collagen C-terminal domain based on type IV collagen NC1 domain was built by using Geno3D, a comparative

Table 1
Spongin Short-Chain Collagen–Related and Type IV Collagen Sequences From Metazoa

Protein Family	Phylum	Class	Species Name	Accession Number	Abbreviation
Spongin short-chain collagen	Porifera	Demospongiae	*Ephydatia mülleri*	P18503	Emu P185
Spongin short-chain collagen related	Porifera	Demospongiae	*Suberites domuncula*	Q9GV99	Sdo Q9GV
				Q8WP36	Sdo Q8WP
	Cnidaria	Hydrozoa	*Hydra magnipapillata*	DT608116	Hma DT60
				CN625320, CN771131	Hma CN62
	Mollusca	Bivalvia	*Pecten maximus*	DN793433	Pma DN79
	Annelida	Clitellata	*Lumbricus rubellus*	CF798595, …	
	Echinodermata	Echinoidea	*Strongylocentrotus purpuratus*	Scaffold25750	Spu 2575
				CX553667, CX683684	Spu CX55
	Chordata	Ascidiacea	*Ciona intestinalis floridae: Branchiostoma floridae*	BW468897	Cin BW46
		Cephalochordata		BW911746	Bfl BW91
Type IV collagen	Porifera	Homoscleromorpha[a]	*Pseudocorticium jarrei*	Q7JMZ8	Pja Q7JM
	Cnidaria	Hydrozoa	*H. magnipapillata*	DN138061, DN240807	Hma DN13
				DN636820, CF777303	Hma DN63
		Anthozoa	*Nematostella vectensis*	c438003055.Contig1	
				c415700716.Contig2	
	Arthropoda	Insecta	*Anopheles gambiae*	Q7PVR8	Aga Q7PV
				BM588632, BM654242	Aga BM58
			Bombyx mori	BP125709, CK511975	Bmo BP12
				CK522520	Bmo CK52
			Drosophila melanogaster	O18407	Dme O184
				P08120	Dme P081
	Nematoda	Chromadorea	*Caenorhabditis elegans*	P17140	Cel P170
				P17139	Cel P179
	Echinodermata	Echinoidea	*S. purpuratus*	Q26640	Spu Q266
				Q07265	Spu Q072
	Chordata	Ascidiacea	*C. intestinalis*	BW439169	Cin BW43
				BW229076, BW428230	Cin BW22
		Cephalochordata	*B. floridae*	BW844807	Bfl BW84
				BW895761	Bfl BW89
		Aves	*Gallus gallus*	Q919K3	Gga a1
				XP_416952	Gga a2
				AAY43819	Gga a3
				XP_422615	Gga a4
				XP_420320	Gga a5
				XP_420322	Gga a6
		Mammalia	*Mus musculus*	NM_009931	Mmu a1
				NM_009932	Mmu a2
				NM_007734	Mmu a3
				NM_007735	Mmu a4
				NM_007736	Mmu a5
				NM_053185	Mmu a6
			Homo sapiens	P02462	Hsa a1
				P08572	Hsa a2
				Q01955	Hsa a3
				P53420	Hsa a4
				P29400	Hsa a5
				Q14031	Hsa a6

[a] As defined by Borchiellini et al. (2004).

molecular modeling program for proteins (Combet et al. 2002). Protein structure of type IV collagen NC1 domain (PDB code 1li1-A, α1 chain) was taken as template for molecular modeling. Sequence alignment of spongin short-chain collagen NC1 domain based on type IV collagen proteins was validated by using phylogenetic and predicted secondary structure information (Geourjon et al. 2001). On the basis of this alignment, distance restraints and dihedral angles were calculated on the template structure. These measurements were performed for all common atoms revealed by alignment of spongin short-chain collagen NC1 domain with the templates. The CNS 1.1 program (Brünger et al. 1998) was used to generate the model by a distance geometry approach similar to that used in modeling from nuclear magnetic resonance experiments. Each structure was regularized by simulated annealing (2,000 steps) and energy minimization (2,000 steps). Ten models were built, all exhibiting closely similar features, and superimposed with the ANTHEPROT 3D package by minimizing the root mean square deviation between α carbons (Geourjon and Deleage 1995). Mirror images were eliminated on the basis of energy calculation. The model retained was that with the lowest energy (−6938 Kcal/mol) and regular chemical features, and its quality was assessed with the PROCHECK tools (Laskowski et al. 1993) (90% residues are located in the favorable region of the Ramachandran plot). These values were consistent with all 10 models. Molecular pictures were drawn with PyMOL (DeLano 2005). For

construction of a structural model based on the type IV collagen NC1 hexamer, 6 copies of the monomer model of spongin short-chain collagen were created, and each monomer was superimposed onto the molecules forming the hexamer in the crystal structure of type IV collagen NC1 hexamer. The superimpositions were performed using the "DaliLite" program from the Dali server (http://www.ebi.ac.uk/dali/Interactive.html). Neither the modeled trimer nor the hexamer model were subjected to molecular dynamics simulations, reasons being 1) the relatively low sequence similarity between the spongin short-chain collagen and type IV collagen NC1 domains and 2) the relatively low amount of secondary structure in the model of spongin short-chain collagen which in itself is a direct consequence of (1).

Alignment and Evolutionary Analysis

Sequences of type IV collagen and spongin short-chain collagen–related NC1 domains (either separately or in combination) were first aligned using ClustalW (Thompson et al. 1994) with BLOSUM alignment matrices and adjusted gap penalties (at the Pole BioInformatique Lyonnais). The resulting initial alignments were scanned using RASCAL (Thompson et al. 2003) and manually improved using the SeaView alignment editor (Galtier et al. 1996). When possible, structural information was incorporated in order to improve alignment accuracy. The alignments were constructed in a 2-stage manner: 1) alignments of complete NC1 domains were first produced [subdomain a plus subdomain b] and 2) the stretches corresponding to the different subdomains were separated, and the 2 resulting alignments were aligned together using information from the consensus sequences (subdomain a over subdomain b). Neighbor-joining (NJ or BIONJ) and maximum likelihood (ML) analyses were performed on the final alignments. For NJ, the trees were made using Phylo_win (Galtier et al. 1996) with pairwise gap removal, 1,000 bootstrap repetitions, and observed divergence or Poisson correction as distance methods. The PHYML v2.4.4 algorithm (Guindon and Gascuel 2003) was applied for the ML analyses, under the JTT or Dayhoff model of sequence evolution. Bootstrap support was based on 100 replicates using the programs SEQBOOT and CONSENSE (majority rule extended) of the PHYLIP package (Felsenstein 1996), to generate data replicates and consensus tree, respectively. Illustrations were drawn using the TreeView program (Page 1996) and then annotated using Adobe Illustrator. Number of synonymous (*K*s) and nonsynonymous (*K*a) nucleotide substitutions per site between homologous DNA sequences were estimated using an ML method as implemented in the codeml program (Goldman and Yang 1994). For each human–mouse and human–chicken orthologous gene pairs, cDNA sequences were aligned in accordance with pairwise amino acid alignments.

Results

Type IV Collagen and Spongin Short-Chain Collagen-Related NC1 Domains Display Distinct Phyletic Distribution but Share Similar Primary Structure Features

With the initial aim of searching proteins related to sponge short-chain collagens in Porifera, we mined public databases with Blast using short-chain collagen NC1 sequence from the freshwater sponge *E. mülleri* as seed. This search led to the discovery of cDNAs encoding putative spongin short-chain collagen–related proteins in *S. domuncula* and, quite unexpectedly, in a number of protostomes with the notable exception of Ecdysozoa, and in invertebrate deuterostomes (table 1). The same analysis carried out with ecdysozoan (drosophila, mosquito, nematodes, and honeybee) or vertebrate (tetraodon, zebrafish, chicken, and human) genomes confirmed the absence of spongin short-chain collagen–related sequences in these animals, using this approach. Use of the NC1 sequences of the 2 spongin short-chain collagen–related proteins from *S. domuncula* or from the newly identified spongin short-chain collagen–related proteins gave the same result. In addition, HMM search against Swiss-Prot using profiles built with complete NC1 domains of spongin short-chain collagen–related proteins recovered only the *E. mülleri* spongin short-chain collagen sequence (P18503).

Our previous work revealed that, intriguingly, spongin short-chain collagen and type IV collagen NC1 domains exhibit a same bipartite architecture and regions with local similarities (Exposito et al. 1990). Indeed, it appears clearly from the schematic view presented in figure 1 that spongin

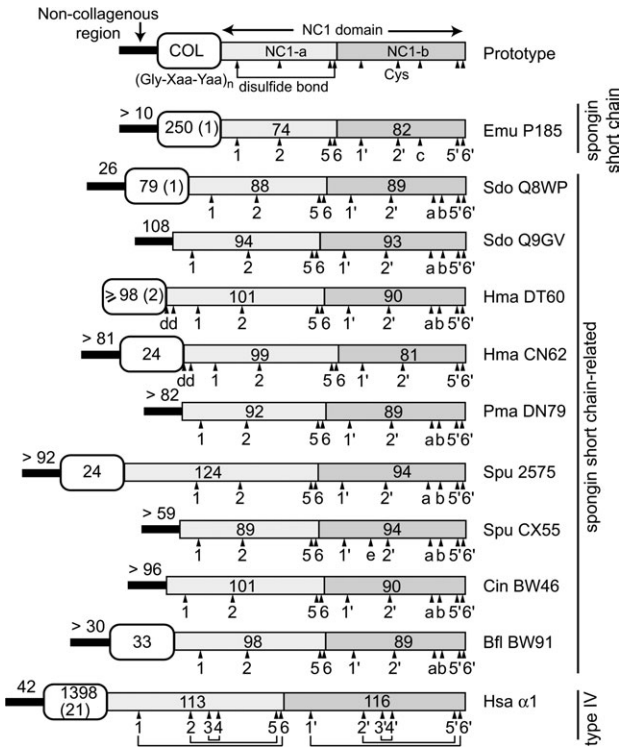


FIG. 1.—Modular organization of *Ephydatia mülleri* spongin short-chain collagen, spongin short-chain collagen–related proteins, and human α1(IV) collagen. The general domain architecture of spongin short-chain collagen–related proteins and type IV collagens is depicted. A summary of the different regions and structural motifs is shown at the top in the form of a prototypal protein. The spongin short-chain collagen–related proteins depicted are those for which complete NC1 information is available. NC1 domains are drawn to scale. Bolded Arabic numbers indicate the length in amino acids of the different regions. Arabic numbers in parentheses correspond to the number of interruptions within the collagenous domains (COL). The critical cysteine residues are represented by black triangles and labeled with letter or Arabic numbers according to their position in multiple sequence alignments.

Table 2
Threading Analysis of Spongin Short-Chain (related) Collagen Sequences

Sequence	Species Name	3D-PSSM[a] *E* Value (1li1 rank)	mGEN Threader[b] *E* Value (1li1 rank)	FUGUE[c] *Z* Score (1li1 rank)
Emu P185	*Ephydatia mülleri*	0.455 (1)	3.26 (2)	4.97 (1)
Sdo Q8WP	*Suberites domuncula*	0.129 (1)	0.163 (1)	4.45 (1)
Sdo Q9GV	*S. domuncula*	0.453 (1)	0.512 (1)	3.31 (1)
Hma DT60	*Hydra magnipapillata*	0.307 (1)	1.593 (3)	4.39 (1)
Hma CN62	*H. magnipapillata*	**0.0465** (1)	0.069 (1)	**16.22** (1)
Pma DN79	*Pecten maximus*	**0.0199** (1)	0.047 (1)	5.05 (1)
Spu CX55	*Strongylocentrotus purpuratus*	**0.0359** (1)	0.028 (1)	5.14 (1)
Spu 2575	*S. purpuratus*	0.0585 (1)	0.103 (1)	5.07 (1)
Cin BW46	*Ciona intestinalis*	**0.028** (1)	0.194 (1)	4.4 (1)
Bfl BW91	*Branchiostoma floridae*	**0.0116** (1)	**0.006** (1)	5.58 (1)
Positive control				
Pja Q7JM	*Pseudocorticium jarrei*	5.42×10^{-39} (1)	2×10^{-5} (1)	45.09 (1)

NOTE.—1li1, PDB file name of crystal structure of hexameric type IV collagen NC1 domain.
[a] *E* values below 0.05 are highly confident. *E* values up to 1.0 are worthy of attention.
[b] Confidence levels: certain, *E* value < 0.001; high, *E* value < 0.01; medium, *E* value < 0.1.
[c] *Z*-score ≥ 6.0 (certain, 99% confidence level), *Z*-score ≥ 4.0 (likely, 95% confidence level). Highly significant scores are in bold.

short-chain collagen (–related) and type IV collagen NC1 domains display similar lengths, have conserved cysteine residues, and are equally subdivided into 2 presumably homologous subdomains. Moreover, like in the sponge *S. domuncula*, other spongin short-chain collagen–related proteins can possess a collagenous region including several Gly-Xaa-Yaa triplets. The different NC1 domains have not been found in combination with other known protein domains. Thus, members of the spongin short-chain collagen–related and type IV collagen protein families could include a collagenous region in addition to a NC1 domain, indicating that they might have homeomorphic evolutionary relationships. We wondered whether spongin short-chain collagen–related proteins would be retrieved using type IV collagen NC1 sequences as seeds in Blast searches. This analysis confirmed the presence of type IV collagen in the sponge class Homoscleromorpha and in all eumetazoan lineages (table 1) but failed to recover any spongin short-chain collagen–related sequence. These data suggested that spongin short-chain collagen–related and type IV collagen NC1 domains were too distantly related to be detected by reciprocal Blast searches.

Type IV Collagen and Spongin Short Chain Collagen NC1 Domains Display Structural Similarities

Secondary Structure Predictions and Threading Experiments

Threading methods are 3D-structure prediction techniques that can reveal more distant relationships than conventional sequence-based methods such as Blast. We decided to take advantage of the solved structures of type IV collagen NC1 domain (Sundaramoorthy et al. 2002; Than et al. 2002) to predict whether spongin short-chain collagen–related sequences can adopt a similar fold. To this end, we used a battery of 3D-1D–fold recognition programs, including 3D-PSSM (Kelley et al. 2000), mGenThreader (Jones 1999) and FUGUE (Shi et al. 2001). The major result of these analyses was that the best scores were observed with 1li1, that is, the PDB code corresponding to the crystal structure of the human type IV collagen NC1 $[(\alpha 1)_2\alpha 2]_2$ hexamer structure (table 2). The best result was obtained for the hydra sequence Hma CN62 with the FUGUE analysis system at the 99% confidence level, indicating a remarkable compatibility of the secondary structures (*Z*-Score of 16.22; table 2). For *E. mülleri* spongin short-chain collagen NC1, FUGUE gave the best result with 1li1 at the 95% confidence level. We also used the tissue inhibitor of metalloproteinase (TIMP-1) sequence as input because a putative structural link between the type IV collagen NC1 domain and TIMP-1 was previously proposed (Netzer et al. 1998). The threading methods used in this work failed to detect any relationships between TIMP and type IV collagen NC1 domain. Taken together, the threading data indicated a substantial degree of compatibility between the query sequences (spongin short-chain collagen–related NC1 domains) and the type IV collagen NC1 structural fold. As shown in fig. S1 (Supplementary Material online), there was a correspondence between the actual secondary structures of type IV collagen NC1 domain and the predicted secondary structure of the *E. mülleri* spongin short-chain collagen NC1 domain, the regions of structural similarities including most of the β-strands. These findings suggest that, despite wide differences in aminoacid sequences (~16% of identity between spongin short-chain collagen and human α1(IV) NC1 domains; table S1, Supplementary Material online), the NC1 domains of spongin short-chain collagen–related and type IV collagens may have similarities in their 3D structures.

Spongin Short-Chain Collagen NC1 Model Construction and Analysis

On the basis of the threading results and 2D predictions, we attempted to model the *E. mülleri* spongin short-chain collagen NC1 domain using 1li1 as template. A structural model of the spongin short-chain collagen NC1 monomer is presented in figure 2*A*, whereas the X-ray derived structure of a type IV collagen NC1 monomer is shown in figure 2*B*. The first observation that could be made is that the β-strands located near the triple-helical junction are clearly retrieved in the spongin short-chain collagen model. This suggests that this ordered region is likely to be rigid, a prerequisite for the initiation of a quaternary

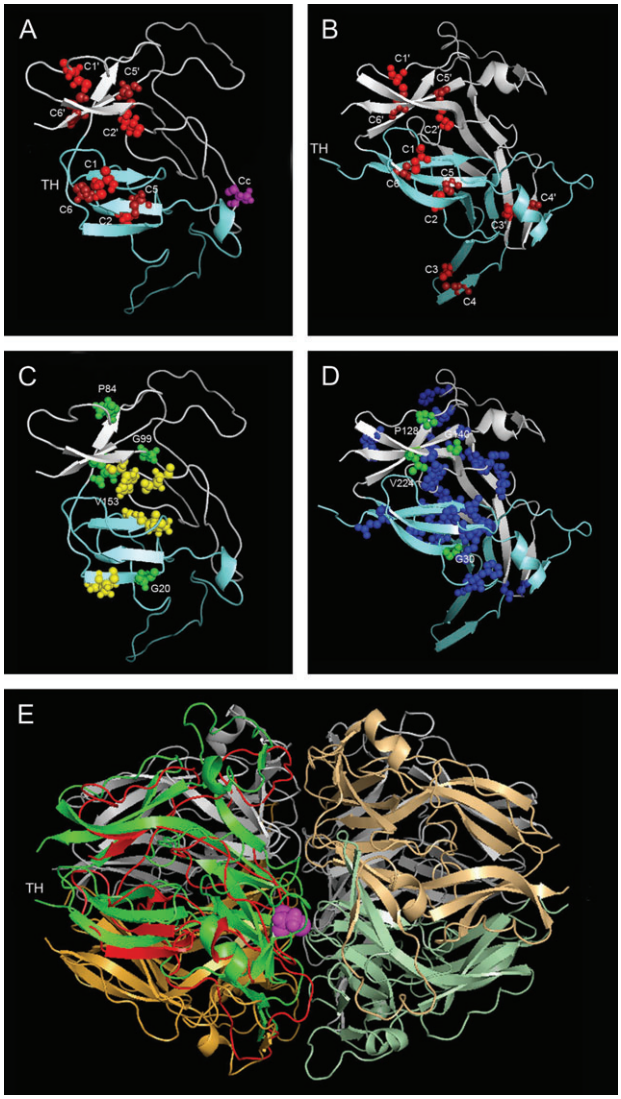


FIG. 2.—Homology-derived model of *Ephydatia mülleri* spongin short-chain collagen NC1 domain. Ribbon diagram of *E. mülleri* spongin short-chain collagen (*A* and *C*) and human α1(IV) collagen (*B* and *D*) NC1 domains. The spongin short-chain collagen NC1 domain has been modeled on the crystal structure (1li1-A) of human α1(IV) collagen NC1 domain. NC1 subdomains a and b are colored in cyan and white, respectively. Position of the triple helix is indicated (TH). Conserved cysteines in each domain (small balls) are colored in red, and residue numbers are indicated (*A* and *B*). Residues that were found to be conserved within the spongin short-chain collagen–related subfamily (*C*, yellow balls), type IV collagen NC1 domains (*D*, blue balls), and between spongin short-chain collagen–related and type IV NC1 domains (*C* and *D*, green balls) are marked. Ribbon plot view of the type IV collagen NC1 hexamer down the 2-fold pseudoexact axis is shown in *E*. Type IV NC1 monomers are colored green (α1A), orange (α1B), and gray (α2) in each individual protomer. Chains A, B, and C make up the left-sided trimer, whereas chains D, E, and F compose the right-sided trimer. The spongin short-chain collagen modeled chain (red) was superimposed with α1(IV) collagen NC1 (green). The protomer–protomer interface is in the longitudinal plane. The "orphan" cysteine residue present within the b subdomain of spongin short-chain collagen NC1 is colored in magenta (*A* and *E*). The figure was made with PyMol.

structure where NC1 trimers are expected to be attached to a rope-like triple helix. Sequence conservation information derived from the complete multiple alignments were mapped onto the spongin short-chain collagen NC1 model and the type IV collagen NC1 structure (fig. 2*C*and *D*). Apart from cysteine residues (addressed below and fig. 2*A* and *B*), 9 and 37 conserved residues were observed within the spongin short-chain collagen–related and type IV collagen sequence clusters, respectively, and 4 amino acids were perfectly conserved between spongin short-chain collagen–related and type IV collagen NC1 domains. Noteworthy, in the structural context, the NC1 residues conserved in spongin short-chain collagen (fig. 2*C*, yellow), type IV collagen (fig. 2*D*, blue), and both sequences (fig. 2*C* and *D*, green) are mostly located at the proximity of the triple-helical junction region. This well-conserved region between spongin short-chain collagen and type IV collagen NC1 domains corresponds in type IV collagen to the β-sheet I, which is formed by the 3 noncontiguous strands (β1, β10, and β2) in both NC1 subdomains (Sundaramoorthy et al. 2002). Thus, our structural model is informative as to 1) the possible homology between spongin short-chain collagen and type IV collagen NC1 domains and 2) highly conserved residues that are probably critical to NC1 domain function.

Next, the occurrence and relative positions of cysteine residues were investigated in the different NC1 domains. Similar cysteine residues within the type IV collagen NC1-a and NC1-b subdomains were named C1-C6 and C1′-C6′, respectively (fig. 1). In type IV collagen NC1, each subdomain is stabilized by 3 intrachain disulfide bonds involving the following pairs: C1-C6, C2-C5, and C3-C4 (Siebold et al. 1988). Moreover, it has been shown that the 2 NC1 prominent regions involved in chain selection are a β-hairpin including the cysteines C3 and C4 and the hypervariable region VR3 (Khoshnoodi et al. 2006 and fig. 3) located between the cysteine residues C4′ and C5′. Notably, the C3-C4 and C3′-C4′ pairs are absent in all the spongin short-chain collagen–related sequences. At the same time, 8 of the 10 spongin short-chain collagen–related sequences displayed 2 additional cysteine residues (Ca and Cb) in their NC1-b subdomain (fig. 1) in a region analogous to VR3. In type IV collagen NC1 protomers, the β-hairpin structure from each monomer swaps into a 4-stranded antiparallel β-sheet from a flanking NC1 domain to form a stable interchain contact (Sundaramoorthy et al. 2002; Than et al. 2002). This swapping motif that plays an important role in the stabilization of type IV collagen NC1 trimers is well conserved between the type IV collagen chains at the sequence level (Khoshnoodi et al. 2006). In contrast, the analogous region in spongin short-chain (–related) collagens shows great variability (fig. 3). Based on a spongin short-chain collagen NC1 hexamer model, monomers that constitute putative protomers are entangled into one another at zones implicated in domain swapping in the crystal structure of type IV collagen NC1 hexamer (fig. S2, Supplementary Material online). Moreover, monomer contact surfaces within the modeled trimer seem possible in that the rather few electrostatic interactions are not of repelling order (fig. S2A, Supplementary Material online).

In type IV collagen NC1 hexamers, hydrophobic and hydrophilic interactions stabilize the protomer–protomer interface. Moreover, it has been shown that in the $[(\alpha1)_2\alpha2]$ mammalian type IV collagen network, the stability of the NC1 hexamer might be reinforced by a covalent cross-link involving the NC1 residues Met^{93} and Lys^{211}

contributed by both protomers (Than et al. 2002, 2005; Vanacore et al. 2005). According to the multiple sequence alignments, all the type IV collagen chains of bilaterian animals possess Met and Lys residues at similar positions (fig. 3). Sponge and hydra (Cnidaria, Hydrozoa) type IV chains lack such residues. However, given their presence in both *N. vectensis* (Cnidaria, Anthozoa) type IV collagen chains, it is most likely that a type IV collagen chain harboring equivalent Met and Lys residues was encoded by a common ancestor of Cnidaria and Bilateria. *Ephydatia mülleri* spongin short-chain collagen and spongin short-chain collagen–related chains also lack these 2 amino acids. Absence of the Met-Lys residues in type IV NC1 chains from sponges and in spongin short-chain collagen–related sequences could either suggest that the corresponding NC1 protomers are assembled into a less stable quaternary structure or that alternative mechanisms exist to stabilize a putative hexamer. In that respect, it is intriguing to note that the "orphan" cysteine residue (Cc) of E. *mülleri* spongin short-chain collagen is predicted to lie at the exterior of the NC1 monomer, facing the putative interface between NC1 trimers, raising the possibility that this residue might be involved in covalent cross-connections between 2 NC1 "protomers" (fig. 2*A* and *E* and fig. S2B, Supplementary Material online). Based on the spongin short-chain collagen NC1 hexamer model, both trimers are within a "realistic" distance of each other, and we observed that a slight rotation of one of the spongin short-chain collagen NC1 trimers around a 3-fold axis perpendicular to the hexamer interface positioned the orphan cysteine residues so that they face one another, a feature which does not seem to be fortuitous.

However, it should be kept in mind that the structural model might not reflect the exact position of the cysteine residues within the actual spongin short-chain collagen (–related) NC1 domain. More generally, great caution should be taken in interpreting these results obtained by comparative protein modeling, due to the low similarity between spongin short-chain and type IV collagen NC1 domains.

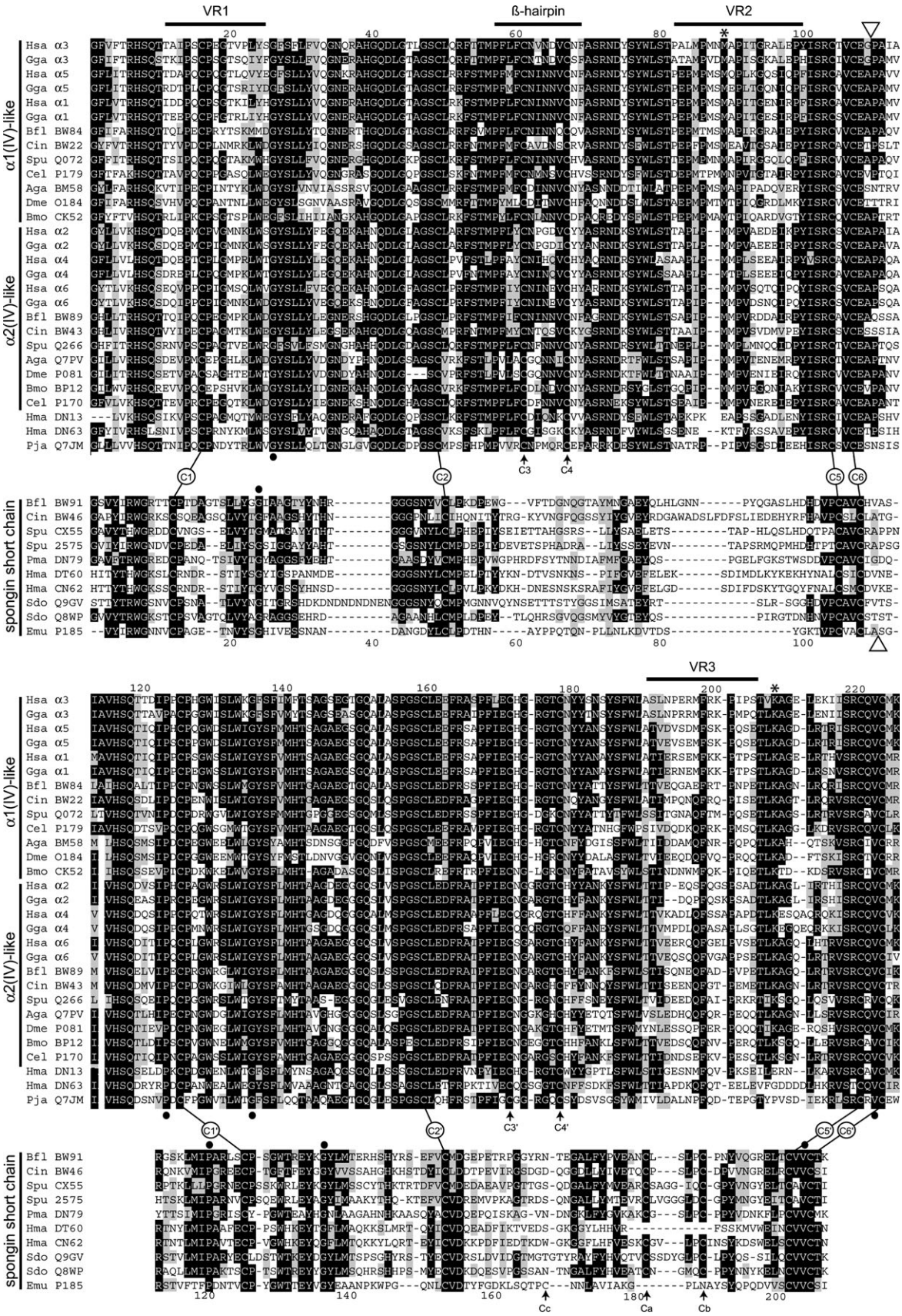
Phylogenetic Analysis

Comparison of modular organization, as well as conservation of critical residues and modeling data, provides strong evidence that spongin short-chain collagen and type IV collagen NC1 domains are structurally related and presumably share a common ancestor. Because spongin short-chain collagen–related and type IV collagen NC1 domains could reasonably be considered as homologous, multiple alignments were used as input for phylogenic analyses using NJ and ML methods. Monophyly of the type IV collagens was extremely well supported in all analyses, as well as the grouping of *E. mülleri* spongin short-chain collagen and spongin short-chain collagen–related sequences. Sequences from sponges were usually retrieved at the basis of the spongin short-chain collagen–related and type IV collagen groups (figs. 4, 5*A*, and 6). Hence, ancestral type IV collagen and spongin short-chain collagen–related NC1 domains must have arisen very early during metazoan evolution and diverged before separation of the poriferan and cnidarian lineages. As spongin short-chain collagen and type IV collagen may be ancient paralogues, we were interested in determining the evolutionary relationships within both protein families.

As previously shown (Mariyama et al. 1992; Netzer et al. 1998), our phylogenetic analyses indicate that type IV collagens are divided into 2 subfamilies termed α1-like (α1, α3 and α5 in vertebrates) and α2-like (α2, α4 and α6 in vertebrates) (figs. 4 and 5*A*). As one of the 2 type IV collagen sequences from Hydra (Hma DN13) could not be unambiguously placed in the different trees, it is unclear at this stage whether the α1-like/ α2-like duplication already took place in this organism or if the emergence of the 2 type IV subfamilies occurred after the Cnidaria–Bilateria split. Hydra might also possess an as yet undiscovered α2-like chain or have lost the corresponding gene. In this regard, it is important to note that the type IV collagen chains of the sea anemone *N. vectensis*, which lies at the basis of the Cnidaria, segregated with the sequences of *Hydra magnipapillata*, disfavoring the hypothesis of a third, α2-like gene, in Cnidaria (data not shown). Although supported by low bootstrap values, segregation of the ecdysozoan type IV collagen sequences inside the α1-like and α2-like groups was the most frequently retrieved tree topology (figs. 4 and 5*A*). Although the type IV α2-like NC1 sequence from *Caenorhabditis elegans* segregates with that of arthropods (figs. 4 and 5*A*), forming a clear ecdysozoa group, the nematode α1-like sequence (Cel P179) segregates with that of Ciona (Cin BW22). This may be due to the high divergence rate reported for nematode and ciona genes in general compared with other species (Mushegian et al. 1998; Holland and Gibson-Brown 2003). Alternatively, this may be indicative of faster divergence rates for α1-like sequences in arthropods (that produce longer branches compared with the α2-like cluster, see fig. 5*A*). Type IV collagen gene diversification has occurred later, in the early evolution of vertebrates, most probably after their divergence with cephalochordates (6 genes were identified in *Tetraodon nigroviridis*, whereas only 2 genes were found in amphioxus). Previous studies have shown that, in mammals, the *col4a1*/*col4a2*, *col4a3*/*col4a4*, and *col4a5*/*col4a6* gene pairs were located on 3 different chromosomes in a head-to-head fashion (Hudson et al. 1993). Based on this atypical genomic organization and on sequence homologies among the various chains, it has been suggested that α3 evolved before the duplication resulting in the α1/α5 pair in the α1-like cluster, and that duplication of an ancestral α4 gene predated the

→

FIG. 3.—Representative multiple alignments of type IV collagen and spongin short-chain collagen–related NC1 domains. The alignments were generated with ClustalW and viewed with Boxshade (http://bioweb.pasteur.fr/docs/softgen.html#BOXSHADE). Black boxes denote sequence identity (threshold = 0.50), whereas gray boxes refer to conservative changes. Dashes indicate gaps. All sequences were from this study. The abbreviations used are presented in table 1. For type IV collagen NC1 domains, position of the variable regions (VR1-3) and of the β-hairpin are indicated at the top of the amino acid sequence alignment; conserved cysteines are numbered as in figure 1. The limits of the subdomains are indicated above the alignments (open arrowhead). The internal homology in the primary structure of spongin short-chain collagen–related and type IV collagen NC1 domains is apparent.


VR1
ß-hairpin
VR2
VR3
α1(IV)-like
α2(IV)-like
spongin short chain
Hsa α3
Gga α3
Hsa α5
Gga α5
Hsa α1
Gga α1
Bfl BW84
Cin BW22
Spu Q072
Cel P179
Aga BM58
Dme O184
Bmo CK52
Hsa α2
Gga α2
Hsa α4
Gga α4
Hsa α6
Gga α6
Bfl BW89
Cin BW43
Spu Q266
Aga Q7PV
Dme P081
Bmo BP12
Cel P170
Hma DN13
Hma DN63
Pja Q7JM
Bfl BW91
Cin BW46
Spu CX55
Spu 2575
Pma DN79
Hma DT60
Hma CN62
Sdo Q9GV
Sdo Q8WP
Emu P185
C1
C2
C3
C4
C5
C6
C3'
C4'
Cc
Ca
Cb

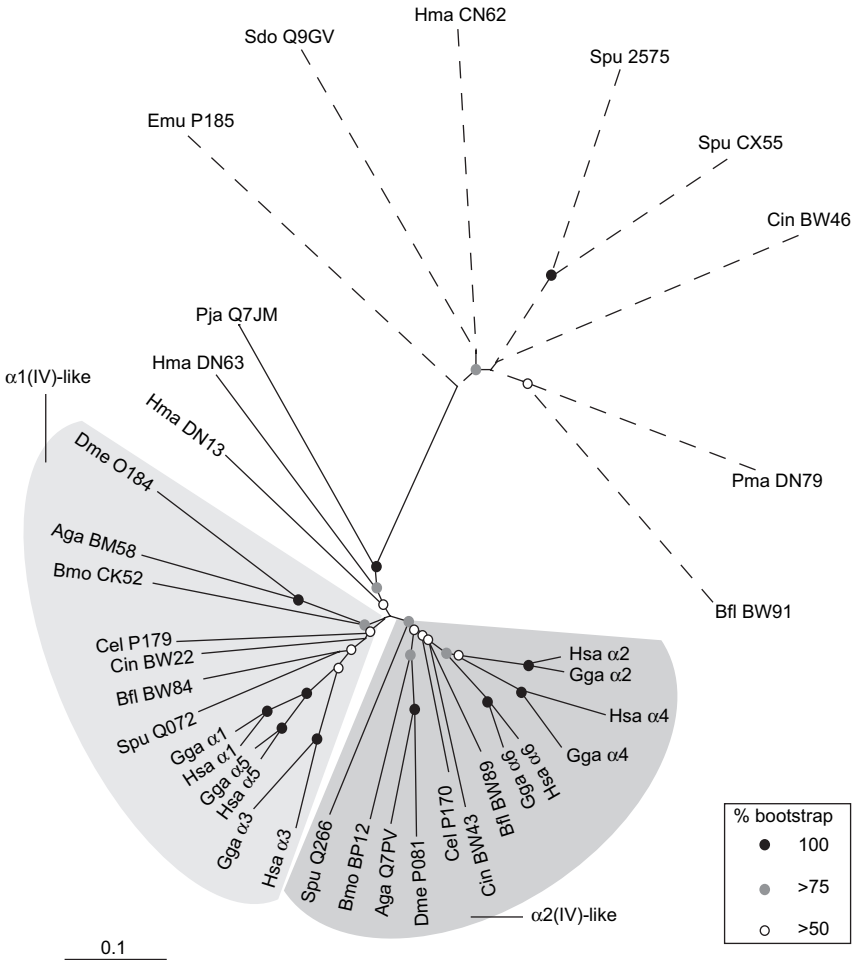


FIG. 4.—Unrooted NJ tree of spongin short-chain collagen–related proteins and type IV collagens. The tree was inferred by the NJ method from comparison of NC1 domain sequences of spongin short-chain collagen–related and collagen IV proteins (198 informative sites). Gap sites were excluded from the analysis. The clusters corresponding to the α1-like and α2-like collagen subfamilies are shaded. The bootstrap values at nodes represent the percentage of 1,000 bootstrap replications. Boostrap probabilities higher than 50% are illustrated as indicated in the figure.

divergence of α2 and α6 in the α2-like clade. Inspection of the chromosomal location of type IV collagen genes in *Gallus gallus* revealed identical pairing. Our phylogenetic reconstruction using chicken and human orthologous chains unambiguously placed α3 at the basis of the vertebrate α1-like cluster, but α6 sequences were often retrieved basal to the α2-like cluster, demonstrating phylogenetic incongruence (see figs. 4 and 5*A* for instance). An NJ analysis (fig. 5*B*) carried out with a reduced multiple alignment including vertebrate sequences from *G. gallus*, *Mus musculus*, and *Homo sapiens* produced a robust tree with a topology congruent with the proposed phylogenetic scheme. It is noteworthy that a significantly higher *K*a/*K*s ratio (table S2; Supplementary Material online) was found in a chicken–human NC1 comparison for *col4a3* (0.11), compared with the median value for genes located in intermediate chromosomes (0.052) and, unexpectedly, compared with its neighboring gene *col4a4* (0.045). Interestingly, this chicken α3 chain that shows evidence of relaxation from purifying selection already evolved autoimmune epitopes as it is recognized by Goodpasture autoantibodies (MacDonald et al. 2006). The situation is repeatable in a human–mouse comparison, with the α3 gene being the least constrained, although in this case the *K*a/*K*s ratio was not increased more than expected. Interestingly, "disease genes" have been reported to evolve with higher *K*a/*K*s ratio (Smith and Eyre-Walker 2003). Nevertheless, it is important to indicate that overall, type IV collagen genes display remarkably low *K*a/*K*s values (mean *K*a/*K*s ratio of type IV collagen genes are 2- to 3-fold less compared with other secreted domains or "metazoan-specific" genes), which is indicative of strong purifying selection (table S2; Supplementary Material online). The α1/α2 pair, which corresponds to the ubiquitously expressed collagen IV chains, exhibited the lowest *K*a/*K*s ratio in both interspecific comparisons. This finding is consistent with previous data reporting stronger selective constraints for housekeeping and broadly expressed genes (Duret and Mouchiroud 2000; Zhang and Li 2004). Notably, human α1 and α2 type IV collagen NC1 domains display more than 75% similarity in amino acids with their *Pseudocorticium jarrei* homologues, illustrating the substantial conservation of type IV collagen NC1.

An NJ tree showing the possible interrelationships between the available spongin short-chain collagen–related sequences (fig. 6) suggest recent duplications of spongin

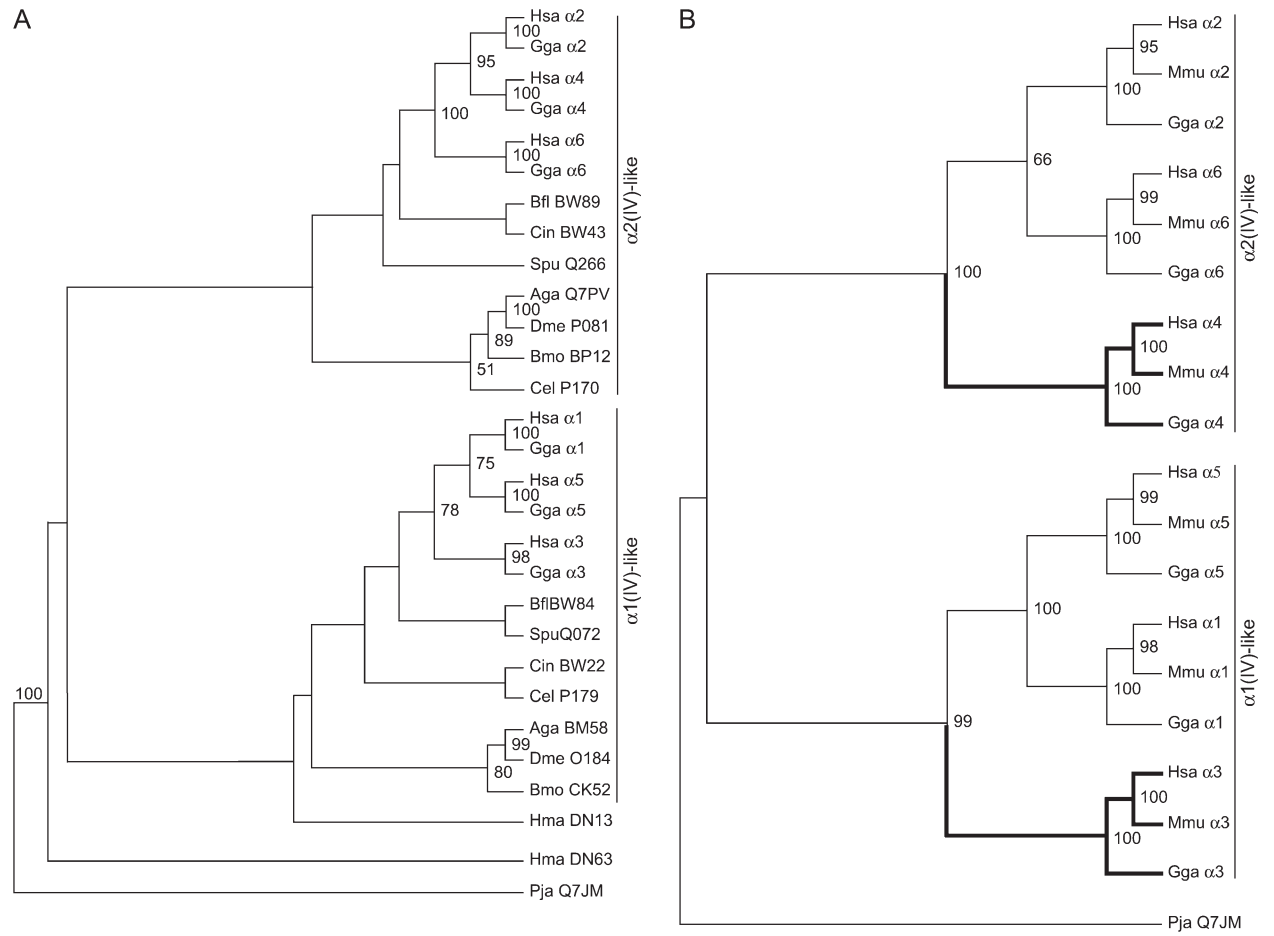


FIG. 5.—Phylogenetic analysis of type IV collagens. (*A*) Maximum likelihood tree of the type IV collagen subfamily inferred from NC1 domain sequences. The tree was inferred by an ML method based on an alignment of 226 amino acids in length (gap sites were ignored from tree inference). The clusters corresponding to the α1-like and α2-like collagen subfamilies are indicated in the right-hand side of the figure. A total of 100 bootstrap replications were used to test the reliability of the tree. Bootstrap percentages ≥50 are shown. Vertical bars delineate the α1-like and α2-like subfamilies. (*B*) Phylogenetic analysis by NJ of vertebrate type IV collagens. Tree was built using a multiple sequence alignment of type IV collagen NC1 sequences from chicken, mouse, and human (221 informative sites) and rooted with a sponge type collagen sequence from *Pseudocorticium jarrei*. Numbers on the nodes indicate the percentage recovery of that node in 1,000 bootstrap replications. Bootstrap probabilities higher than 50% are shown.

short-chain collagen–related genes in several organisms, namely, hydra and sea urchin. Unfortunately, owing to the lack of sequence data, phylogeny of the spongin short-chain collagen–related family can hardly be resolved further.

A novel series of multiple alignments was done using spongin short-chain collagen–related and collagen IV NC1 subdomains instead of complete domains, and NJ and ML phylogenetic trees were derived. For each protein family, sequences corresponding to the first subdomain clustered as one monophyletic group and sequences corresponding to the second subdomain formed a similar cluster (fig. 7). Trees built by using more accurate multiple alignments of either spongin short-chain collagen–related or collagen IV NC1 subdomain sequences also strongly supported the separate clustering of each subdomain. In other words, the subdomains of spongin short-chain collagen–related NC1 are more similar to one another than to the corresponding subdomain in collagen IV NC1. Likewise, there is significantly more similarity between the a and b subdomains of type IV collagen NC1 than there is between these subdomains and the corresponding subdomains of spongin short-chain collagen–related NC1. These observations could be interpreted as evidence that division into 2 homologous subdomains resulted from 2 independent tandem duplication events in the spongin short-chain collagen–related and type IV collagen clades. In favor of this hypothesis is the fact that contiguous subdomains are more distantly related in spongin short-chain collagen–related proteins than in type IV collagen subdomains (fig. 8). However, pairwise percent identity scores (tables S3 and S4, Supplementary Material online) and overall similarity values (see figs. 2*C* and *D* and 3) indicate that this may actually be due to faster divergence rates for spongin short-chain collagen–related NC1 sequences compared with type IV collagen NC1 sequences. Tree topologies demonstrating separate clustering of homologous spongin short-chain collagen–related and type IV collagen subdomains were likely in light of the great amino acid divergence between each family. As NC1 domains of spongin short-chain collagen–related and type IV collagen chains are both N-terminally flanked by triple helix, the hypothesis of a single, initial duplication

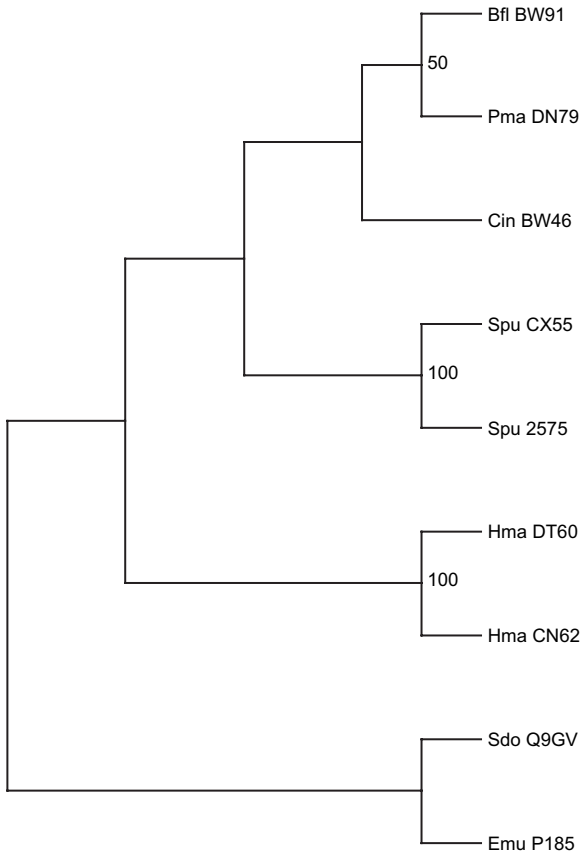


FIG. 6.—NJ tree of the spongin short-chain collagen–related subfamily. Based on an alignment of the NC1 domain sequences of 206 amino acids in length, a phylogenetic tree was inferred by the NJ method (1,000 bootstrap replicates) and rooted with the sponge sequences (*Ephydatia mülleri* and *Suberites domuncula*).

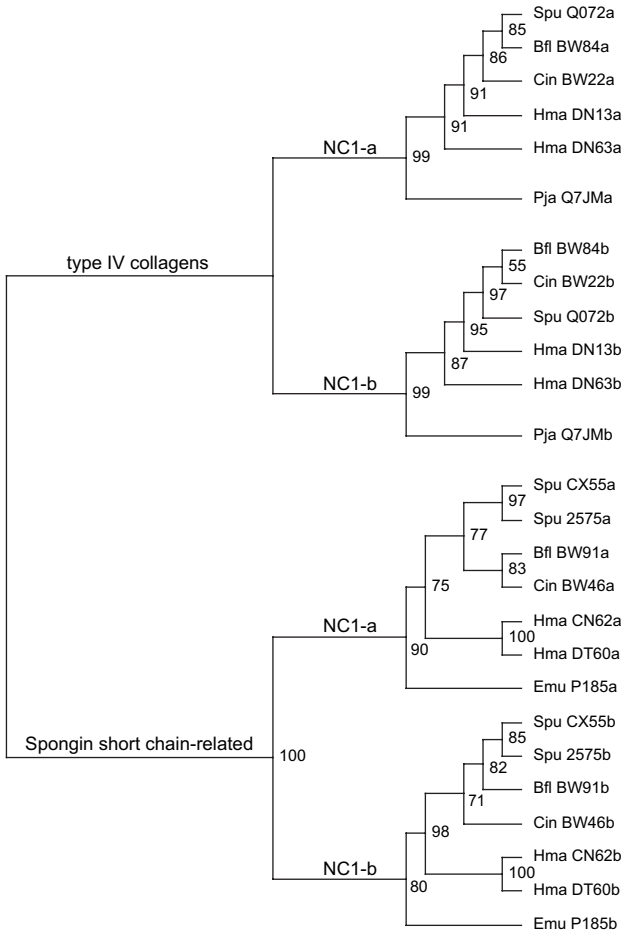


FIG. 7.—Phylogenetic analysis by BioNJ of spongin short-chain collagen–related and type IV collagen subdomain sequences. Phylogenetic reconstruction was performed using a multiple sequence alignment between the a and b NC1 subdomains of spongin short-chain collagen–related proteins and type IV collagens from invertebrates (90 informative sites). Homologous a and b subdomains grouped robustly within each protein subfamily when more sequences were added to the alignments or when the ML method was applied. The orthology group containing the sequences Spu Q266, Bfl BW89, and Cin BW43 was not included in this tree. Bootstrap values are shown only when they are greater than 50%.

resulting in one complete NC1 sequence subsequently fused to a triple-helical motif seems therefore more parsimonious.

Discussion

To prospect for the presence of proteins including a specific module in a species, use of Blast programs is successful in most circumstances. However, as exemplified in this work, Blast analyses may sometimes be insufficient to trace the natural history of a protein module (Schmid and Tautz 1997; Schmid and Aquadro 2001; Domazet-Loso and Tautz 2003; Mueller et al. 2004). In this report, we suggest that *E. mülleri* spongin short-chain collagen NC1 domain is homologous to the corresponding domain in type IV collagen. This conclusion is based to a significant extent upon the demonstration that these domains 1) are equally subdivided into 2 subdomains of equal lengths, 2) contain conserved cysteines, and 3) display common structural motifs identified using 2D and 3D predictions. Noteworthy, spongin short-chain collagen and type IV collagen NC1 domains have undergone such a drift that they are not picked up by classical automated domain detection procedures (e.g., ProDom, PFAM, SCOP, PROSITE, and Interpro).

Extracellular matrix proteins are mainly multimodular and are often defined as mosaic entities with each type of module present in multiple copies in one protein and/or in several protein families. Although domains used in the building of extracellular proteins are usually domains of great mobility (Tordai et al. 2005; Patthy 1999), the spongin short-chain collagen/type IV collagen NC1 does not appear to be a mobile domain, that is, it is retrieved from the available sequences with a unique domain partner (the collagen triple helix) and in a conserved architecture. This domain therefore contributed to an ancient multimodular protein, the collagen, but apparently no longer participated in novel domain combinations during metazoan evolution. Interestingly, the situation is analogous for the C-propeptide in fibrillar collagen which, like type IV collagen NC1 domain, is involved in chain selection and in initiation of triple helix formation (Lees et al. 1997; Myllyharju and Kivirikko 2001).

A model for the evolution of the spongin short-chain collagen/type IV collagen NC1 domain is presented in

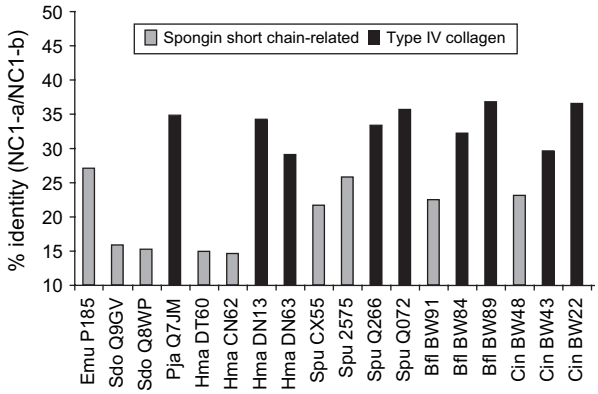


FIG. 8.—Percentage of identity between contiguous NC1 subdomains for spongin short-chain collagen–related and type IV collagen proteins. Gray and Black boxes represent spongin short-chain collagen–related and type IV collagen chains, respectively. The identity matrix was generated using ClustalW.

figure 9. In this scenario, the sequence encoding the NC1 structural unit made up of 2 homologous subdomains was produced by ancient tandem duplication. This event, leading to the 2-fold repeated structural pattern observed in modern spongin short-chain collagen–related proteins and type IV collagen NC1 domains, occurred probably in the very early evolution of animals, before the parazoan–eumetazoan split. The nature (and possible function) of the ancestral sequence, which gave rise to the NC1 internal repeat, is not known. Moreover, our phylogenetic analysis did not allow us to infer which of the subdomains (a or b) was the primordial building block. The structure of the putative protodomain, rich in β-sheets, raises the possibility that it might have already been involved in protein–protein interactions and oligomerization at the extracellular level (Wang and Hecht 2002; Siepen et al. 2003). Relevant to this is the fact that the NC1 subdomains of spongin short-chain collagen–related proteins and collagen IV are disulfide-bonded β-rich polypeptides, these features being common in extracellular modules that face the oxidative environment of the extracytoplasmic space (Martin et al. 1998). Partition of modern NC1 domains into 2 subdomains seems to constitute an essential feature for both structure and function, as we were not able to retrieve any sequences encoding isolated subdomains. Crystallographic data indicate that the β-strands located near the triple-helix junction or close to the hexamer interface are contributed by different subdomains. Therefore, structural requirements driving trimeric association and oligomerization may be sufficient to explain why both subdomains are needed for the NC1 domain in order to achieve its function.

The initial tandem replication event was followed by gene duplication creating 2 copies that diverged to become the spongin short-chain collagen–related and type IV collagen ancestral genes. As domain combinations are usually formed only once (Vogel et al. 2005), it is most parsimonious to consider that the spongin short-chain collagen–related and type IV collagen genes evolved by duplication of an NC1 domain already combined to a triple-helical motif, rather than emergence from independent recombination events. It is tempting to speculate that the ancestral gene

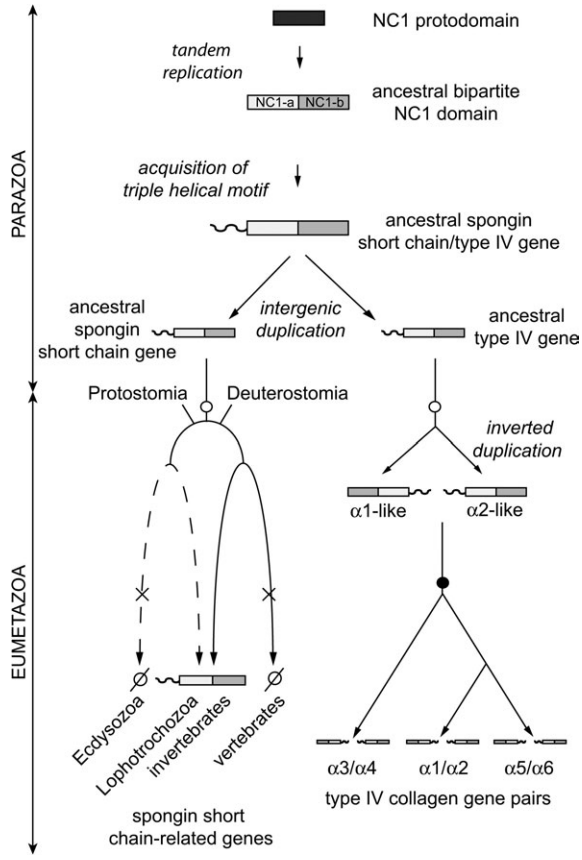


FIG. 9.—A hypothesis for the natural history of the spongin short-chain collagen–related and type IV collagen families during animal evolution. We propose to assign the following polarity to the evolutionary events related to the NC1 domain of the spongin short-chain collagen–related and type IV collagen families. An ancient tandem duplication, predating the divergence of Parazoa and Eumetazoa, generated the internal repeat found within spongin short-chain collagen–related and type IV collagen NC1 domains. This initial tandem replication was followed by acquisition of the collagen triple-helical motif. Indeed, from the structural data, it is unlikely that the NC1 subdomain alone could initiate triple helix formation. The ancestral spongin short-chain collagen–related/type IV collagen gene duplicated prior to the common ancestor of all extant bilaterians to produce daughter genes that evolved separately. Type IV collagen gene family increased first by an inverted duplication (producing head-to-head gene pairing), and then by 2 rounds of duplication in vertebrates to give rise to 3 bigene clusters. The evolution of the spongin short-chain collagen–related family is poorly defined because of the paucity of sequence data in extant species. However, the spongin short-chain collagen–related gene might have been lost (presumably independently) in the lineages leading to nematodes, arthropods, and vertebrates. NC1 domains are shown as open rectangles, and triple-helical motif is represented as a tail. Black circle indicates the Cephalochordata–Vertebrata split, whereas the Radiata–Bilateria split is represented by an open circle. Inferred gene losses are shown by crosses.

was more related to spongin short-chain collagen than to type IV collagen, as it evolved in early metazoans devoid of basement membranes (such as the demosponges). In this hypothesis, cells from early-emerging multicellular animals first evolved spongin short-chain collagen–related proteins as part of some basic mechanisms for sticking to an extracellular surface, before type IV collagen emerged as an essential component for attachment to the basal lamina, which presumably evolved later.

The spongin short-chain collagen–related/collagen IV duplicated genes underwent different fates during

eumetazoan evolution, including lineage-specific gene duplications (e.g., spongin [Exposito et al. 1991], spongin short-chain collagen–related genes in hydra and sea urchin, see figs. 4 and 7). This scenario of gene duplication from an ancestral half-domain sequence, followed by subsequent gene duplication and diversification, is reminiscent of the evolution of β/α barrels in the microbial world (Lang et al. 2000).

Importantly, our data mining analysis suggest that the spongin short-chain collagen–related gene has been lost in the common ancestor of the ecdysozoa lineage and in the common ancestor of the vertebrate lineage. Analysis of priapulids, which are placed basal to nematodes and arthropods in the Ecdysozoa, and of early vertebrates (e.g., hagfishes and lampreys) will help providing clues to these possible events of gene loss. It is intriguing that vertebrates, that produce mineralized tissues, and moulting invertebrates, which have an external nonmineralized skeleton (note that arthropods have chitin-made cuticles while the nematode exoskeleton is formed by non–type IV collagen proteins) may be devoid of spongin short-chain collagen–related proteins. If this apparent absence is not the mere result of missing data, it is tempting to speculate that spongin short-chain collagen–related ancestral gene loss in the ancestors of these organisms played a role in differentiating such specialized tissues. In any case, spongin short-chain collagen–related genes are likely to be less essential than collagen IV genes because deletions might have eliminated them from several eumetazoan genomes. Alternatively, organisms from these lineages may contain spongin short-chain collagen–related genes that are too divergent to be uncovered by sequence analysis using current tools, that is, these genes have evolved rapidly in vertebrates and Ecdysozoa and no longer have recognizable similarity. As a general rule, the primary structure of complete spongin short-chain collagen–related NC1 domains have been poorly conserved, even in closely related species, whereas the sequences of type IV collagen NC1 have been more preserved during metazoan evolution. Thus, sequence evolution rates and propensity for gene loss may be correlated in the system described here, with spongin short-chain collagen–related sequences evolving faster than collagen IV NC1 sequences. This observation is in line with recent works suggesting that weakly constrained proteins are lost during evolution significantly more often than highly constrained ones (Kamath et al. 2003; Krylov et al. 2003).

Assuming that spongin short-chain collagen–related proteins are involved in extracellular attachment, like spongin short-chain collagens, this marked sequence divergence between the different spongin short-chain collagen–related NC1 domains may reflect the diversity of substrata available for attachment in various invertebrate lineages. In that respect, it would be of great interest to determine the expression profile of spongin short-chain collagen–related genes (vs. type IV collagen), and the functions of their encoded products both in sponges and, most importantly, in nonsponge organisms (e.g., hydra, ciona, and amphioxus). What function could proteins related to spongin short-chain collagens have in protostomes and invertebrate deuterostomes? To date, results on expression pattern are only available in *Ciona intestinalis* and were generated by large-scale automated in situ hybridization (http://ghost.zool.kyoto-u.ac.jp/). These experiments reveal that a *C. intestinalis* spongin short-chain collagen–related gene (Cin BW46, expressed sequence tag cluster CLSTR03436r1) is expressed in juvenile animals, in epithelial cells, and in body wall muscle but do not inform on the tissue distribution of the corresponding protein. As a matter of fact, in absence of experimental data, it is not obvious what role could play spongin short-chain collagen–related proteins in organisms that possess basement membranes. Although they could be suspected of involvement in cell-matrix adhesion, intercellular cohesion, and organismal organization, spongin short-chain collagen–related proteins may also subserve more specialized functions. Another aspect is whether or not spongin short-chain collagen–related NC1 domains are involved in protomer formation and assembly into hexamers. Determination of the precise subcellular localization and interaction partners of spongin short-chain collagen–related proteins together with biochemical characterization will hopefully offer insightful information into these important issues.

Conclusion

Spongin short-chain collagens and type IV collagen are among the oldest modular proteins unique to Metazoa because they are already present in Porifera and Cnidaria. In modern multicellular animals, spongin gives a sponge its flexibility and support, whereas collagen gives both properties to a tissue. Spicules and extracellular matrix both integrate cells into 3D structures, emphasizing the functional analogy existing between substratum attachment and basement membrane attachment. In this work, we reported the discovery of a novel family of proteins related to sponge short-chain collagens in a number of nonsponge invertebrates, which may have homologous relationships with type IV collagens in their NC1 domain. Remote homology detection was followed by phylogenetic analysis, revealing that type IV collagens and spongin short-chain collagen–related proteins have had separate evolutionary histories. Because extracellular matrix attachment is thought to have played crucial roles in the evolution of multicellular animals, deciphering the phylogeny and function of these proteins is of considerable interest.

Supplementary Material

Supplementary Tables S1–S4 and Figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

A.A. is a recipient of a fellowship from the Centre National de la Recherche Scientifique. V.N. is supported by a grant from Institut National de la Recherche Agronomique.

Literature Cited

Aho S, Turakainen H, Onnela ML, Boedtker H. 1993. Characterization of an intronless collagen gene family in the marine sponge *Microciona prolifera*. Proc Natl Acad Sci USA. 90:7288–7292.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Borchiellini C, Chombard C, Manuel M, Alivon E, Vacelet J, Boury-Esnault N. 2004. Molecular phylogeny of Demospongiae: implications for classification and scenarios ofcharacter evolution. Mol Phylogenet Evol. 32:823–837.

Borza DB, Bondar O, Ninomiya Y, Sado Y, Naito I, Todd P, Hudson BG. 2001.The NC1 domain of collagen IV encodes a novel network composed of the α1, α2, α5, and α6 chains in smooth muscle basement membranes. J Biol Chem. 276:28532–28540.

Boutaud A, Borza DB, Bondar O, Gunwar S, Netzer KO, Singh N, Ninomiya Y, Sado Y, Noelken ME, Hudson BG. 2000. Type IV collagen of the glomerular basement membrane. Evidence that the chain specificity of network assembly is encoded by the noncollagenous NC1 domains. J Biol Chem. 275:30716–30724.

Boute N, Exposito JY, Boury-Esnault N, Vacelet J, Noro N, Miyazaki K, Yoshizato K, Garrone R. 1996. Type IV collagen in sponges, the missing link in basement membrane ubiquity. Biol Cell. 88:37–44.

Brünger AT, Adams PD, Clore GM, et al. (14 co-authors). 1998. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr. 54:905–921.

Combet C, Jambon M, Deleage G, Geourjon C. 2002. Geno3D: automatic comparative molecular modelling of protein. Bioinformatics. 18:213–214.

DeLano WL. 2005. MacPyMOL: A PyMOL-based molecular graphics application for MacOS X [Internet]. PyMol version 0.99. San Francisco, CA: DeLano Scientific LLC. Available from: http://www.pymol.org.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 13:2213–2219.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.

Erickson AC, Couchman JR. 2000. Still more complexity in mammalian basement membranes. J Histochem Cytochem. 48:1291–1306.

Exposito JY, Cluzel C, Garrone R, Lethias C. 2002. Evolution of collagens. Anat Rec. 268:302–316.

Exposito JY, Garrone R. 1990. Characterization of a fibrillar collagen gene in sponges reveals the early evolutionary appearance of two collagen gene families. Proc Natl Acad Sci USA. 87:6669–6673.

Exposito JY, Le Guellec D, Lu Q, Garrone R. 1991. Short chain collagens in sponges are encoded by a family of closely related genes. J Biol Chem. 266:21923–21928.

Exposito JY, Ouazana R, Garrone R. 1990. Cloning and sequencing of a Porifera partial cDNA coding for a short chain collagen. Eur J Biochem. 190:401–406.

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_ WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci. 12:543–548.

Garrone R. 1984. Formation and involvement of extracellular matrix in the development of sponges, a primitive multicellular system. In: Trelstad RL, editor. The role of extracellular matrix in development. New York: Alan R. Liss. p. 461–477.

Garrone R. 1985. The collagen of porifera. In: Bairati A and Garrone R, editors. Biology of invertebrate and lower vertebrate collagens. London: Plenum Press. p. 157–175.

Geourjon C, Combet C, Blanchet C, Deleage G. 2001. Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci. 10:788–797.

Geourjon C, Deleage G. 1995. ANTHEPROT 2.0: a three-dimensional module fully coupled with protein sequence analysis methods. J Mol Graph. 13:209–212, 199–200.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hamano Y, Kalluri R. 2005. Tumstatin, the NC1 domain of α3 chain of type IV collagen, is an endogenous inhibitor of pathological angiogenesis and suppresses tumor growth. Biochem Biophys Res Commun. 333:292–298.

Holland LZ, Gibson-Brown JJ. 2003. The *Ciona intestinalis* genome: when the constraints are off. Bioessays. 25:529–532.

Hudson BG, Reeders ST, Tryggvason K. 1993. Type IV collagen: structure, gene organization, and role in human diseases. J Biol Chem. 268:26033–26036.

Hudson BG, Tryggvason K, Sundaramoorthy M, Neilson EG. 2003. Alport's syndrome, Goodpasture's syndrome, and type IV collagen. N Engl J Med. 348:2543–2556.

Jones DT. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol. 287:797–815.

Kamath RS, Fraser AG, Dong Y, et al. (13 co-authors). 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature. 421:231–237.

Kelley LA, MacCallum RM, Sternberg MJE. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol. 299:499–520.

Khoshnoodi J, Sigmundsson K, Cartailler JP, Bondar O, Sundaramoorthy M, Hudson BG. 2006. Mechanism of chain selection in the assembly of collagen IV: a prominent role for the α2 chain. J Biol Chem. 281:6058–6069.

Krasko A, Lorenz B, Batel R, Schröder HC, Müller IM, Müller WE. 2000. Expression of silicatein and collagen genes in the marine sponge *Suberites domuncula* is controlled by silicate and myotrophin. Eur J Biochem. 267:4878–4887.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.

Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M. 2000. Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science. 289:1546–1550.

Laskowski RA, Moss DS, Thornton JM. 1993. Main-chain bond lengths and bond angles in protein structures. J Mol Biol. 231:1049–1067.

Lees JF, Tasab M, Bulleid NJ. 1997. Identification of the molecular recognition sequence which determines the type-specific assembly of procollagen. EMBO J. 16:908–916.

MacDonald BA, Sund M, Grant MA, Pfaff KL, Holthaus K, Zon LI, Kalluri R. 2006. Zebrafish to humans: evolution of the α3-chain of type IV collagen and emergence of the autoimmune epitopes associated with Goodpasture syndrome. Blood. 107:1908–1915.

Mariyama M, Kalluri R, Hudson BG, Reeders ST. 1992. The α4(IV) chain of basement membrane collagen. Isolation of cDNAs encoding bovine α4(IV) and comparison with other type IV collagens. J Biol Chem. 267:1253–1258.

Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. 1998. Protein folds and functions. Structure. 6:875–884.

Mueller JL, Ripoll DR, Aquadro CF, Wolfner MF. 2004. Comparative structural modeling and inference of conserved protein classes in Drosophila seminal fluid. Proc Natl Acad Sci USA. 101:13542–13547.

Mushegian AR, Garey JR, Martin J, Liu LX. 1998. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res. 8:590–598.

Myllyharju J, Kivirikko KI. 2001. Collagens and collagen-related diseases. Ann Med. 33:7–21.

Netzer KO, Suzuki K, Itoh Y, Hudson BG, Khalifah RG. 1998. Comparative analysis of the noncollagenous NC1 domain of type IV collagen: identification of structural features important for assembly, function, and pathogenesis. Protein Sci. 7:1340–1351.

Ortega N, Werb Z. 2002. New functional roles for non-collagenous domains of basement membrane collagens. J Cell Sci. 115:4201–4214.

Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 12:357–358.

Patthy L. 1999. Genome evolution and the evolution of exon-shuffling—a review. Gene. 238:103–114.

Schmid KJ, Aquadro CF. 2001. The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics. 159:589–598.

Schmid KJ, Tautz D. 1997. A screen for fast evolving genes from Drosophila. Proc Natl Acad Sci USA. 94:9746–9750.

Schröder HC, Krasko A, Batel R, Skorokhod A, Pahler S, Kruse M, Müller IM, Müller WE. 2000. Stimulation of protein (collagen) synthesis in sponge cells by a cardiac myotrophin-related molecule from *Suberites domuncula*. FASEB J. 14:2022–2031.

Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 310:243–257.

Siebold B, Deutzmann R, Kühn K. 1988. The arrangement of intra- and intermolecular disulfide bonds in the carboxyterminal, non-collagenous aggregation and cross-linking domain of basement-membrane type IV collagen. Eur J Biochem. 176:617–624.

Siepen JA, Radford SE, Westhead DR. 2003. β edge strands in protein structure prediction and aggregation. Protein Sci. 12:2348–2359.

Simpson TL. 1984. Collagen fibrils, spongin, matrix substances. In: TL Simpson, ed. The cell biology of sponges. New York: Springer-Verlag. p. 216–254.

Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. Gene. 318:169–175.

Söder S, Pöschl E. 2004. The NC1 domain of human collagen IV is necessary to initiate triple helix formation. Biochem Biophys Res Commun. 325:276–280.

Sullivan JC, Ryan JF, Watson JA, Webb J, Mullikin JC, Rokhsar D, Finnerty JR. 2006. StellaBase: the *Nematostella vectensis* Genomics Database. Nucleic Acids Res. 34(Database issue):D495–D499.

Sundaramoorthy M, Meiyappan M, Todd P, Hudson BG. 2002. Crystal structure of NC1 domains. Structural basis for type IV collagen assembly in basement membranes. J Biol Chem. 277:31142–31153.

Than ME, Bourenkov GP, Henrich S, Mann K, Bode W. 2005. The NC1 dimer of human placental basement membrane collagen IV: does a covalent crosslink exist? Biol Chem. 386:759–766.

Than ME, Henrich S, Huber R, Ries A, Mann K, Kühn K, Timpl R, Bourenkov GP, Bartunik HD, Bode W. 2002. The 1.9-Å crystal structure of the noncollagenous (NC1) domain of human placenta collagen IV shows stabilization via a novel type of covalent Met-Lys cross-link. Proc Natl Acad Sci USA. 99:6607–6612.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thompson JD, Thierry JC, Poch O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics. 19:1155–1161.

Timpl R, Wiedemann H, van Delden V, Furthmayr H, Kühn K. 1981. A network model for the organization of type IV collagen molecules in basement membranes. Eur J Biochem. 120:203–211.

Tordai H, Nagy A, Farkas K, Banyai L, Patthy L. 2005. Modules, multidomain proteins and organismic complexity. FEBS J. 272:5064–5078.

Vanacore RM, Friedman DB, Ham AJ, Sundaramoorthy M, Hudson BG. 2005. Identification of S-hydroxylysyl-methionine as the covalent cross-link of the noncollagenous (NC1) hexamer of the α1α1α2 collagen IV network: a role for the post-translational modification of lysine 211 to hydroxylysine 211 in hexamer assembly. J Biol Chem. 280:29300–29310.

Vogel C, Teichmann SA, Pereira-Leal J. 2005. The relationship between domain duplication and recombination. J Mol Biol. 346:355–365.

Wang W, Hecht MH. 2002. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric β-sheet proteins. Proc Natl Acad Sci USA. 99:2760–2765.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 21:236–239.

David Irwin, Associate Editor

Accepted August 14, 2006