The α4(IV) Chain of Basement Membrane Collagen

ISOLATION OF cDNAs ENCODING BOVINE α4(IV) AND COMPARISON WITH OTHER TYPE IV COLLAGENS*

(Received for publication, July 18, 1991)

Mariko Mariyama‡, Raghuram Kalluri‡, Billy G. Hudson‡, and Stephen T. Reeder§

From the §Howard Hughes Medical Institute and the ‡Departments of Internal Medicine and Genetics, Yale University School of Medicine, New Haven, Connecticut 06510 and the ‡Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, Kansas 66103

Renal basement membranes are believed to contain five distinct type IV collagens. An understanding of the specific roles of these collagens and the specificities of their interactions will be aided by knowledge of their comparative structures. Genes for α1(IV), α2(IV), α3(IV), and α5(IV) have been cloned and the deduced peptide sequences compared. A fifth chain, α4(IV), has been identified in glomerular and other basement membranes. Using a polymerase chain reaction-based strategy and short known peptide sequences from the noncollagenous domain (NC1), we have cloned and characterized partial bovine cDNAs of α4(IV). Sequence analysis shows that this molecule has characteristic features of type IV collagens including an NH2-terminal Gly-X-Y domain which is interrupted at several points and a COOH-terminal NC1 domain with 12 cysteine residues in positions identical to those of other type IV collagens. Within the NC1 domain bovine α4(IV) has 70, 59, 58, and 53% amino acid identity with human α2(IV), α1(IV), α5(IV), and α3(IV), respectively. Alignment of the peptides also shows that α4(IV) is most closely related to α2(IV). Nevertheless, in the extreme COOH-terminal region of the NC1 domain there are specific regions of identity to α4(IV). Cloning of the region of α4(IV) that encodes the NC1 domain allows comparison of all five type IV collagens and highlights certain regions that are likely to be important in the specificities of NC1-NC1 interactions and in other discriminant functions of these molecules.

It is now apparent that there are at least five basement membrane (type IV) collagen chains: α1(IV), α2(IV), α3(IV), α4(IV), and α5(IV). Genes for α1(IV) and α2(IV) have been cloned and the primary amino acid structures determined in their entirety (see Hostikka and Tryggvason, 1988 for a summary). The overall structure of the chains is similar, each having a noncollagenous or globular domain of ~230 residues at the COOH terminus (NC1), an extensive collaginous region that forms a triple helix with two other type IV collagen chains and a 7 S domain at the NH2-terminal end. The triple-helical region, which is ~1400 residues long, has the consensual sequence Gly-X-Y, where X and Y are nonconserved amino acids; in the case of type IV collagens, the triplet arrangement is interrupted at several points.

The NC1 domains of the α1(IV) and α2(IV) chains are highly conserved at the amino acid level (Hostikka and Tryggvason, 1988), and the genes appear to have evolved from a common ancestral precursor. Bacterial collagenase releases the NC1 domains from other components of basement membrane as hexamers, each comprised of three NC1 domains of two heterotrimeric (Weber et al., 1984). The NC1 domains can be further separated on the basis of molecular weight and charge. Hudson and colleagues (Wieslander et al., 1984; Butkowski et al., 1985, 1987; Langeveld et al., 1988; Saus et al., 1988; Gunwar et al., 1990) identified four separate monomers from bovine glomerular and lens basement membranes and showed, on the basis of short stretches of peptide sequence, that these corresponded to four separate chains: the well characterized abundant species, α1(IV) and α2(IV) and two novel chains designated α3(IV) and α4(IV). A similar array of monomers has subsequently been demonstrated in human renal basement membrane (Yoshioka et al., 1985; Kieppel et al., 1986; Wieslander et al., 1987; Butkowski et al., 1990).

In order to study the structure, function, and assembly of basement membrane collagen, it is necessary to know the primary structures of all of the α-chains. We have recently isolated bovine and human partial cDNAs for the α3(IV) chain and have demonstrated that its NC1 domain has many features in common with the α1(IV) and α2(IV) chains including conservation of 12 cysteine residues that have been found in all type IV collagens studied to date (Morrison et al., 1991a, 1991b). Recently, a fifth type IV collagen, α5(IV), whose existence had not been suspected from biochemical or immunological studies of glomerular basement membrane (GBM), was cloned on the basis of homology with a highly conserved region of known type IV molecules (Hostikka et al., 1990; Pihlajaniemi et al., 1990). Of the five type IV collagens known to exist, the genes of four have therefore been isolated. Here we describe the use of the polymerase chain reaction (PCR) to isolate partial cDNAs for the last of the known type IV collagens, α4(IV). The strategy was based on the availability of short stretches of peptide sequence from the NC1 domain of α4(IV).

The abbreviations used are: NC1, noncollagenous domain; PCR, polymerase chain reaction; GBM, glomerular basement membrane; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis.

*This work was supported by the Howard Hughes Medical Institute and by Grant DK18381 from the National Institute of Diabetes and Digestive and Kidney Disease. 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) M77480.

‡ To whom correspondence should be addressed: Boyer Center for Molecular Medicine, Yale University School of Medicine, P. O. Box 9812, New Haven, CT 06536-0812. Tel.: 203-737-4416; Fax: 203-624-8213.

1 The abbreviations used are: NC1, noncollagenous domain; PCR, polymerase chain reaction; GBM, glomerular basement membrane; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis.
EXPERIMENTAL PROCEDURES

Primer Design—Oligonucleotide primer sequences are given in Fig. 1. Pooled (or degenerate) sense (4F-1 and 4F-2) and antisense (5R-1 and 5R-2) primers were designed to correspond to the coding sequences of short known amino acid sequences from the NC1 domain of the bovine a4(IV) chain. All codon usages were included except that rare codons occurring at the 5′ end of primers were ignored. In addition, to reduce the degeneracy of each primer pool, between two and four primer pools were synthesized separately for both the sense (e.g. 4F-1 and 4F-2) and antisense (e.g. 5R-1 and 5R-2) directions.

A second set of pooled antisense primers (4R-1–4R-4) was designed so that the 15 residues at the 3′ end correspond to a highly conserved amino acid sequence (CQVCVM) shared by the NC1 domains of all known vertebrate type IV collagens (Fig. 1). Since a minimum of 18 nucleotides are required for stable primer annealing, an additional three nucleotides (coding for lysine) were added at the 5′ end of the primer (Fig. 1). The recognition sequences of EcoRI or BamHI were added to the 5′ end of primers to facilitate subsequent cloning of PCR products.

PCR Protocol—First, PCR was performed using 50 ng of bovine renal cDNA template with 25 pmol of either 4F-1 or 4F-2 (sense primers) and either 4R-1 or 4R-2 (antisense primers). All combinations of sense and antisense primers were tried. A stepwise decreasing annealing temperature protocol was used: 1 min at 52 °C for the first three cycles, at 48 °C for the next three cycles, at 46 °C for the next three cycles and at 43 °C for 35 cycles (Morrison et al., 1991a). Denaturation was carried out at 92 °C for 1 min and extension at 72 °C for 1 min.

A second set of PCR was performed using one fiftieth of the PCR product of the first reaction as a template with all combinations of primers 4F-1, 4F-2 and 5R-1, 5R-2. The 5R-1 and 5R-2 primers (Fig. 1) were based on the peptide sequence of an elastase fragment of the a4(IV) NC1 domain (see below). The PCR cycle conditions were: 1 min at 92 °C, 1 min at 52 °C, and 1 min at 72 °C for 30 cycles. A PCR product of the predicted size was excised from a 1.2% agarose gel and purified according to the method of Gunwar et al. (1990). The isolated DNA was electrophoresed on gels cast a 1% agarose gel containing formaldehyde, blotted on an oligo(dT) column (Pharmacia, Uppsala, Sweden). RNA was reverse transcribed to cDNA and then digested with Pseudomonas aeruginosa elastase (1:63 enzyme:substrate ratio) in the presence of 20 mM final concentration. One pg of this digest was analyzed by SDS-PAGE using 4-22% gradient gels (Laemmli, 1970) (Fig. 2). The gel was stained with Coomassie Blue for 30 min and destained in 10% acetic acid, 20% methanol. The filter was washed with distilled water, stained with Coomassie Blue, destained in 50% methanol, 10% acetic acid, washed with water, and dried. The protein bands were carefully excised and used directly for sequencing. Amino-terminal sequence analysis was performed at the Microsequencing Facility of the Department of Biological Chemistry, UCLA School of Medicine.

Library Screening—A 250-base pair PCR product was labeled with [α-32P]dCTP by random primer labeling and used to screen an oligo(dT)-primed λgt11 bovine lens cDNA library. A total of 100 clones were screened. The filters were also hybridized with a probe representing nucleotides 4425–4475 of the human a2(IV) cDNA (Hestikka and Tryggvason, 1988) and strongly hybridizing clones were discarded. After purification of the positives, each single plaque was eluted into 500 μl of SM buffer (100 mM NaCl, 8 mM MgSO4, 50 mM Tris-Cl (pH 7.5), and 0.01% gelatin). 5 μl of each was used as a template for PCR with primers complementary to the β-galactosidase portion of the λgt11 template. In order to reduce the frequency of PCR products that contain misincorporation errors, at least 5000 copies of template were used in the PCR reactions. The amplified product was digested with EcoRI and cloned into the EcoRI site of pBSII (Stratagene).

DNA Sequencing—The sequences of the cloned cDNAs and PCR products were obtained using plasmid DNA as template. T7 and T3 sequencing primers and 18-nucleotide primers (designed from known sequences of the inserts once these had been obtained) were used in the dyeoxy method with T7 polymerase (Sequenase). Because the cDNA was recovered from bacteriophage clones by PCR (see above) and because of the risk of misincorporation errors during amplification, the sequence was confirmed (with the exception of nucleotides 0–332) in at least two independent clones (Fig. 3).

Northern Analysis—For Northern analysis, RNA was isolated from snap-frozen tissue of a 60-day-old calf using acid guanidinium thiocyanate/phenol/chloroform extraction procedures (Chomczynski and Sacchi, 1987). Poly(A)-enriched RNA was isolated from total RNA on an oligo(dT) column (Pharmacia, Uppsala, Sweden). RNA was electrophoresed on a 1% agarose gel containing formaldehyde, blotted to nitrocellulose, and hybridized with gel-purified MM544, the bovine a4(IV) probe.

RESULTS

The amino acid sequence of a 30-residue stretch of the extreme NH2-terminal end of the NC1 domain of a novel type IV collagen, designated a4(IV), was previously determined by Gunwar et al. (1990): GPPGFPGYGGLFLVLSQTD-GEPTXPMG (where the identity of X could not be determined). Prior attempts to use this sequence to design an oligonucleotide hybridization probe for a4(IV) were unsuccessful. Therefore, an attempt was made to use PCR to amplify a segment of DNA coding for the 30-residue segment using highly degenerate 18 nucleotide sense and antisense primers coding for the ends of the region. This strategy had been successful in isolating a3(IV) sequences (Morrison et al.,...
Although several products of the correct size were obtained, sequence analysis showed that none of these was capable of encoding the amino acid sequence obtained by Gunwar et al. (1990). Therefore, a second PCR-based strategy was attempted in which it was assumed that a five-amino acid stretch (CQVCM) which is conserved in all vertebrate type IV molecules described hitherto would also be conserved in α4(IV). Primer pools (4R-1 through 4R-4) corresponding to all possible coding sequences for this region were used as an antisense primers (Fig. 1). Two primer pools (4F-1 and 4F-2), each coding for a portion of the peptide sequenced by Gunwar et al. (1990), were used as sense primers (Fig. 1). PCR reactions using this combination of primers gave a large number of products. In order to select the correct product a nested primer strategy was required. This strategy made use of additional peptide sequence from the region between 4F-1 and 4R-1, obtained by analysis of an elastase fragment of the α4(IV) NC1 domain.

The purified α4(IV) NC1 monomer was partially digested with P. aeruginosa elastase. Digestion fragments of between 14 and 15 kDa were obtained. The NH₂-terminal sequence of the 15-kDa fragment was established as LAEDEISPYISR. A weaker serine signal was present at the second position from the NH₂ terminus. The 15-kDa fragment was assumed to represent the COOH-terminal half of the α4(IV) NC1 domain. A new set of antisense primers (5R-1 and 5R-2) were based on this sequence (Fig. 1). Since this short peptide contains only two and three sequence differences from the human and murine α2(IV) sequences, respectively (Hostikka and Tryggvason, 1988; Saus et al., 1989), the antisense primer pool was designed so that the extreme 3' end of the primers corresponded to the position of one of the 2 residues at which α2(IV) and α4(IV) differ. The PCR reaction yielded a single product (MM28) of about the expected size (261 base pairs). This product was used to screen a bovine lens cDNA library and 17 positive clones were obtained. DNA sequence (which is identical to the sequence of the cDNA (MM544) obtained subsequently and is therefore not shown) revealed that MM28 is similar to human and murine α2(IV) over its entire length. Since it was anticipated that some of the positive cDNA clones would in fact contain α2(IV) fragments, the plaques were counter-hybridized with a human α2(IV) probe. The four clones that were not strongly positive with α2(IV) were selected (MM544, MM2, MM10, MM12). These clones were shown, by restriction mapping and sequence analysis, to contain overlapping cDNA fragments (Fig. 3).

**Nucleotide Sequence of α4(IV) cDNA**—The DNA sequence of MM544 and the other overlapping clones (Fig. 3) encode the COOH-terminal end of a peptide having characteristic features of a type IV collagen (Figs. 4 and 5). At the NH₂-terminal end are 222 residues containing the classical collagenous Gly-X-Y motif with three interruptions. At the COOH-terminal end is a 231-residue NC1 domain with 12 cysteine residues in the same relative positions as those found in all other type IV collagens that have been analyzed hitherto. The deduced amino acid sequence is identical to that obtained by direct amino acid sequencing in 38 of 42 positions. In two positions (amino acid residues 219 and 318 of Fig. 4), one of two possible residues determined by peptide analysis was compatible with the nucleotide sequence. The direct and deduced sequences differ at position 323. This nucleotide sequence was identical in four independent cDNA clones, ruling out the possibility of an incorporation error introduced during the PCR reactions. One of the 42 residues (residue 243 of Fig. 4) could not be determined by peptide sequencing (Gunwar et al., 1990).

**Comparative Sequence Analysis**—Overall the sequence shows 70, 59, 58, and 53% amino acid identity with NC1 domains of the human α2(IV), α1(IV), α5(IV), and α3(IV) chains, respectively. Within the NC1 domain, α4(IV) is most closely related to α2(IV) both in terms of the percentage amino acid identity and in terms of structural similarities such as the presence of an additional amino acid at position 402 (of the α4(IV) sequence) and the absence of amino acids (corresponding to residues 1532 and 1534 of the human α1(IV) sequence) where α1(IV), α3(IV), and α5(IV) have additional residues. α4(IV) is unique in having additional residues in positions 419 and 435 (of the α4(IV) sequence). Within the triplets of the collagenous region (for which deduced amino acid sequence is available), α4(IV) shows no particular sequence relationship with other α(IV) chains except for the presence of the triplet motif. There are two interruptions in the Gly-X-Y motif at similar positions in α1(IV), α2(IV), α4(IV), α5(IV), and α5(IV). The α2(IV), α3(IV), and α4(IV) chains each has an additional short interruption within a 9-residue interval (Fig. 5). Another striking feature of the α4(IV) collagenous domain is the presence of 4 cysteine residues, 3 of which are within the repetitive Gly-X-Y region. Although cysteine residues have been found within the Gly-X-Y interruptions and within the 7 S domains of other type IV collagens, α4(IV) is atypical in having cysteine residues in the Gly-X-Y units of the COOH-terminal half of the molecule.

**Northern Analysis**—The bovine cDNA clone MM544, which encodes 453 residues of α4(IV), was used to probe a Northern blot of poly(A)-enriched RNA from bovine kidney and liver. Fig. 6 shows that the gene codes for a single transcript of ~10.4 kilobases in these tissues.

**DISCUSSION**

Collagenase digests of both bovine GBM and bovine lens basement membrane reveal four separate monomers that have been termed M1α, M1β, M2*, and M3 (Butkowski et al., 1985, 1991a).
Fig. 4. The composite nucleotide sequence and deduced peptide sequence of bovine \(\alpha(IV)\) cDNA obtained from the overlapping clones shown in Fig. 3. The vertical arrow marks the beginning of the NC1 domain.

The regions corresponding to the PCR primers are underlined, and the primer names are in italics. The portion of the gene corresponding to the known peptide sequence is marked (<...>).

1987; Langeveld et al., 1988; Saas et al., 1988; Gunwar et al., 1990). A similar array of collagenase fragments has been observed in human GBM where the corresponding moieties have been termed M24, M26, M28**, and M28*, respectively (Butkowski et al., 1990). The first three of these have been shown to be the COOH-terminal (NC1) portions of three collagens of the type IV class. The regions corresponding to the PCR primers are underlined, and the primer names are in italics. The portion of the gene corresponding to the known peptide sequence is marked (<...>).
probes. The positions of RNA size markers of 1.4-9.5 kilobases are shown.

ruptions characteristic of type IV collagens. It has a 231-residue noncollagenous COOH-terminal domain (NC1) with between 53 and 70% similarity with other vertebrate α-chains. The 12 cysteine residues which are important in intra- and intermolecular disulfide bonds, and which are found in conserved positions in all type IV collagens studied to date, are also found in α4(IV). As with other type IV collagens, the NC1 domain of α4(IV) appears to have arisen by tandem duplication of an ancestral hemi-NC1 region. The presence of this structural feature indicates that α4(IV) also arose from an ancestral precursor that is common to all known type IV collagens and negates the hypothesis that it is part of a genetically distinct collagen family (Kim et al., 1991).

The bovine α4(IV) NC1 domain has the greatest peptide sequence identity with corresponding regions of the murine (70%) and human (70%) α2(IV) chains (Hostikka and Tryggvason, 1988; Saus et al., 1989). Moreover, within the NC1 domain at least, α4(IV) is more readily aligned with α2(IV) than with other α-chains. The sequences and structural similarities between α2(IV) and α4(IV) on the one hand and between α1(IV), α3(IV), and α5(IV) on the other, allow the vertebrate type IV collagens that have been defined so far to be tentatively divided into two families: α1- and α2-like. Although in Drosophila only one type IV collagen has been found, the α1:α2 dichotomy is present in Caenorhabditis elegans, where clb-1 is α2-like and clb-2 is α1-like (Guo and Kramer, 1989). Differences in function between the various α-chains have not as yet been determined, but such differences are likely to exist and each functional subtype may correspond to a structurally defined family.

Functional differences may also be reflected as structural heterogeneity within basement membrane (Langeveld et al., 1988) and as differences in the tissue distribution of type IV collagens. Within the kidney the distribution of the α3(IV) and α4(IV) chains is similar but distinct from the common pattern shared by α1(IV) and α2(IV) (Butkowski et al., 1989; Kleppel et al., 1989). The α3(IV) and α4(IV) chains were also found to co-localize in other tissues. For example, Sanes et al. (1990) found that these chains are found together in the basement membranes of synaptic muscle fibers but not in extrasynaptic muscle fibers, endoneurial or perineurial nerve, or artery. The α1(IV) and α2(IV) chains, however, were found in basement membrane at all these sites but were noticeably less abundant in synaptic muscle fibers. These data, taken together, suggest that α3(IV) and α4(IV) may be present in the same triple-helical molecule. The sequence similarities between α2(IV) and α4(IV) and between α1(IV) and α3(IV) suggest in turn that α3(IV) and α4(IV) may “substitute” for α1(IV) and α2(IV), respectively, in some heterotrimers. A role for α5(IV) will also need to be included in this model.

The primary structure of α4(IV) NC1, deduced herein, together with knowledge of the primary structures of the other α-chain NC1 domains, permits a comparative analysis to identify regions of NC1 that may govern specificity of protomer assembly and protomer-protomer interaction. The co-expression of five chains in GBM allows for as many as 35 possible protomer subtypes (triple-helical molecules that differ in chain composition) and as many as 630 possible combinations of adjoining protomers through NC1-NC1 interactions. However, discriminatory molecular interactions must operate during assembly of the collagen superstructure to limit the number of possibilities. The specificity of interaction must be of at least two types, one at the level of protomer assembly (chain selection) and the other at the level of protomer-protomer association.

The available evidence suggests that the structural basis

---

**Fig. 6. Northern analysis of α4(IV).** Poly(A)-enriched RNA from a 60-day-old calf was electrophoresed, transferred to a nitrocellulose membrane, and hybridized with MMS644, the α4(IV) cDNA probe. K, kidney; Li, liver. A transcript of ~10.4 kilobases was detected (arrow). The positions of RNA size markers of 1.4-9.5 kilobases are shown.

**Fig. 7.** A composite diagram depicting the conserved and variable regions among the NC1 domains of mammalian α1(IV), α2(IV), α3(IV), α4(IV), and α5(IV) chains of type IV collagen. The chains were optimally aligned as indicated in Fig. 5. The number of residues reflects the maximal number of residue positions required to achieve optimal alignment. The disulfide bonds, marked by short white bars, are arranged as they occur in α1(IV) and α2(IV) (Siebold et al., 1988). For clarity, only one of two possible disulfide configurations is shown. The arrow marks the collagenous-NC1 junction. Filled circles designate positions where identical amino acids occur in all five α-chains; open circles designate positions where a different residue is found in at least one of the chains. Prominent regions of sequence variation between the chains (located in one-half of loops C and C', within loops A and A', and at the NH₂ terminus) are designated by solid lines.
for the two types of specificity resides within the NC1 domain of each α-chain. This evidence includes the capacity of monomers to reassociate to form hexamers (Timonededa et al., 1990), the existence of at least two distinct populations of NC1 hexamer (Saus et al., 1988), and the preponderence of homodimers in the dimer fraction of dissociated hexamers (Gunwar et al., 1991). Moreover, the type IV NC1 domain has a structure analogous to the COOH-terminal globular domains of type I-III collagens which specify chain selection, registration of the COOH-terminal Gly-X-Y sequences and formation of a nucleus of triple-helical confirmation for the zipper-like folding of the collagenous domain (Engel and Prockop, 1991). On an a priori basis, conserved amino acid residues among the five NC1 domains reflect general features required for protomer assembly and protomer-protomer association whereas variable residues reflect discriminatory interactions that govern the assembly of specific protomer subtypes and the association of specific types of protomer. Conserved and variable residues were identified by comparative analysis of the five NC1 domains. To facilitate the comparison, the sequences were aligned in the context of folded NC1 monomers with disulfide bonds (Fig. 7). The arrangement of disulfide bonds in all NC1 domains was assumed to be identical to that of α1(IV) and α2(IV) (Siebold et al., 1988). The most likely candidates for specification of NC1-NC1 interactions lie within the prominent variable regions located in one-half of loops C and C′ and within loops A and A′ and in the 14 NH2-terminal residues of NC1.

Knowledge of the amino acid sequence of the NC1 domain of α4(IV) allows its evolutionary relationships to other type IV collagens to be inferred (Fig. 8). The similarity between α4(IV) and α2(IV) suggests that they arose from a common α2-like precursor chain which had already diverged from the precursor(s) of the α1, α3, and α5 chains. In view of the head-to-head genomic arrangement of their genes on human chromosome 13 (Pöschl et al., 1988), it is very likely that α1(IV) and α2(IV) arose from a common ancestor by a duplication/inversion mechanism. The α4(IV) chain must, therefore, have evolved by duplication from an α2-like precursor chain after the initial duplication/inversion event. The genomic organization of α1(IV) and α2(IV) allows coordinate regulation of both genes by common bidirectional regulatory elements. The evidence, cited above, that α3(IV) co-localizes with α4(IV) in many basement membranes, suggests that these genes may also be coordinately regulated. Study of the genomic localization of the human α4(IV) gene will indicate whether it shares transcription regulatory sequences with the α3(IV) gene which has recently been assigned to human chromosome 2 bands q35-q37 (Morrison et al., 1991b).

REFERENCES