Structure of the Human Type IV Collagenase Gene*

(Received for publication, November 16, 1989)

Firkko Huhtala‡, Louise T. Chow§, and Karl Tryggvason‡

From the Biocenter and the Department of Biochemistry, University of Oulu, Oulu SF-90570, Finland and the Department of Biochemistry, University of Rochester School of Medicine and Dentistry, Rochester, New York 14642

The structure of the gene for human 70-kDa type IV collagenase (gelatinase) was determined. Three overlapping genomic clones were isolated and shown to contain 0.4 kilobase (kb) of the 5′-flanking region, the 27-kb structural gene, and 4.5 kb of the 3′-flanking region. The gene has 13 exons that vary in length from 110 to 901 base pairs (bp) and 12 introns that range from 175 to 4350 bp. Alignment of intron locations demonstrated that introns 1–4 and 8–12 of the type IV collagenase gene coincide with intron locations in the interstitial collagenase and stromelysin genes, indicating a close structural relationship of these metalloproteinase genes. Exons 5–7 are each 174 bp in size, and each codon for one complete internal repeat that resembles the collagen-binding domains of fibronectin. The transcription initiation site was determined by primer extension and S1 nuclease analyses. Analysis of the 0.4-kb 5′-flanking region of the gene showed that, in contrast to the genes of interstitial collagenase and stromelysin, there is no TATA box or 12-O-tetradecanoylphorbol-13-acetate-responsive element present in the promoter region, whereas there are two GC boxes. There is no CAAT box, but a potential binding site (CCCCAGGCG) for the transcription factor AP-2 is located in the first exon.

Type IV collagenase (gelatinase) is a 70-kDa protein that belongs to a family of mammalian extracellular neutral metalloproteinases that are able to degrade a number of matrix proteins (1, 2). Other well-characterized members of the family are interstitial collagenase (3) specific for fibrillar collagens and stromelysin (transin) (4) that can degrade several noncollagenous matrix components in addition to type IV collagen. The existence of a 92-kDa type IV collagenase (gelatinase) has also been reported (5). Type IV collagenase is synthesized as a preproenzyme of 660 residues with a 29-residue signal peptide and an 80-residue propeptide (9, 12). The calculated molecular weight of the proenzyme form is 70,984 and that of the active enzyme is 62,067. The proenzyme can be converted to an active form by proteolytic cleavage or with organic mercurials (7, 8). The amino acid sequence is partially homologous to that of interstitial collagenase (13) and stromelysin (4, 14), also known as transin (15), but there are also distinct differences. The resemblance involves 456 residues of the 631-residue type IV procollagenase, i.e., a 192-residue amino-terminal end sequence and a 264-residue carboxyl-terminal end sequence (9, 12). Together, these 456-residue sequences have a significant identity to the entire sequences of interstitial collagenase (469 residues) and stromelysin (477 residues) (7).

However, type IV collagenase differs from the other enzymes in that it has a unique 174-residue central domain that consists of three internal repeats that resemble the collagen-binding domains of fibronectin (9, 16). Type IV collagenase is presumably required for the turnover of basement membranes in vivo, but increased secretion has also been shown to be linked with the invasive potential of tumor cells (1, 2, 17). The latter may significantly facilitate the penetration of disseminating tumor cells through tissue compartments of the body.

The genes for interstitial collagenase and stromelysin are not only structurally related, but, additionally, their gene loci are linked on the long arm of chromosome 11 (18). In contrast, we have recently shown that the type IV collagenase gene is located on 16q21 (12). This difference may in part explain the divergence of the type IV collagenase gene from the other metalloproteinase genes.

Knowledge about the structure and regulation of the gene is a prerequisite for a fundamental understanding of the role of type IV collagenase in normal and pathological states. In this study, we have isolated and determined the complete structure of the human type IV collagenase gene. The data show that the exon/intron structure resembles that of the gene for interstitial collagenase but that it also has features with pathological states such as inflammation and malignant growth.

Type IV collagenase degrades type IV collagen, the major structural component of basement membranes, in a specific manner, and it also has gelatinase activity (7, 9). The type IV collagen molecule is an ~400-nm-long triple helical rod with a large globular noncollagenous domain (NC-domain) at the carboxyl terminus. Type IV collagenase cleaves this substrate molecule at a single site 121 ± 12 nm from the amino terminus into ¼ and ¾ size fragments (10). The actual cleavage site has not been determined, but it is believed to be Gly–Ile and Gly–Leu bonds that are located within the same region of the α1(IV) and α2(IV) chains, respectively (11). The 92-kDa enzyme has also been shown to generate similar cleavage fragments (5). Type IV collagenase is synthesized as a preproenzyme of 660 residues with a 29-residue signal peptide and an 80-residue propeptide (9, 12). The calculated molecular weight of the proenzyme form is 70,984 and that of the active enzyme is 62,067. The proenzyme can be converted to an active form by proteolytic cleavage or with organic mercurials (7, 8). The amino acid sequence is partially homologous to that of interstitial collagenase (13) and stromelysin (4, 14), also known as transin (15), but there are also distinct differences. The resemblance involves 456 residues of the 631-residue type IV procollagenase, i.e., a 192-residue amino-terminal end sequence and a 264-residue carboxyl-terminal end sequence (9, 12). Together, these 456-residue sequences have a significant identity to the entire sequences of interstitial collagenase (469 residues) and stromelysin (477 residues) (7).

However, type IV collagenase differs from the other enzymes in that it has a unique 174-residue central domain that consists of three internal repeats that resemble the collagen-binding domains of fibronectin (9, 16). Type IV collagenase is presumably required for the turnover of basement membranes in vivo, but increased secretion has also been shown to be linked with the invasive potential of tumor cells (1, 2, 17). The latter may significantly facilitate the penetration of disseminating tumor cells through tissue compartments of the body.

The genes for interstitial collagenase and stromelysin are not only structurally related, but, additionally, their gene loci are linked on the long arm of chromosome 11 (18). In contrast, we have recently shown that the type IV collagenase gene is located on 16q21 (12). This difference may in part explain the divergence of the type IV collagenase gene from the other metalloproteinase genes.

Knowledge about the structure and regulation of the gene is a prerequisite for a fundamental understanding of the role of type IV collagenase in normal and pathological states. In this study, we have isolated and determined the complete structure of the human type IV collagenase gene. The data show that the exon/intron structure resembles that of the gene for interstitial collagenase but that it also has features with pathological states such as inflammation and malignant growth.
Human Type IV Collagenase Gene

FIG. 1. Summary of exon/intron structure of human type IV collagenase gene. Three overlapping human genomic clones (HC-1, HC-2, and HC-3) that span over 30 kb were isolated and characterized as described under "Experimental Procedures." Top, the exons are depicted by boxes, and they are numbered starting from the 5'-end of the gene. The black boxes correspond to the exons coding for the three internal repeats. The introns and flanking sequences are depicted as a solid line. Middle, genomic clones were subjected to restriction analysis, and restriction sites for EcoRI (E) and BamHI (B) are indicated. A scale in kilobases is shown below.

FIG. 2. Structure of human type IV collagenase gene. Three overlapping human genomic clones (HC-1, HC-2, and HC-3) that span over 30 kb were isolated and characterized as described under "Experimental Procedures." Top, the exons are depicted by boxes, and they are numbered starting from the 5'-end of the gene. The black boxes correspond to the exons coding for the three internal repeats. The introns and flanking sequences are depicted as a solid line. Middle, genomic clones were subjected to restriction analysis, and restriction sites for EcoRI (E) and BamHI (B) are indicated. A scale in kilobases is shown below.

EXPERIMENTAL PROCEDURES

Isolation and Characterization of Genomic Clones—Human genomic libraries in λ phage EMBL3 (HL1006 and HL1067J, Clontech) were screened using the 32P-labeled human type IV collagenase cDNA clone pK191 (12) as probe, and positive clones were isolated using standard methods (19). Three positive overlapping genomic clones were further characterized by restriction mapping, and suitable fragments were subcloned into M13 derivatives for sequencing. The exon, intron structure was analyzed by nucleotide sequencing and heteroduplex electron microscopy.

DNA Sequencing—The DNA sequence was determined from single-stranded DNA isolated in M13 derivatives by the dideoxy method (20) using Sequenase and T4 DNA polymerase enzymes according to the procedure recommended by the manufacturer (United States Biochemical Corp.). Either M13 universal primers or specific oligonucleotide primers derived from appropriate regions of the cDNA clones (12) were used in the sequencing reactions.

Heteroduplex Analysis—Genomic DNAs (HC-1, HC-2, and HC-3) were purified from phage using standard methods (19). The cDNA clone pK191 (12) was linearized with SfiI, which cuts in the vector 5' to the cloning site. The asymmetric localization of the cDNA in the vector allowed an unambiguous determination of the orientation of exon/intron segments in the heteroduplexes. The phage DNAs and the cDNA were denatured with NaOH, neutralized with Tris-HCl, and renatured in 50% (v/v) formamide, 0.1 M Tris-HCl, pH 8.5, 0.01 M EDTA at room temperature. The resulting heteroduplexes were mounted on parlodion grids, contrast-enhanced, and visualized by electron microscopy as described previously (21, 22). Length standards were circular single-stranded X174 DNA and renatured linear pK191 DNA in the same electron photograph.

Primer Extension and S1 Nuclease Mapping—Primer extension analysis was performed using an end-labeled 20-mer oligonucleotide complementary to the human type IV collagenase cDNA sequence (nucleotides 204–223) as primer and total RNA (30 µg) isolated from HT-1080 cells as template for reverse transcription (23). The primers were extended with reverse transcriptase under standard conditions (23). In the S1 nuclease protection assay, single-stranded M13 template DNA, containing a 2.3-kb SalI-EcoRI genomic fragment spanning the 5'-end of the type IV collagenase gene, was hybridized with the labeled nucleotide. The 640-nucleotide single-stranded radioactive probe was isolated from a denaturating agarose gel after SalI digestion. The labeled probe was used in the primer extension and S1 nuclease assays as primer.

RESULTS

Characterization of Genomic Clones—Three overlapping genomic λ phage clones (HC-1, HC-2, and HC-3) that together span ~32 kb of genomic DNA were isolated (Fig. 1) using cDNA clones coding for human type IV collagenase as a probe.
plex analysis. The results revealed that the three genomic clones, Southern blotting, and electron microscopic heteroduplexes contained the entire gene of 27 kb and, additionally, the region between the genomic clones and the cDNA clones was used to sequence to determine the exact sizes of the exons. Relevant segments of the clones were subcloned and sequenced to determine the exact sizes of the exons.

0.4 kb of the 5'-flanking and 4.5-kb of the 3'-flanking sectionally, electron microscopic analysis of heteroduplexes between synthetic oligonucleotide primers that were designed with the exception of exon 2, all translated exons start with the second or third base of a codon. Primer extension and Sl nuclease mapping analyses (see below) and nuclease-negative control with yeast tRNA (10 pg); lane 7, Sl nuclease I, primer extension with HT-1080 total RNA (30 pg); lane 6, Sl nuclease mapping with RNA recovered from HT-1080 cells known to synthesize type IV collagenase. A 20-mer oligonucleotide complementary to bases 204-223 in the gene (Fig. 3) was used as primer. The primer extension resulted in the protection of a single fragment of 415 bp of the flanking region (Fig. 3) revealed several interesting features of this gene. There was no TATA or CAAT box. However, there were two GGGCGG consensus sequences (GC boxes) for the binding of the transcription factor Sp1 (27). One GGGCGG sequence started at position -89 and a second one at position -69 (Fig. 3). We could not identify a transcription factor AP-1 binding sequence that can confer TPA induction (28), and, furthermore, a computer search did not reveal consensus sequences (see Ref. 29) for the glucocorticoid-response element, transcription factor C/EBP, CAMP inducibility (CREB), the octamer transcription factors OCT-1 and OCT-2, or the serum-response element. EB

Fig. 3. Nucleotide sequence of 5'-end of human type IV collagenase gene and coding sequence-derived amino acid sequences. The initiation for transcription was determined by primer extension and Sl nuclease assays. The bent arrows indicate the two potential transcription initiation sites. The numbering of nucleotides starts at the first transcription initiation site. The putative Sp1 binding sites (GGGCGG) are boxed. The translation start site is depicted with a straight arrow, and the cleavage site of the signal peptide is shown by a triangle. The transcription factor AP-2 consensus sequence is indicated with a broken line, and the first eight nucleotides of the first intron are indicated with a solid line.

Fig. 4. Transcription start sites of human type IV collagenase gene. Transcription start sites were determined by primer extension assays. The RNA was isolated from HT-1080 cells. Lane 1, primer extension with 10 pg of mRNA; lane 6, S1 nuclease-negative control with yeast tRNA (10 pg); lane 7, S1 nuclease assay with HT-1080 RNA (30 pg). In the sequencing reaction (lanes 2-5), the same 20-mer oligonucleotide was used as in the primer extension and Sl nuclease assays (12). These analyses established that the human type IV collagenase gene contains 13 exons that vary in size from 110 to 901 bp and introns ranging in size from 175 to 4350 bp (Figs. 1 and 2). With the exception of exon 2, all translated exons start with the second or third base of a codon. Primer extension and Sl nuclease mapping analyses (see below) and nucleotide sequencing revealed that exon 1 contains 443 bp. Exons 3 and 4 code for a region of the protein that is related to the amino-terminal end of the interstitial collagenase and stromelysin proenzymes. Exons 5-7 are all 174 bp in size, and each of them codes for one complete internal repeat that resembles the collagen-binding domains of fibronectin. Exon 8 contains sequences for the putative zinc-binding region, which is highly conserved in all metalloproteinases from bacteria to man (14, 26). Exons 8-13 code for 246 residues from the carboxyl-terminal end of the enzyme that has extensive homology to interstitial collagenase and stromelysin.

Exon/Intron Structure—The nucleotide sequences of the exons were primarily determined using "exon hopping" (25) with synthetic oligonucleotide primers that were designed based on sequences known from the cDNA clones. Additionally, electron microscopic analysis of heteroduplexes between the genomic clones and the cDNA clones was used to determine the complete exon/intron pattern of the gene. These analyses established that the human type IV collagenase gene contains 13 exons that vary in size from 110 to 901 bp and introns ranging in size from 175 to 4350 bp (Figs. 1 and 2). With the exception of exon 2, all translated exons start with the second or third base of a codon. Primer extension and Sl nuclease mapping analyses (see below) and nucleotide sequencing revealed that exon 1 contains 443 bp. Exons 3 and 4 code for a region of the protein that is related to the amino-terminal end of the interstitial collagenase and stromelysin proenzymes. Exons 5-7 are all 174 bp in size, and each of them codes for one complete internal repeat that resembles the collagen-binding domains of fibronectin. Exon 8 contains sequences for the putative zinc-binding region, which is highly conserved in all metalloproteinases from bacteria to man (14, 26). Exons 8-13 code for 246 residues from the carboxyl-terminal end of the enzyme that has extensive homology to interstitial collagenase and stromelysin. Exon 13 also contains the entire 3'-untranslated region coding sequence.

5'-End and -Flanking Region—To determine the initiation site of transcription, we employed primer extension analysis and Sl nuclease mapping with RNA recovered from HT-1080 cells known to synthesize type IV collagenase. A 20-mer oligonucleotide complementary to bases 204-223 in the gene (Fig. 3) was used as primer. The primer extension resulted in the protection of two major bands that were separated by 9 bases (Fig. 4), indicating that two sites can be utilized for the start of transcription. The first site is 290 nucleotides and the second site is 280 nucleotides upstream from the translation initiation codon (ATG) (Fig. 3). Based on the intensity of the bands (Fig. 4), the second site is the major start site. The Sl nuclease experiment resulted in the protection of a single fragment with a start site corresponding to the second site observed in the primer extension experiment (Fig. 4).

Sequencing analysis of the 5'-end untranscribed region and of 415 bp of the flanking region (Fig. 3) revealed several interesting features of this gene. There was no TATA or CAAT box. However, there were two GGGCGG consensus sequences (GC boxes) for the binding of the transcription factor Sp1 (27). One GGGCGG sequence started at position -89 and a second one at position -69 (Fig. 3). We could not identify a transcription factor AP-1 binding sequence that can confer TPA induction (28), and, furthermore, a computer search did not reveal consensus sequences (see Ref. 29) for the glucocorticoid-response element, transcription factor C/EBP, CAMP inducibility (CREB), the octamer transcription factors OCT-1 and OCT-2, or the serum-response element. However, a potential binding site consensus sequence (CCCCAGGC) for the transcription factor AP-2 (30) was present in exon 1 starting at position +157 (Fig. 3).

DISCUSSION

Characterization of the structural gene for human 70-kDa type IV collagenase demonstrates both differences from and similarities to the genes for mammalian interstitial collagenase and stromelysin. The type IV collagenase gene has 13 exons as opposed to 10 in the two other genes (Fig. 5). The gene is 27 kb, which is more than three times larger than the human (31) and rabbit (32) interstitial collagenase genes, and it is also significantly larger than the rat stromelysin (transin) genes (33, 34). The size difference is obviously not due only
Human Type IV Collagenase Gene

Human Interstitial Collagenase Gene

Rat Transin Gene

Human Fibronectin Gene

**FIG. 5.** Comparison of structure of human type IV collagenase gene with those of human interstitial collagenase (31), rat stromelysin (transin) (33), and human fibronectin (16, 35) genes. The exons are shown as shaded boxes numbered from the 5'-end, and the introns are depicted by broken lines. Exons corresponding to each other are aligned. Sizes of exons (in base pairs) are indicated below the boxes, and sizes of introns (in base pairs) are indicated above the introns. The intron sizes of the stromelysin and fibronectin genes are approximated from Refs. 33 and 35, respectively.

**FIG. 6.** Illustration of conserved location of intervening sequences in human type IV collagenase, human interstitial collagenase, and rat stromelysin (transin-1) genes. Nucleotide sequences of exons at the intron junctions are shown in open-ended boxes, with the derived amino acid sequences shown below. The introns are depicted by shaded boxes. Sequences of the interstitial collagenase and stromelysin genes are taken from Refs. 31 and 33, respectively.

**FIG. 7.** Comparison of amino acid sequences encoded by exons 5–7 in type IV collagenase gene with those encoded by exons 8 and 9 in human fibronectin gene. The two fibronectin type II repeats (FN II-1 and FN II-2) contain 1 and 2 more amino acids, respectively, than the three repeats (CLG4-1, CLG4-2, and CLG4-3) in type IV collagenase. Amino acids conserved in all repeats are boxed, with the cysteine residues in shaded boxes.

to the presence of the unique domain consisting of internal repeats, but also to the larger intron sizes. For example, introns 1–3, 8, 11, and 12 in the type IV collagenase gene are all considerably longer than their counterparts in the other metalloproteinase genes. Despite these differences, the genes clearly belong to the same family based on structural similarities in the parts of the genes that code for homologous regions of the proteins. Alignment of sequences at intron junctions (Fig. 6) demonstrates that introns 1–4 and 8–12 of the type IV collagenase gene all coincide with locations in the inter-
stital collagenase and stromelysin genes (31–34). The slight differences in the sizes (Fig. 5) of exons depend on whole codons that have been either inserted or deleted.

This study demonstrates that the three 58-residue internal repeats that resemble the collagen-binding domains of fibronectin are each encoded by a single 174-bp exon. Fibronectin contains two internal repeats of this type (type II domains), each of which is encoded by separate exons (35). The sequence identity among the three repeats in type IV collagenase is ~50%, which is of the same order as the sequence identity to the type II domains in fibronectin (Fig. 7). The cysteine residues are conserved in all repeats. It seems likely that the divergence of the type IV collagenase gene has involved an uptake of such elements from another locus. The function of the internal repeats in type IV collagenase is not known, but they possibly confer substrate specificity to the enzyme by providing binding sites for type IV collagen and gelatin.

This study shows that, despite the close intergene relationship, the 5'-flanking region of the type IV collagenase gene differs considerably from those of the genes for human and rabbit interstitial collagenase and human and rat stromelysin (32–34, 36, 37). These differences may provide an explanation for their differential gene expression. For example, cultured human melanoma cells (A2058) and fibrosarcoma cells (HT-1080) express type IV collagenase (9), but not stromelysin (4). Even the human promoter TPA does not induce expression of stromelysin in these cells (4). Furthermore, normal human bronchial epithelial cells that become malignant after transformation with the Ha-ras oncogene express type IV collagenase, but not stromelysin or interstitial collagenase (9). These observations support previous findings about linkage between type IV collagenase and the malignant phenotype (9, 17, 38–40). However, type IV collagenase is also expressed to some extent in cultured human endothelial cells, keratinocytes, and fibroblasts (9, 41).

One striking finding was that the type IV collagenase gene does not contain a TATA box, which is present in the interstitial collagenase and stromelysin genes. The lack of a TATA box has been observed in most but not all housekeeping genes, the polyoma virus late promoter, the adenovirus 2 DNA-binding protein promoter, and the SV40 late promoter (32–46). Also, genes for the basement membrane proteins, type IV collagen (47) and laminin B1 (48) and B2 chains (49), do not have a TATA box. Promoters without TATA boxes generally have heterogeneous initiation sites for transcription, as was also shown to be the case for the type IV collagenase gene in this study. Another feature of the type IV collagenase gene is the presence of two Sp1 binding sequences (GGCCGC) in the 5'-flanking region, whereas there are no such sequences in the other two metalloproteinase genes. All three metalloproteinase genes are similar in that they lack a CAAT box.

Interestingly, the type IV collagenase gene lacks the TPA-response element sequence (TGAACTCGA) that can serve as a binding site for the nuclear transcription factor AP-1 (28). This consensus sequence is present in both the interstitial collagenase and stromelysin genes at the same location, starting ~40 bp upstream from the TATA box and ~70 bp from the start site for transcription (36, 37). Both interstitial collagenase and stromelysin expression are directly induced in fibroblasts by TPA, an effect that is mediated through the AP-1 protein (28). The effects of TPA on type IV collagenase expression are as yet controversial. TPA has been reported to increase both type IV collagenase activity (41) and mRNA levels (50) in human skin fibroblasts, but the absence of induction has also been reported (9). The possible role of the AP-2 binding site (CCCCAGGC) in the first exon of the type IV collagenase gene remains to be determined. The AP-2 protein appears to mediate transcription activation through phorbol esters such as TPA and protein kinase C as well as CAMP-dependent protein kinase A (30).

REFERENCES
32. Fini, M. E., Plucinski, I. M., Mayer, A. S., Gross, R. H., and
Structure of the human type IV collagenase gene.
P Huhtala, L T Chow and K Tryggvason


Access the most updated version of this article at http://www.jbc.org/content/265/19/11077

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/265/19/11077.full.html#ref-list-1