Volume 330, number 2, 122–128 FEBS 12948 September 1993

cDNA isolation and partial gene structure of the human α4(IV) collagen chain

Manabu Sugimoto[a], Toshitaka Oohashi[b], Hidekatsu Yoshioka[b], Nobuhiko Matsuo[a], Yoshifumi Ninomiya[b,*]

[a]*Department of Ophthalmology, Okayama University Medical School, Okayama 700, Japan*
[b]*Department of Molecular Biology and Biochemistry, Okayama University Medical School, 2-5-1, Shikata-cho, Okayama 700, Japan*

Received 19 July 1993

A novel collagen IV chain, α4(IV), has recently been identified in basement membranes. We describe part of the primary structure of the human α4(IV) polypeptide for the first time, which has been determined by cloning and sequencing of cDNAs encoding 241 amino acid residues of the COL domain and 231 residues of the NC1 domain. We also characterized a genomic DNA fragment containing 4 exons coding for the entire NC1 domain. Among five known α chains of collagen IV, the α4(IV) chain is distinct from the other four chains. However, it is more similar to the α2 (IV) chain than to the α1(IV), α3(IV) and α5(IV) chains in terms of amino acid sequence homology, domain structure of polypeptides and exon/intron structure of the genes, suggesting the presence of two phylogenetically distinct subclasses of collagen IV α chains; one composed of α2 and α4 chains and the other of α1, α3 and α5 chains.

Basement membrane; Type IV Collagen; mRNA; Gene structure; cDNA

1. INTRODUCTION

Underlying most epithelium and endothelium is a thin, fuzzy layer and some fibrils that collectively make up the basement membrane. Constituents of the basement membrane have been shown to be made up of collagens, proteoglycans and other glycoproteins. Collagen IV is one of the major constituents of basement membranes [1,2]. It is a hetero or homotrimeric molecule containing three distinct domains: the 7S domain at the amino terminus, the central triple-helical(COL) domain and the NC1 domain located at the carboxyl terminus [2]. Recent techniques in protein chemistry [3–5] and molecular biology [6] have made it possible to permit unambiguous identification of new collagen chains present in high molecular aggregates in basement membranes. So far five distinct but related α chains: α1(IV), α2(IV), α3(IV), α4(IV) and α5(IV) have been designated as collagen IV chains due to their characteristic primary structure. Intriguingly, two sets of genes have been reported to co-localize in the same human chromosomal locus; α1(IV) gene(COL4A1) and α2 (IV) gene(COL4A2) sit on opposite strands on chromosome 13q34, sharing a common promoter region [7–9] and human chromosome 2q36 is the region for both α3(IV)(COL4A3) [10] and α4(IV) collagen genes(COL4A4) [11].

In this manuscript we report for the first time the primary structure of the 231 amino acid residues of the NC1 domain and the 241 residues of the COL domain of the human α4(IV) chain and the partial gene structure of COL4A4. The exon/intron structure of the 3′ part of the gene is similar to that of COL4A2 [12] but distinct from that of COL4A1 [13], COL4A3 [14] and COL4A5 [15]. This suggests that the COL4A4 gene is phylogenetically related to COL4A2 gene rather than to the other three genes encoding the α1(IV), α3(IV) and α5(IV) chains.

2. MATERIALS AND METHODS

2.1. *Cell culture and RNA blotting*

Corneal endothelial cells and lens epithelial cells were isolated from rabbit eyes and cultured as described [16]. Total cellular RNA was isolated from confluent cells with the guanidinium thiocyanate method [17]. Poly(A)$^+$ RNA was isolated by oligo(dT)-cellulose chromatography. Total RNA extracted from human lung and kidney was kindly provided from Dr. Y. Muragaki, Wakayama Medical College. For Northern-blot analysis, RNAs were electrophoresed on 1% agarose gels, blotted onto nylon filters, and hybridized with various probes using a standard method [17]. 28 S and 18 S ribosomal RNAs were used as size markers.

2.2. *cDNA cloning*

A λgt10 cDNA library (HL1047a, Clontech Laboratories Inc.) from human whole eye was screened using the previously isolated cDNA, pYKb, coding for the rabbit α4(IV) chain [11]. A few among 10 positive clones were characterized as described below. Recombinant DNA isolation, filter hybridization and phage purification were performed as described [17].

*Corresponding author. Fax: (81) (86) 222 7768.

The nucleotide sequence data reported in this paper will appear in the DDBJ, EMBL and GenBank Nucleotide Sequence Databases under accession number D17391.

Abbreviations: SSC, 0.15 M NaCl, 0.015 M sodium citrate (pH 7.0); bp, base pair(s); kb, kilobase(s)

Published by Elsevier Science Publishers B.V.

2.3. *Screening of human genomic DNA*

A total genomic DNA library (HL1006d, Clontech Laboratories Inc.) was screened for human related genomic DNAs using a *Bam*HI/*Eco*RI (476 bp) fragment of MS2 as a probe. Regular conditions were used for prehybridization and hybridization (5 × SSC, 1% *N*-laurylsarcosine at 65°C, overnight) and for washing(3 × SSC, 0.5% *N*-laurylsarcosine at 65°C for 30 min, twice and 3 × SSC at 65°C for 30 min, twice) [16]. Isolation and purification of phage DNA was performed according to standard methods [17]. Restriction enzyme fragments of isolated clones were subcloned into pBluescript vectors for further characterization.

2.4. *Nucleotide sequence analysis*

Nucleotide sequencing was performed using the dideoxy chain termination technique [18] using [^{35}S]dATP and the Sequenase kit (United States Biochemicals, Cleveland, OH) either on double-stranded vectors [19] or on single-stranded M13 vectors [20]. Sequence-specific oligonucleotides were used as primers when necessary. Some DNA fragments were deleted by exonuclease III and mung bean nuclease and self-ligated to the original vector using Kilo-Sequence Deletion kit (Takara Biochemicals, Kyoto, Japan). The nucleotide sequence from each strand was determined to confirm the data. The MacVector software was used to analyze the sequence data obtained.

3. RESULTS AND DISCUSSION

3.1. *Isolation of cDNAs encoding human α4(IV) chain*

We have previously isolated and characterized cDNAs coding for the rabbit α4(IV) chain [11]. One of the rabbit cDNAs, pYKb, was used to screen λgt10 cDNA libraries from whole eye for human α4(IV)cDNAs. As shown in Fig. 1, three cDNAs, MS2, MS3, and MS39 were isolated from the library. Complete nucleotide sequence analysis demonstrated that one reading frame remained open all through the insert(1174 nucleotides) of the MS2 clone and the deduced amino acid sequence contained glycine residues in every third amino acid (Fig. 2), typical for collagenous polypeptides. It encodes 241 residues of the triple-helical (COL) domain and 150 residues of the non-triple helical (NC1) domain. Identification of the clone was based upon complete identity between a part of the amino acid sequence deduced from the cDNA and the amino acid sequence from the junction of the COL and NC1 domain of the α4(IV) polypeptide [3,11]. This was also confirmed by a comparison with the amino acid sequence from the coding region of an exon of the human α4(IV) gene, COL4A4, located on chromosome 2 [11].

3.2. *α4(IV) transcript is larger than other collagen IV transcripts*

Unexpectedly, RNA species which specifically hybridized to the insert of the cDNA, MS2, were much larger than those hybridized to cDNAs coding for the other α-chains of collagen IV. As shown in Fig. 3, it hybridized to several species of RNA from human lung and kidney. The major two bands were approximately 10 kb and 7 kb in size. α4(IV) mRNA is expressed more in kidney than in lung. A stretch of adenyl residues were found at the 3′ end of the two cDNAs, MS39 and MS3, indicating the presence of at least two different RNA transcripts, as found in other collagen mRNAs [21,22]. MS3 harbors a fairly long stretch (approximately 2 kb) of the 3′-untranslated region, which contains several poly(A) signals (Fig. 2). These signals could be utilized and generate several mRNA species. Although we do not know the structure of the 5′ two-thirds of the mRNA, the polypeptide encoded by the three cDNAs shown in Fig. 1 appears to be similar to the other four collagen IV chains. It would be interesting to know if the mRNA size difference is due to the longer amino terminal region or to the longer 5′-untranslated region for the α4(IV) transcript.

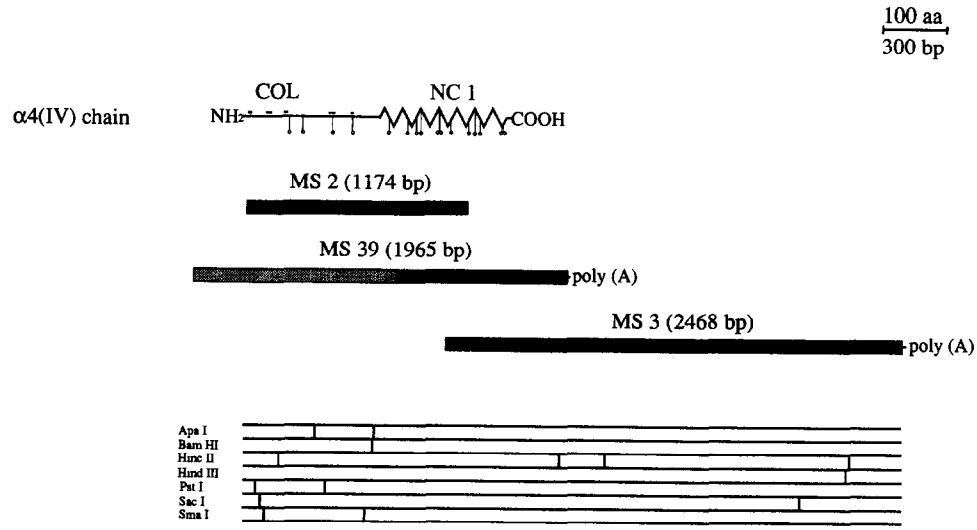

Fig. 1. Diagram representing the domain structure of the human α4(IV) collagen chain translation product as deduced from the nucleotide sequence of the cDNAs, MS2, MS39 and MS3, and the restriction enzyme maps of the clones. Closed squares above the α4(IV) domain structure indicate the location of the five imperfections of the Gly-X-Y triplet structure. Location of the cysteinyl residues are indicated by open circles. Note that the 5′ side of the MS39 (indicated by shadowing) is a copy of the intron; details are explained in section 3.

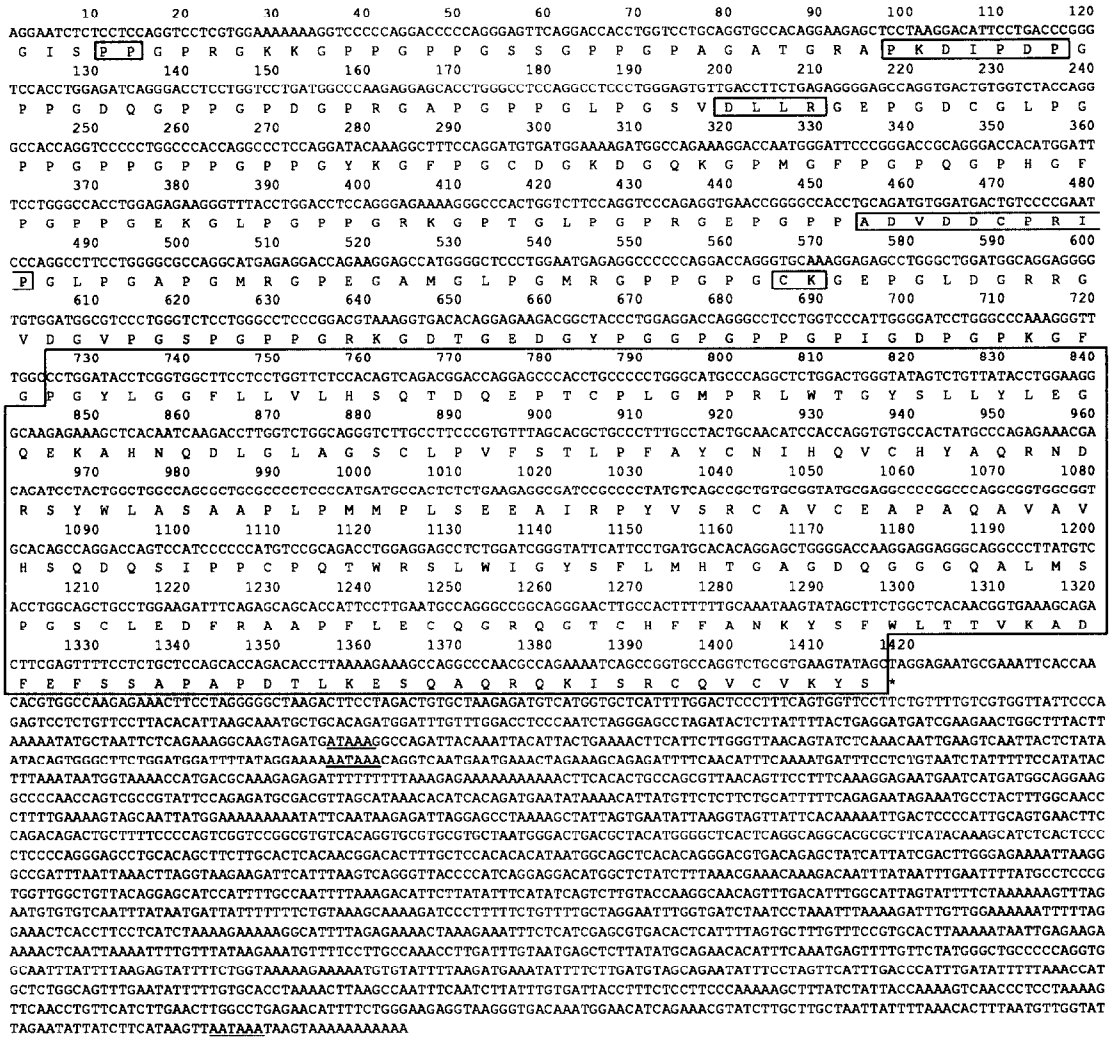

Fig. 2. Nucleotide sequence of the cDNAs coding for the α4(IV) collagen chain and amino acid sequence of the conceptual translation product. Numbers indicate nucleotide numbers counting from the 5′ nucleotide of the insert of the MS2 clone. The non-triple-helical NC1 (nucleotides 725–1,417) is indicated by the large boxed area. The position of the five imperfections in the Gly-X-Y repeat structure, a termination codon (nucleotide numbers 1,418–1,420) and three poly(A) signals in the 3′-untranslated region are indicated by small boxes, a star and short underlines, respectively.

3.3. *Primary structure of the human α4(IV) chain*

The composite nucleotide sequence of 3,559 bases from the overlapping 3 cDNAs of MS2, MS3 and MS39 and its translation product are shown in Fig. 2. The translation product covered by the 3 cDNAs is not complete (see Fig. 1). The cDNAs encode 241 amino acid residues of the triple-helical (COL) domain, 231 residues of the NC1 domain of the polypeptide of human α4(IV) chain and the 3′-untranslated region. The COL domain contains 72 Gly-X-Y triplets, 16 of which 22% are Gly-Pro-Pro triplets that stabilize collagen triple-helices, presumably utilizing hydroxyprolines located at Y positions. In the middle of the stretch of the triple-helical domain, a stretch of Gly-Pro-Pro repeated 4 times was found. Five imperfections of Gly-X-Y repeats are found in the COL domain. Four cysteinyl residues exist within COL domain: two of them (nucleotide numbers 227–229 and 296–298) are in the triple-helical domain but the other two (numbers 470–472 and 566–568) are within the two imperfections of the Gly-X-Y repeats. None of the a chains of collagen IV contains cysteinyl residues within the COL domain in this region except for the α3 chain, which contains one cysteinyl residue in triple-helical domain [23] as shown in Fig. 4. The carboxyl-terminal non-triple-helical domain (NC1) for the α4(IV) chain contains 231 amino acid residues. When the other NC1 domains are compared as shown in Fig. 5, the total length is almost the same ranging from a maximum of 232 residues for α3 and a minimum of 227 residues for α2 . Twelve cysteinyl residues are aligned in position. When the α4(IV)NC1 sequence was compared to other NC1 sequences, 60, 72, 54, and 60%

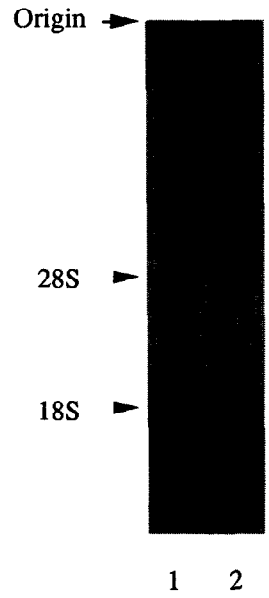

Fig. 3. Northern blot analysis of total RNA from human lung (lane 1) and kidney (lane 2) with the α4(IV)-specific cDNA probe, MS2. Twenty μg of the total RNA was applied onto a 1% agarose gel, electrophoresed and blotted onto a nylon filter. 18 S and 28 S rRNAs were used as size markers as indicated. Note that a transcript of ~ 10 kb is detected in lung and kidney.

similarity was observed with α1, α2, α3, and α5(IV)NC1s, respectively. Among the five human NC1 domains, higher homology was observed at both nucleotide and amino acid levels between the α4 and α2 than between the α4 and α1, α3 and α5NC1s. When α4(IV) chains from rabbit, human, and bovine chains are compared with each other, they show a similar domain structure. 92 and 91% sequence similarity was obtained between human and rabbit [11] and the human and bovine [24] NC1 domains at the amino acid level, respectively.

3.4. *Human α4(IV) genomic clone isolation*

A total human genomic library was screened by MS2

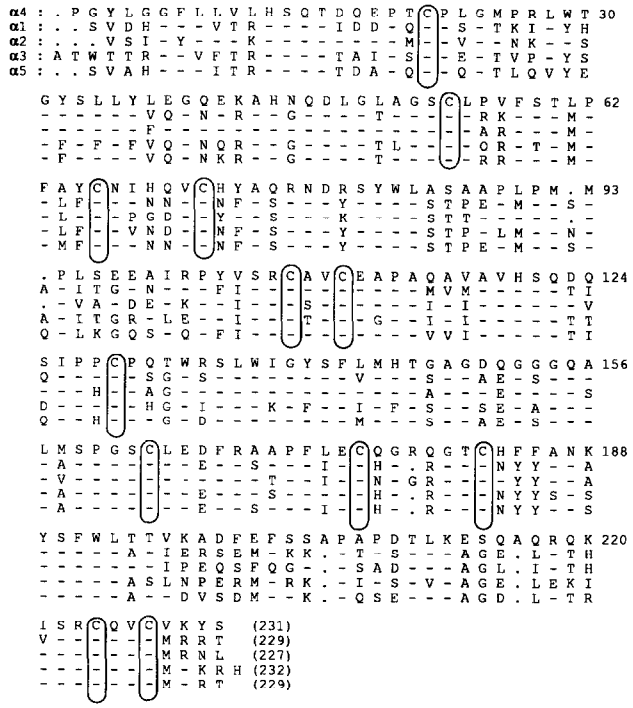

Fig. 5. Amino acid sequence comparison between the human α4(IV) NC1 domain and NC1 domains from four other human collagen IV α chains. Amino acid residues are numbered starting from the amino terminal end of the human α4(IV)NC1 domain. The cDNA-deduced amino acid sequence of the human α4(IV)NC1 (first row) domain is compared with the NC1 sequence from α1(IV) (second row) [27], α2(IV) (third row) [28], α3(IV) (fourth row) [10] and α5(IV) (fifth row) [29]. Dashes indicate that the amino acid sequence is the same as for the human α4(IV), otherwise the substituted residue is indicated. Gaps indicated by dots have been introduced to maintain alignment. Note that all the cysteinyl residues are conserved among the five a chains.

to isolate the gene fragment for the α4(IV) chain. One of the three positive clones, YN14a (15 kb in size, shown in Fig. 6), was characterized further. An *Eco*RI/*Sal*I fragment of the YN14a clone was subcloned into pBluescript. The restriction enzyme digestion profile

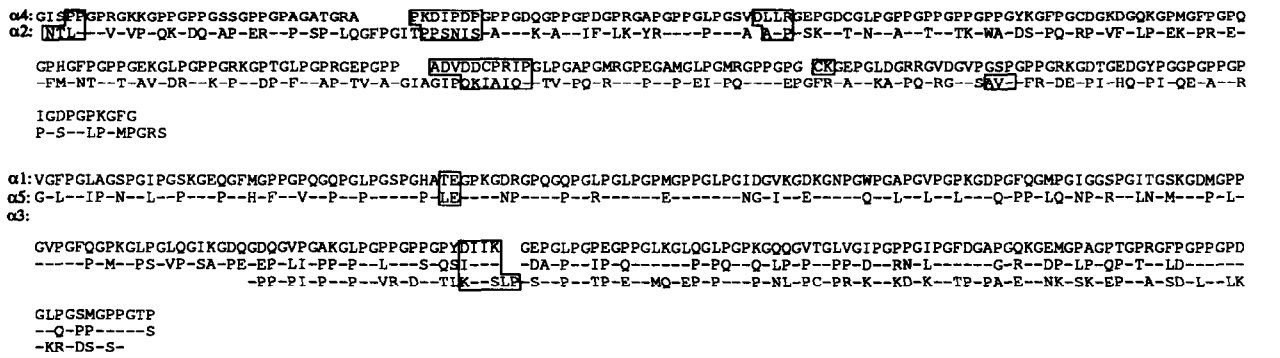

Fig. 4. Amino acid sequence comparison in COL domains between α4(IV) and the other collagen IV. The amino acid residues in the COL domains are aligned to have best homology. Dashes indicate that the amino acid sequence is the same as for the top amino acid sequence; otherwise the substituted residue is indicated. The position of the imperfection of the Gly-X-Y are indicated by boxed areas. Note that the amino acid sequence of the α4(IV) chain is similar to that of the α2 (IV) but different from that of the other three a(IV) chains with odd numbers, which have higher similarity in sequence among themselves.

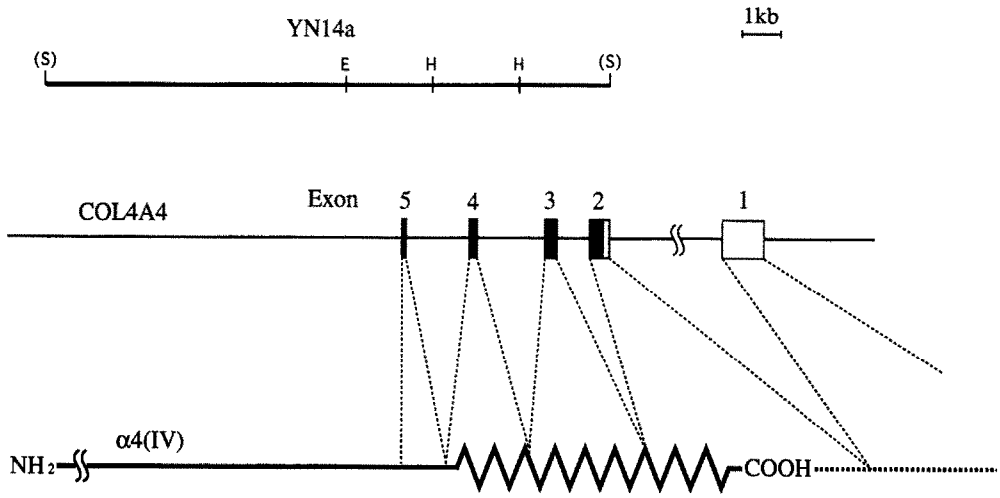

Fig. 6. Diagram indicating the partial gene structure for the α4(IV) chain. The isolated genomic clone covers the four exons that are arbitrary designated counting from the 3′-end exon. Exon 2 encodes the 3′ part of the NC1 domain and part of the 3′-untranslated region (indicated by open box), whereas exon 3 codes for the middle part of the NC1 domain. Exons 4 and 5 are described in a previous report [11]. Restriction sites are indicated by capital letters: E, *Eco*RI; H, *Hin*cII and S, *Sal*I (cloning site).

and Southern-blot analysis together with nucleotide sequence analysis demonstrated that the NC1 domain was encoded by three separate exons: E2, E3 and E4 as shown in Fig. 6.

We have previously reported [11] that a genomic clone, YK31 (15 kb in size) contained two exons encoding a carboxyl part of the COL domain and the amino end of the NC1 domain. We isolated and characterized a new genomic fragment, YN14a, and discovered that the fragment was the same as before but that it contained two more exons located at the 3′ side of the previous exons. We renamed the numbers for these exons counting from the 3′ end as shown in Fig. 6.

Sequences determined at the boundaries between exons and introns follow the consensus rules with minor exceptions (Fig. 7). Exon 3 contains 95 complete codons and one split codon at the 5′ end. Exon 2 contains 87 complete codons, a termination codon and 245 nucleotides of the 3′-untranslated region. The exon/intron structure of the α4(IV) is similar to that of the α2(IV) gene but quite different from that of the α1(IV), α3(IV) and α5(IV) genes (Fig. 8). Positions of exon/intron boundaries with split codons are well conserved among the two subclasses of the genes shown in Fig. 8. The nucleotide sequence of the 5′ side of one of the cDNAs, MS 39, was different from that of MS2. It turned out to be the same as the sequence of the intron that was located at the 5′ side of exon 3, indicating that the cDNA was a copy of one of the premature RNA species.

3.5. *Which α chains of collagen IV can constitute which collagen IV molecules?*

Collagen IV molecules, which were assumed to contain two α1(IV) chains and one α2 (IV) chain, now also harbor α3(IV), α4(IV) and α5(IV) chains. Stoichiometry of collagen IV molecules could be different among tissues, suggesting the presence of subclass molecules of collagen IV with a unique chain composition. Recent analysis of the NC1 domains from the glomerular basement membranes suggested the existence of more than three different trimers: one composed of α1(IV)NC1 and α2 (IV)NC1 domains, a second composed of α3(IV)NC1 and α4(IV)NC1, and a third with a mixed composition of all four a(IV)NC1 subunits [25]. It has also been shown that α3(IV) and α4(IV) chains are co-

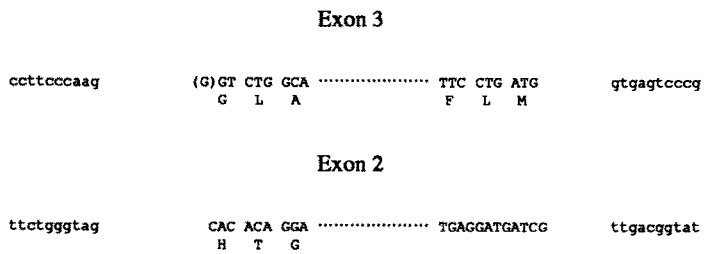

Fig. 7. Nucleotide sequences of exon 2, exon 3, and neighboring introns. Capital letters indicate the exon sequences, and small letters are the intron sequences. Parenthesis indicates the split codons.

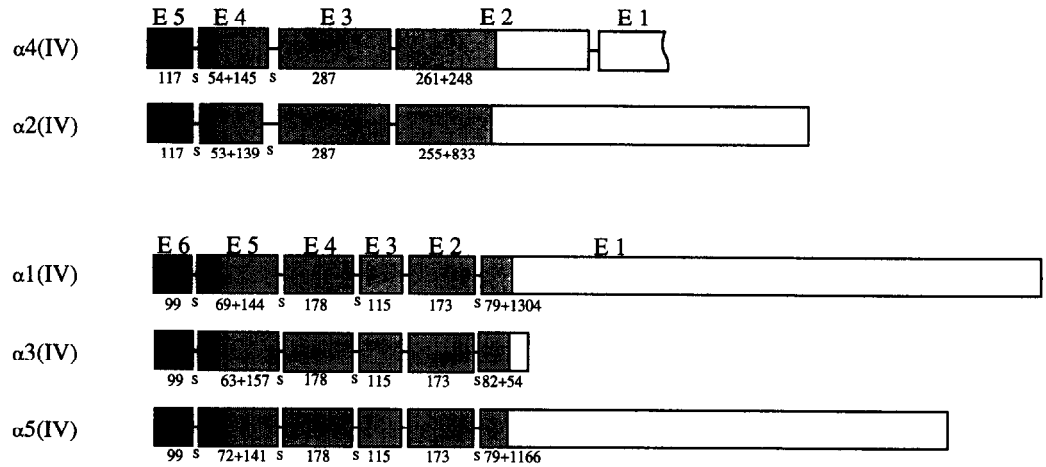

Fig. 8. Comparison of the exon structure located at the 3′ part of the collagen IV genes. Exons are indicated by boxes and introns by the interconnecting bars. Exons are numbered from the 3′ end. Numbers listed underneath each exon indicate nucleotide numbers of individual exons. When codons are split by the neighboring two exons, small letters 's' are drawn between the two exons. Note that the two classes of genes are separated; one for α2 (IV) and α4(IV) genes, and the other for α1(IV), α3(IV), and α5(IV) genes.

localized and are only present in basement membranes in the kidney, eye, cochlea, lung, and brain, whereas the α1(IV) and α2 (IV) chains are present in all basement membranes [26]. As shown in Figs. 4 and 5, sequence similarity among the five α chains of collagen IV and the locations of the imperfections within the triple-helical domains together with a distinct gene structure suggest that there are two different subclasses of α chains: one composed of α2 and α4 chains, and a second with α1, α3 and α5 chains. Theoretically these five chains can make up 35 different trimer molecules composed of unique chain combinations. The molecules can be connected in various arrangements and differ in distribution in basement membranes of different tissues. We now know the primary structure of at least the NC1 domains of all five a chains in human. The cloning of human α4(IV) cDNA provides a probe for molecular studies of human basement membrane diseases such as autosomal forms of the Alport syndrome that may involve abnormalities in this α4 chain. It is also possible to raise chain-specific monoclonal antibodies that recognize specific α chains to investigate the stoichiometry of collagen IV molecules and the tissue distribution of the molecules containing the α4(IV) chain.

Acknowledgements: This work was supported in part by Grant EY 07334 from the National Institute of Health, Scientific Research Grant 04454564 from the Ministry of Education, Science, and Culture, Japan and a grant for the Promotion of Science from the Ciba-Geigy Foundation (Japan).

REFERENCES

[1] Martin, C.R., Timpl, R. and Kuhn, K. (1988) Adv. Protein Chem. 39, 1–50.
[2] Hudson, B.G., Wieslander, J., Wisdam, B.J.and Noelken, M.E. (1989) Lab. Invest. 61, 256–269.
[3] Butkowski, R.J., Shen, G. -Q., Wieslander, J., Michael, A.F., and Fish, A.J. (1990) J. Lab. Clin. Med. 115, 365–373.
[4] Wieslander, J., Langeveld, J., Butkowski, R., Jodlowski, M., Noelken, M. and Hudson, B.G. (1985) J. Biol. Chem. 260, 8564–8570.
[5] Saus, J., Wieslander, J., Langeveld, J., Quinones, S. and Hudson, B.G. (1988) J. Biol. Chem. 263, 13374–13380.
[6] Hostikka, S.L., Eddy, R.L., Byers, M.G., Hoyhtya, M., Shows, T.B. and Tryggvason, K. (1990) Proc. Natl. Acad. Sci. USA 87, 1606–1610.
[7] Poschl, E., Pollner, R. and Kuhn, K. (1988) EMBO J. 7, 2687–2695.
[8] Burbelo, P.D., Martin, G.R. and Yamada, Y. (1988) Proc. Natl. Acad. Sci. USA 85, 9679–9682.
[9] Soininen, R., Houtari, M., Hostikka, S.L., Prockop, D.J. and Tryggvason, K. (1988) J. Biol. Chem. 263, 17217–17220.
[10] Morrison, K.E., Mariyama, M., Yang-Feng, T.L. and Reeders, S.T. (1991) Am. J. Hum. Genet. 49, 545–554.
[11] Kamagata, Y., Mattei M.-G. and Ninomiya, Y. (1992) J. Biol. Chem. 267, 23753–23758.
[12] Hostikka, S.L. and Tryggvason, K. (1987) FEBS Lett. 224, 297–305.
[13] Soininen, R., Houtari, M., Ganguly, A., Prockop, D.J. and Tryggvason, K. (1989) J. Biol. Chem. 264, 13565–13571.
[14] Quinones, S., Bernal, D., Garcia-Sogo, M., Elena, S.F. and Saus, J. (1992) J. Biol. Chem. 267, 19780–19784.
[15] Zhou, J., Hostikka, S.L., Chow, L.T. and Tryggvason, K. (1991) Genomics 9, 1–9.
[16] Yamaguchi, N., Benya, P.D., van der Rest, M. and Ninomiya, Y. (1989) J. Biol. Chem. 264, 16022–16029.
[17] Maniatis, T., Fitsch, E.F., Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, 2nd Edn., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[18] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[19] Hattori, M. and Sakaki, Y. (1986) Anal. Biochem. 152, 232–238.
[20] Norrander, J., Kempe, T. and Messing, J. (1983) Gene 26, 101–105.
[21] Myers, J.C., Dickson, L.A., de Wet, W., Bernard, M.P., Chu, M.-L., Di Liberto, M., Pepe, G., Sangiorgi, F., O. and Ramirez, F. (1983) J. Biol. Chem. 258, 10128–10135.
[22] Ninomiya, Y. and Olsen, B.R. (1984) Proc. Natl. Acad. Sci. USA 81, 3014–3018.

[23] Turner, N., Mason, P.J., Brown, R., Fox, M., Povey, S., Rees, A. and Pusey, C.D. (1992) J. Clin. Invest. 89, 592–601.
[24] Mariyama, M., Kalluri, R., Hudson, B.G. and Reeders, S.T. (1992) J. Biol. Chem. 267, 1253–1258.
[25] Johansson, C., Butkowski, R. and Wieslander, J. (1992) J. Biol. Chem. 267, 24533–24537.
[26] Kleppel, M.M., Santi, P.A., Cameron, J.D., Wieslander, J. and Michael, A.F. (1989) Am. J. Pathol. 143, 813–825.
[27] Soininen, R., Haka-Risuku, T., Prockop, D.J. and Tryggvason K. (1987) FEBS Lett. 225, 118–194.
[28] Hostikka, S.L. and Tryggvason K. (1988) J. Biol. Chem. 263, 19488–19493.
[29] Zhou, J., Hertz, J.M., Leinonen, A. and Tryggvason K. (1992) J. Biol. Chem. 267, 12475–12481.