Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4044–4048, June 1987
Biochemistry

Structure of the glycosaminoglycan domain in the type IX collagen-proteoglycan

(recombinant DNA/nucleotide sequencing/peptide sequencing/collagen genes)

DANNY MCCORMICK*, MICHEL VAN DER REST†, JUDITH GOODSHIP*‡, GUILLERMINA LOZANO*§, YOSHIFUMI NINOMIYA*, AND BJORN R. OLSEN*

*Department of Anatomy and Cellular Biology, Harvard Medical School, Boston, MA 02115; and †Genetics Unit, Shriners Hospital, Montreal, PQ, Canada H3G 1A6

Communicated by Jerome Gross, March 20, 1987

ABSTRACT **Type IX collagen represents 5–20% of the total collagen in hyaline cartilage. The molecules of this collagen are composed of three genetically distinct polypeptide subunits. One of these subunits, α2(IX), contains covalently bound glycosaminoglycan (chondroitin sulfate or dermatan sulfate). We report here on the structure of the glycosaminoglycan attachment site of type IX collagen-proteoglycan. We show, by a combination of cDNA and peptide sequencing, that the attachment region contains the sequence Gly-Ser-Ala-Asp, located within the noncollagenous domain NC3 of the α2(IX) chain. By comparing the exons encoding the NC3 domain in the α2(IX) and α1(IX) genes, we find that the exon coding for the glycosaminoglycan attachment site in the α2(IX) gene is 48 base pairs long, whereas the homologous α1(IX) exon is 33 base pairs. The NC3 domain is, therefore, five amino acid residues longer in α2(IX) than in α1(IX). The extra sequence in α2(IX), Val-Glu-Gly-Ser-Ala, provides a simple explanation for the kink observed at the NC3 domain of type IX molecules when examined by electron microscopy. The inserted block of amino acid residues also provides the NC3 domain of α2(IX) chains with a serine residue, not present in α1(IX) that serves as attachment site for a glycosaminoglycan side chain. Our data show that the amino acid sequence that surrounds the glycosylated serine residue in type IX collagen-proteoglycan differs from glycosylated sequences in noncollagenous core proteins. The data also provide strong evidence that glycosylation of type IX collagen is not a chance glycosylation of a serine residue in a noncollagenous domain, but is a specific post-translational modification of this unusual collagen molecule.**

Type IX collagen (1), which represents 5–20% of the total collagen in hyaline cartilage, is unusual in several respects. First, in type IX molecules three triple-helical collagenous domains are interspersed by noncollagenous domains (1, 2). Second, the molecules are composed of three genetically distinct polypeptide subunits. These subunits, designated α1(IX), α2(IX), and α3(IX) chains, are the products of genes whose exon structure is distinctly different from that of fibrillar collagens (3). Third, type IX collagen is also a proteoglycan and is identical to the proteoglycan PG-Lt, isolated from embryonic chicken epiphyseal cartilage (4–6). It has been established that one or perhaps two chondroitin sulfate or dermatan sulfate chains are covalently linked to the α2(IX) chain (7, 8). The glycosaminoglycan chain or chains are located at the noncollagenous domain NC3, between the amino-terminal and central collagenous domains of the molecule, at which a kink in the molecule is observed (9, 10). This evidence comes from electron microscopic observations after rotary shadowing of type IX molecules in the presence of monoclonal antibodies (9). Irwin and Mayne (9) observed binding at the NC3 kink region of type IX molecules of an antibody that recognizes stubs of chondroitin sulfate generated after digestion with chondroitinase ABC. In support of this conclusion is the finding that treatment of type IX collagen with pepsin creates a nick or gap in the α2(IX) chain in the NC3 domain that leads to loss of sulfated glycosaminoglycan from the molecule (1, 10, 11).

In the present paper we have analyzed the primary structure of the NC3 domain of α1(IX) and α2(IX) chains inferred from genomic and cDNA sequencing as well as peptide sequencing. Our data provide a simple explanation for the kink observed at the NC3 domain of type IX collagen molecules (9, 10) and demonstrate that the NC3 domain of the α2(IX) chain contains an attachment site for a single glycosaminoglycan chain within the amino acid sequence Gly-Ser-Ala-Asp.

MATERIALS AND METHODS

Preparation of RNA and cDNA Synthesis. RNA for cDNA synthesis was extracted from 17-day chicken embryo sternal cartilage by the guanidine thiocyanate method of Chirgwin *et al.* (12). Poly(A)$^+$ RNA was then selected by oligo(dT) cellulose chromatography (13).

The synthesis of cDNA was performed using a synthetic 18-mer deoxynucleotide primer (Applied Biosystems, Foster City, CA, 380A DNA synthesizer). The sequence of this primer is identical to the 5′ end of the coding sequence of exon 21 of the α2(IX) collagen gene, as inferred from sequence analysis of the genomic clone GL858 (3). Synthesis of the first strand of cDNA was carried out with 7.5 μg of poly(A)$^+$ RNA as described by Okayama and Berg (14); however, incubation of the reaction mixture was carried out at 42°C, and human placental ribonuclease inhibitor was added to a final concentration of 72 units/ml. Second-strand synthesis was performed essentially as described by the supplier of the cDNA synthesis kit (Amersham), but included *Escherichia coli* DNA polymerase I at a final concentration of 1.2 μg/ml, as well as [α-^{32}P]dCTP at a specific activity of 200 μCi/ml (1 Ci = 37 GBq) in the reaction mixture. Blunt-ended cDNA molecules were generated by incubation of the cDNA with 10 units of mung bean nuclease in 30 mM NaOAc, pH 4.6/50 mM NaCl/10 mM $ZnCl_2$ in 100 μl, for 10 min at 37°C (15). The cDNA was then methylated, ligated to *Eco*RI linkers, and religated into *Eco*RI digested, dephosphorylated arms of the bacteriophage λgt10.

Screening of Chicken cDNA and Genomic DNA Libraries. For screening of the cDNA library the hybridization probe

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡Present address: Department of Paediatric Genetics, Institute of Child Health, London, England.
§Present address: Department of Molecular Biology, Princeton University, Princeton, NJ 08544.

was a nick-translated 1308-base-pair (bp) *Pst* I–*Eco*RI restriction fragment of the genomic clone GL858. This probe contains intronic sequences as well as exons 23–27 from the α2(IX) gene. Approximately 8500 plaques were screened, and 4 positive recombinant phages were isolated. One of these, pDM222, is described in the present paper.

The nick-translated insert of the recombinant plasmid pYN1738 (2) was used to screen a genomic library for fragments of the α1(IX) collagen gene. The genomic library was constructed by partially digesting chicken DNA with *Eco*RI and cloning the fragments into bacteriophage λ Charon 4A (G.L. and G. Vasios, unpublished data).

Filters were screened using hybridization probes in sarcosyl (16).

Construction of Plasmid Subclones and DNA Sequencing. The DNAs of the recombinant phages YN910 and GL858 (3) were digested with *Eco*RI or *Bam*HI, and restriction fragments were subcloned in pBR322 as described (3). Nucleotide sequence analysis of the cDNA pDM222 and GL858 subclones was performed using the dideoxy chain-termination technique (17, 18). The chemical cleavage method was used to sequence a subclone of YN910 (19).

Purification of the Pepsin Fragments Containing α2(IX) COL2 and COL3 Domains and Their Tryptic Fragments. These fragments, called HMW C3 and HMW C4 in former publications, were purified as described by Reese *et al.* (10). The fragments were digested with *L*-1-tosylamido-2-phenylethyl chloromethyl ketone-trypsin (Millipore) in a 1:30 (wt/wt) enzyme/substrate ratio in 0.2 M $(NH_4)HCO_3$ at 37°C for 4 hr and purified by reversed-phase HPLC in two steps: first, 9 mM trifluoroacetic acid and second, 10 mM heptafluorobutyric acid as the ion-pairing agent, as described (20).

Amino Acid Analysis. Dried aliquots of the HPLC fractions were hydrolyzed for 20 hr at 110°C in gaseous 6 M HCl under nitrogen vacuum, and their amino acid compositions were determined on a DIONEX D-500 automatic amino acid analyzer as described (21).

Amino Acid Sequence Analysis. Amino acid sequences were determined by automated Edman degradation in an Applied Biosystem 470A gas-phase sequencer, using the trifluoroacetic acid conversion program. The phenylthiohydantoin derivatives of the amino acids were identified by HPLC as described by Lazure *et al.* (22).

RESULTS AND DISCUSSION

We have reported (3) the isolation and partial nucleotide sequence analysis of a chicken genomic clone, GL858, that contains coding sequences of the α2(IX) collagen chain. Determination of ≈10 kilobases of continuous nucleotide sequences within this clone (unpublished data) and search of this sequence for the presence of collagenous exons allowed a tentative assignment of exons coding for the NC3 domain of α2(IX). To definitively identify the exon-intron junctions in the NC3 region of the gene and to determine the amino acid sequence of NC3 and the flanking collagenous domains COL2 and COL3, we have analyzed an α2(IX) cDNA, pDM222. This cDNA was constructed using a synthetic 18-mer oligonucleotide primer complementary to the 5′ end of exon 21 (as counted from the 3′ end of the gene).

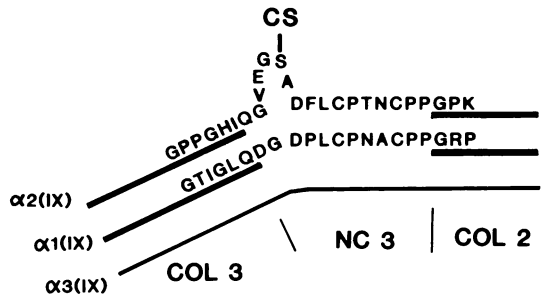

FIG. 2. Diagram showing the amino acid sequences (one-letter code) of the α1(IX) and α2(IX) collagen chains within the NC3 and parts of the neighboring COL2 and COL3 domains. The α2(IX) sequence was derived from a combination of cDNA, genomic DNA, and tryptic peptide sequences. The α1(IX) sequence is based on the sequence of the cDNA pYN1738 (2), and it was confirmed by sequencing the genomic clone YN910. Note the presence of a serine residue in α2(IX) NC3, representing the attachment site for a glycosaminoglycan side chain (CS) within the molecule.

The cDNA pDM222 encodes the amino-terminal region of the α2(IX) COL2 domain, the NC3 domain, and the carboxyl-terminal part of the COL3 domain (Fig. 1). A comparison of the cDNA sequence with the gene sequence shows that the cDNA extends from the oligonucleotide primer (5′ end of exon 21) to the 5′ end of exon 25.

Translation of the nucleotide sequence of pDM222 (Fig. 1) shows that the NC3 domain of α2(IX) consists of 17 amino acid residues in a noncollagenous sequence. This sequence has two interesting features. First, the domain is 5 amino acid residues longer in the α2(IX) than in the α1(IX) chain (Fig. 2). This may provide a simple explanation for the sharp kink observed at the NC3 domain of type IX molecules seen by electron microscopy (9, 10). Second, the α2(IX) sequence contains a single serine residue that must serve as the attachment site for the glycosaminoglycan chain in type IX

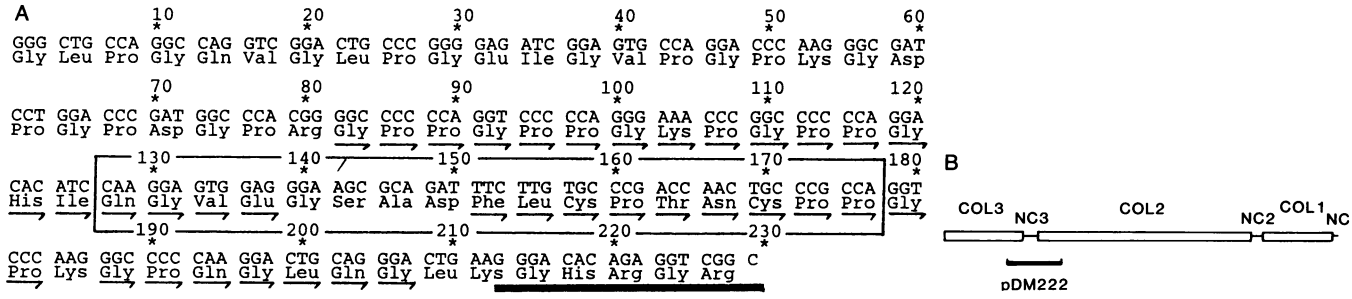

FIG. 1. (*A*) Nucleotide sequence of the cDNA pDM222 and the amino acid sequence of the conceptual translation product. Noncollagenous sequence of NC3, boxed-in area. Short horizontal arrows, amino acid residues that were confirmed by direct amino acid sequencing of two tryptic peptides. For the second peptide, no phenylthiohydantoin-amino acid derivatives could be recovered in cycle 12. The nucleotide sequence shows a lysine residue in this position. We believe that hydroxylation and glycosylation of this residue explains the "hole" in the amino acid sequence in this cycle. This is further supported by the presence of hydroxylysine, in the appropriate concentration, in the acid hydrolysate of this peptide. The sequence of the oligonucleotide primer is underlined. (*B*) Diagram showing the location of the cDNA pDM222 relative to the collagenous (COL1, COL2, and COL3) and noncollagenous (NC1, NC2, and NC3) domains of the α2(IX) collagen chain.

collagen. In contrast, the shorter NC3 domain of the α1(IX) chain contains neither serine nor threonine.

To provide more direct evidence for the involvement of the serine-containing sequence of α2(IX) NC3 (Fig. 2) in glycosaminoglycan attachment, we have determined the amino acid sequence of two tryptic peptides isolated from the fragment

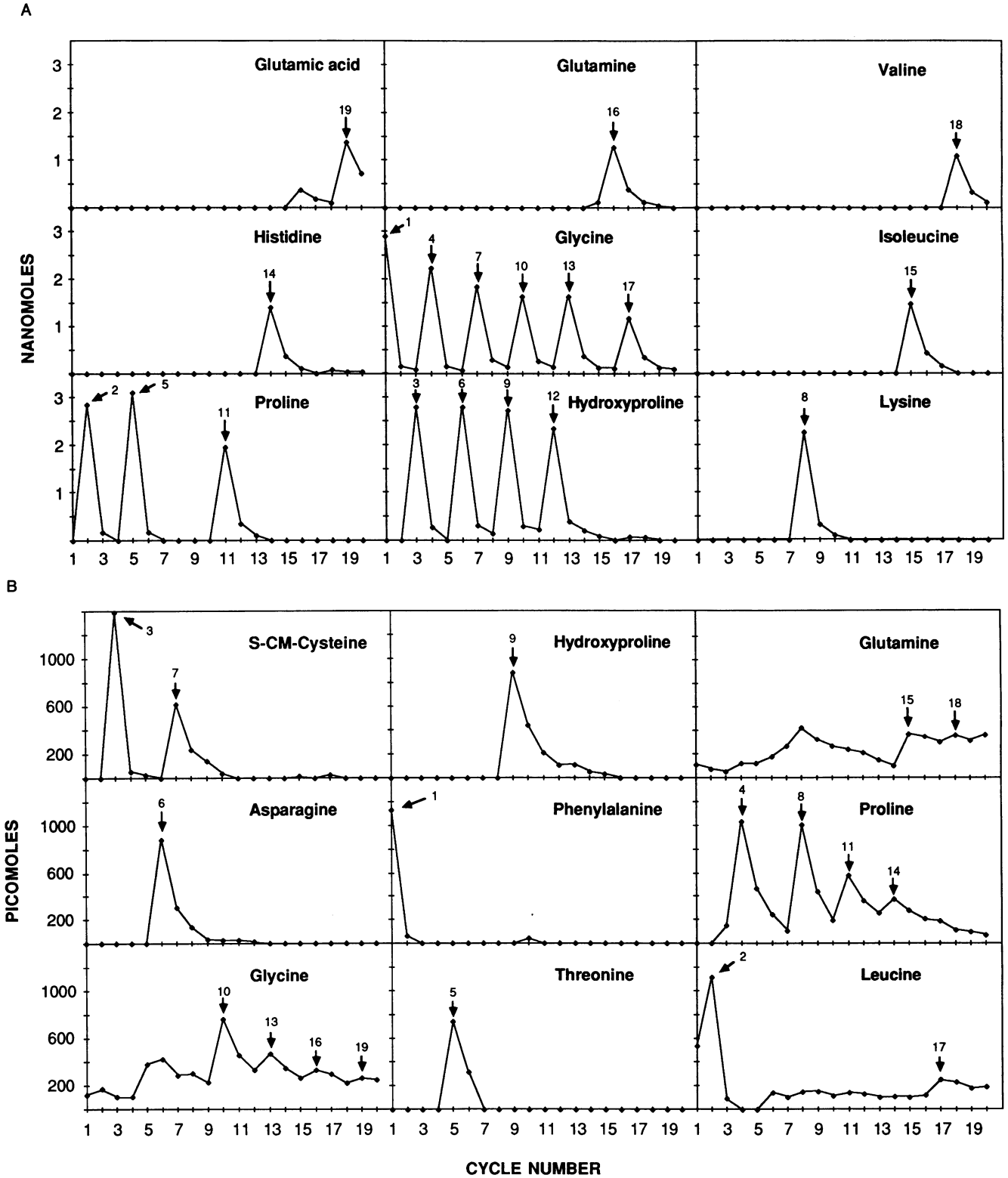

FIG. 3. Amino acid sequence analysis of tryptic peptides isolated from the C4 (*A*) and C3 (*B*) components of HMW (1). For the sequence analyses nanomole quantities of peptides were subjected to several cycles in the gas-phase sequencer. The recovery (in picomoles or nanomoles) of phenylthiohydantoin derivatives of each of nine amino acid residues is given for each cycle. The cycle numbers are indicated along the abscissas of the diagrams for each kind of amino acid residue. The assignment of different residues to each of the cycles is indicated by numbers placed above the peaks in the diagrams. For the tryptic peptide isolated from C4 (*A*) the sequence analysis was carried out to the carboxyl-terminal amino acid residue.

HMW. HMW is a fragment of type IX collagen (1) isolated by extraction of cartilage with pepsin (10, 11). Data has demonstrated that HMW does not contain chondroitin sulfate (10, 11) and that the α2(IX) chain within HMW is cleaved within the NC3 domain (10, 23), whereas the α1 and α3(IX) chains are not cleaved. Thus, there are (at least) two pepsin-sensitive sites within the α2(IX) NC3 domain, and the glycosaminoglycan chain must be attached to one of the amino acid residues that is removed by pepsin.

We have shown (23) that the peptide components C4 and C3 (10) within HMW represent the COL3 and COL2 portions of the α2(IX) chain, respectively. To obtain amino acid sequences from the α2(IX) chain, we isolated and sequenced the carboxyl-terminal tryptic peptide of C4 and the amino-terminal tryptic peptide of C3. In initial experiments, direct amino-terminal sequencing of C3 was attempted twice with two separate preparations. With a crude preparation, multiple sequences were observed, while, with an HPLC purified preparation, a double sequence was observed. This suggested the presence of multiple pepsin cleavage sites and/or contaminating or cross-linked peptides. Therefore, the amino-terminal tryptic peptide of C3 was isolated by HPLC (using *S*-carboxymethyl-cysteine as a compositional marker for the peptide) and sequenced. The carboxyl-terminal tryptic peptide of the C4 component was identified by comparing the amino acid compositions of all tryptic peptides from C4 (1) with those predicted from the α2(IX) cDNA pDM222 (Fig. 1). As demonstrated in Figs. 1 and 3, the sequence analyses showed that the α2(IX) chain of HMW contains a gap between C4 and C3 with four amino acid residues, Gly-Ser-Ala-Asp, being removed by pepsin. Since the removal of these four residues also leads to a loss of chondroitin sulfate (see above), we conclude that the glycosaminoglycan chain is attached to one of the four residues in this sequence.

To investigate the structural basis for the different sizes of the α2(IX) and α1(IX) NC3 domains on the gene level, we compared the nucleotide sequences of the α1 and α2(IX) genes in the NC3-coding region. As discussed above, extensive sequence analysis of the genomic clone GL858 and comparison with the cDNA pDM222 allowed identification of the appropriate exons in the α2(IX) gene. We have reported (3) the isolation of a genomic clone, YN623, containing coding sequences of the α1(IX) chain. Unfortunately, YN623 codes only for the two most 3′ exons in the α1 gene. Therefore, we used the cDNA pYN1738 (2) to screen a chicken genomic library for additional α1(IX) clones. One of these clones, YN910, contained sequences coding for the α1(IX) NC3 domain and was characterized further.

We have completely sequenced a 2.3-kilobase *Eco*RI fragment of YN910. A comparison of this sequence with that of the cDNA pYN1738 (2) allowed unambiguous identification of the exons that code for NC3 and the surrounding regions of the α1(IX) chain. The data are summarized in Fig. 4. As seen from this figure, the NC3 domain of both α1 and α2(IX) chains is encoded by two exons. Each of these exons, numbered 22 and 23 as counted from the 3′ ends of the genes, encodes both collagenous and noncollagenous sequences. Exon 22 is 57 bp long in both genes. However, exon 23 is 48 bp in the α2(IX) gene and 33 bp in the α1(IX) gene. Therefore, the difference in the length of NC3 in the two chains is due entirely to the difference in size of exon 23 (Fig. 2). A comparison of the nucleotide and derived amino acid sequences of exon 23 in the two genes suggests that the 5′ and 3′ ends of exon 23 are homologous between the two genes, and that the α2 exon contains an internal insertion of 5 amino acid residues relative to the α1 exon (Fig. 4). This finding suggests that the attachment of a glycosaminoglycan side chain to the α2(IX) collagen chain does not represent a chance glycosylation of a serine residue in a noncollagenous sequence but is a specific post-translational modification of type IX collagen.

The amino acid sequence around the serine residue of the α2(IX) NC3 domain is clearly different from sequences around glycosaminoglycan attachment sites in noncollagenous core proteins (24, 25). It is known that the Ser-Gly dipeptide sequence can serve as an efficient acceptor substrate for xylosyltransferase *in vitro* (26), and Ser-Gly repeats are characteristic features of core proteins that contain multiple glycosaminoglycan side chains (24, 25, 27). We do not yet know whether the sequence Gly-Ser-Ala, present in the α2(IX) NC3 domain, serves as an acceptor substrate for xylosyltransferase *in vitro*. Obviously, it will be important to compare such a peptide with Ser-Gly-containing peptides in xylosyltransferase acceptor assays. Until the results of such assays are available, we cannot exclude the possibility that glycosylation of type IX collagen is catalyzed by an enzyme that is different from the xylosyltransferase that is responsible for glycosylation of Ser-Gly sequences in other proteoglycan core proteins. The establishment of the glycos-

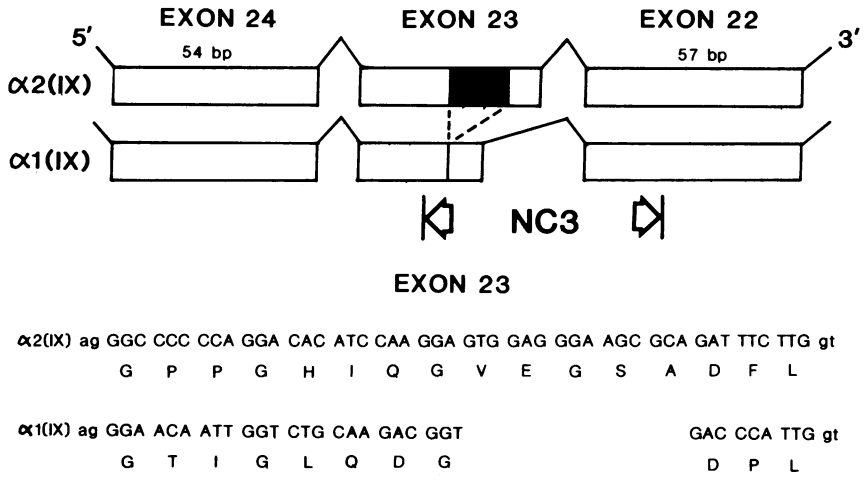

FIG. 4. Diagram showing the exon structure of the α1(IX) and α2(IX) genes for the region encoding the NC3 domain of type IX collagen. The NC3 domain is encoded by exons 22 and 23. Exons 22 and 24 have the same size in the two genes. However, while exon 23 is 48 bp long in the α2(IX) gene, it is only 33 bp in the α1(IX) gene. The nucleotide sequences of the two exons 23 are shown below the diagram together with the corresponding amino acid sequences (one-letter code). Maximal sequence homology is obtained when the two sequences are aligned so that the α2(IX) sequence contains a 15-nucleotide (5-amino acid) insertion relative to the α1(IX) sequence. The position of this insertion in exon 23 of the α2(IX) gene is indicated by the black box.

aminoglycan attachment site sequence of type IX collagen as reported here should allow detailed studies on the glycosylation of the molecule and the role of the proteoglycan domain of this extracellular matrix component.

This work was supported in part by Research Grant AR36820 from the National Institutes of Health, by Grant MT-7796 from the Medical Research Council of Canada, by the Shriners of North America, and by King Edward's Hospital Fund (J.G.). M.v.d.R. is a senior scholar of the Fonds de la Recherche en Sante du Quebec.

1. van der Rest, M., Mayne, R., Ninomiya, Y., Seidah, N. G., Chretien, M. & Olsen, B. R. (1985) *J. Biol. Chem.* **260,** 220–225.
2. Ninomiya Y. & Olsen, B. R. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3014–3018.
3. Lozano, G., Ninomiya, Y., Thompson, H. & Olsen, B. R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4050–4054.
4. Noro, A., Kimata, K., Oike, Y., Shinomura, T., Maeda, N., Yano, S., Takahashi, N. & Suzuki, S. (1983) *J. Biol. Chem.* **258,** 9323–9331.
5. Vaughan, L., Winterhalter, K. H. & Bruckner, P. (1985) *J. Biol. Chem.* **260,** 4758–4763.
6. Bruckner, P., Vaughan, L. & Winterhalter, K. H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2608–2612.
7. Huber, S., van der Rest, M., Bruckner, P., Rodriguez, E., Winterhalter, K. H. & Vaughan, L. (1985) *J. Biol. Chem.* **261,** 5965–5968.
8. Konomi, H., Seyer, J. M., Ninomiya, Y. & Olsen, B. R. (1986) *J. Biol. Chem.* **261,** 6742–6746.
9. Irwin, M. H. & Mayne, R. (1986) *J. Biol. Chem.* **261,** 16281–16288.
10. Reese, C. A., Wiedemann, H., Kuhn, K. & Mayne, R. (1982) *Biochemistry* **21,** 826–830.
11. Reese, C. A. & Mayne, R. (1981) *Biochemistry* **20,** 5443–5448.
12. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299.
13. Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 1408–1412.
14. Okayama, H. & Berg, P. (1982) *Mol. Cell. Biol.* **2,** 161–170.
15. Kroeker, W. D. & Kowalski, D. (1978) *Biochemistry* **17,** 3236–3243.
16. Overbeck, P. A., Merlino, G. T., Peters, N. K., Cohn, V. H., Moore, G. P. & Kleinsmith, L. J. (1981) *Biochim. Biophys. Acta* **656,** 195–205.
17. Sanger, R., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
18. Norrander, J., Kempe, T. & Messing, J. (1983) *Gene* **26,** 101–105.
19. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
20. van der Rest, M. & Fietzek, P. P. (1982) *Eur. J. Biochem.* **125,** 491–496.
21. Ninomiya, Y., Showalter, A. M., van der Rest, M., Seidah, N. G., Chretien, M. & Olsen, B. R. (1984) *Biochemistry* **23,** 617–624.
22. Lazure, C., Seidah, N. G., Chretien, M., Lallier, R. & St.-Pierre, S. (1983) *Can. J. Biochem.* **61,** 287–292.
23. Ninomiya, Y., van der Rest, M., Mayne, R., Lozano, G. & Olsen, B. R. (1985) *Biochemistry* **24,** 4223–4229.
24. Ruoslahti, E., Bourdon, M. & Krusius, T. (1986) *Functions of the Proteoglycans*, Ciba Foundation Symposium 124 (Wiley, Chichester), pp. 260–271.
25. Bourdon, M., Krusius, T., Campbell, S., Schwartz, N. & Ruoslahti, E. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 3194–3198.
26. Coudron, C., Koerner, T., Jacobsson, I., Roden, L. & Schwartz, N. B. (1980) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **39,** 1671.
27. Bourdon, M. A., Oldberg, A., Pierschbacher, M. & Ruoslahti, E. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1321–1325.