THE JOURNAL OF BIOLOGICAL CHEMISTRY

Vol. 265, No. 12, Issue of April 25, pp. 6884–6889, 1990
Printed in U.S.A.

Sequence of Human Syndecan Indicates a Novel Gene Family of Integral Membrane Proteoglycans*

(Received for publication, December 6, 1989)

Markku Mali, Panu Jaakkola, Anna-Maija Arvilommi, and Markku Jalkanen‡

From the Department of Medical Biochemistry, Institute of Biomedicine, University of Turku, Turku SF-20520, Finland

The structure of human syndecan, an integral membrane proteoglycan, has been determined by cloning its full-length cDNA, which codes for the entire 310-amino acid-long core protein, including the NH_2-terminal signal peptide. Similar to mouse syndecan (Saunders, S., Jalkanen, M., O'Farrell, S., and Bernfield, M. (1989) *J. Cell Biol.* 108, 1547–1556), the core protein of human syndecan can be divided into three domains: a matrix-interacting ectodomain containing putative glycosaminoglycan attachment sites, a 25-residue hydrophobic membrane-spanning domain, and a 34-residue cytoplasmic domain. Several interesting conserved structures were revealed by comparing the human syndecan sequence to the murine one. (i) Although the ectodomains are only 70% identical, all putative glycosaminoglycan attachment sites are identical (two of them belong to the consensus sequence SG*X*G and three others to (E/D)GSG(E/D), as are also (ii) the single putative *N*-glycosylation site and (iii) the proteinase-sensitive dibasic RK site adjacent to the extracellular face of the transmembrane domain. Furthermore, (iv) the transmembrane domain is 96% identical, as the only change in human syndecan was an alteration of an alanine residue to glycine; and finally, (v) the cytoplasmic domain is 100% identical, including 3 identically located tyrosine residues. Comparison of transmembrane and cytoplasmic domains to a third cell-surface proteoglycan, 48K5 from human lung fibroblasts (Marynen, P., Zhang, J., Cassiman, J., Vanden Berghe, H., and David, C. (1989) *J. Biol. Chem.* 264, 7017–7024), indicates that the transmembrane and cytoplasmic domains are similar also in this molecule regardless of the presence of a totally nonhomologous ectodomain. Thus, the transmembrane and cytoplasmic domains are unique for these cell-surface proteoglycans, which we propose to be members of a novel gene family of syndecans.

Matrix receptors mediate cell attachment to the extracellular matrix (ECM),[1] which may participate in a variety of such cellular phenomena as proliferation, differentiation, and transformation. The response of cells to changes in the ECM is thus based on the use of these matrix receptors, which recognize a matrix ligand and initiate transmission of a signal intracellularly. Although this cascade is so far largely unknown, analysis of the structure of matrix receptors will reveal some clues as to how these receptors may function. One well-characterized family of matrix receptors are the integrins, which are composed of a variety of α- and β-subunits (Hynes, 1987; Ruoslahti and Pierschbacher, 1987). The classical matrix receptor of this family is the fibronectin receptor (Argraves *et al.*, 1987), which mediates cell binding via its β-subunit to the RGD sequence of the cell-binding domain of fibronectin (Pytela *et al.*, 1985; D'Souza *et al.*, 1988). The cytoplasmic domain of the integrin β-subunit has been shown to be phosphorylated by a retrovirus-encoded protein-tyrosine kinase (Hirst *et al.*, 1986) or protein kinase C (Freed *et al.*, 1989), and it has been postulated that this phosphorylation of the cytoplasmic domain may regulate the affinity of integrins for cytoskeletal components (Freed *et al.*, 1989). One of these components has been recently identified as the 100-kD protein fibulin (Argraves *et al.*, 1989).

The other well-characterized matrix receptor is the cell-surface proteoglycan syndecan (Saunders *et al.*, 1989). It was originally isolated from a mouse mammary epithelial cell line (NMuMG) as a hybrid proteoglycan containing both heparan sulfate and chondroitin sulfate (Rapraeger *et al.*, 1985; Jalkanen *et al.*, 1985). In these cells, syndecan is polarized at their basolateral epithelial surfaces, where it links the cytoskeleton to the ECM (Rapraeger *et al.*, 1986). Cells, however, can release themselves from this interaction by a proteolytic cleavage that results in the shedding of the matrix-binding ectodomain from the membrane-binding domain (Jalkanen *et al.*, 1987). As the ectodomain of syndecan contains glycosaminoglycans (GAG), like heparan sulfate, it can bind to a variety of ECM molecules including fibrils of collagen types I, III, and V (Koda *et al.*, 1985), fibronectin (Saunders and Bernfield, 1988), and thrombospondin (Sun *et al.*, 1989). The recent cloning of syndecan has indicated that it has a 34-residue cytoplasmic domain which contains 3 tyrosine residues (Saunders *et al.*, 1989). No information is available if cells can modify these tyrosines, but the similarity of the cytoplasmic domain in size (but not in sequence) with the integrin β-subunit suggests some similarities in their function as matrix receptors.

In this paper, we present the cDNA and protein sequences of human syndecan from a normal human breast cell line (HBL-100). The preservation of the human syndecan protein sequence compared to that of mouse syndecan was 100, 96, and 70% for the cytoplasmic domain, the membrane domain, and the ectodomain, respectively. This indicates that the primary structure of both the cytoplasmic and membrane domains must be closely related to syndecan function. In the

* This work was supported by the Academy of Finland and the Finnish Cancer Union. 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) J05392.

‡ To whom correspondence should be addressed: Dept. of Medical Biochemistry, University of Turku, Kiinamyllynk. 10, Turku SF-20520, Finland. Tel.: 358-21-6337444; Fax: 358-21-331126. BITNET "BLL-MJ@FINTUVM."

[1] The abbreviations used are: ECM, extracellular matrix; GAG, glycosaminoglycan(s); bp, base pairs; kb, kilobase pair(s).

ectodomain, preserved areas were all the putative glycosaminoglycan attachment sites and also the proteinase-sensitive dibasic sequence adjacent to the extracellular face of the transmembrane domain. Comparison of human syndecan to another human cell-surface proteoglycan, 48K5 from human lung fibroblasts (Marynen *et al.*, 1989), indicates a similar kind of preservation of the membrane and cytoplasmic domains in this molecule, although it has a totally nonhomologous ectodomain. Thus, we propose that these molecules belong to the same family of integral membrane proteoglycans, the syndecans. They can provide different kinds of adhesive properties for cells as the ectodomain can vary both in the size of the core (Lories *et al.*, 1989) and in the type of glycosylation (Sanderson and Bernfield, 1988). For their activity as a signal mediator, they may use a similar, but so far largely unknown, mechanism related to the conserved structure of the transmembrane and cytoplasmic domains.

MATERIALS AND METHODS

cDNA Libraries—The human breast cell line HBL-100 (American Type Culture Collection) was maintained in bicarbonate-buffered Dulbecco's modified Eagle's medium (GIBCO) in the presence of 10% fetal calf serum (GIBCO). For preparation of RNA, cells were plated on four 80-cm^2 Nunc tissue culture bottles. After cells were grown to 80% confluency (4 days), cells were briefly washed with ice-cold phosphate-buffered saline and solubilized in 4 M GIT buffer (4 M guanidine isothiocyanate in 5 mM sodium citrate (pH 7.0), 0.1 M β-mercaptoethanol, and 0.5% *N*-laurylsarcosine). Total RNA was prepared by CsCl density centrifugation (Chirgwin *et al.*, 1979). Poly(A) RNA was purified by chromatography on oligo(dT)-cellulose (type 7; Pharmacia LKB Biotechnology Inc.). cDNA was prepared using the RNase H method (Cubler and Hoffman, 1983) according to the Bethesda Research Laboratories cDNA Synthesis System's Instruction Manual. cDNA was blunt-ended with T4 DNA polymerase (Boehringer Mannheim). Blunt-ended cDNA was size-fractionated by 1% agarose gel electrophoresis, and cDNA fragments over 500 bp were isolated by isotachophoresis (Öfverstedt *et al.*, 1984). cDNA was inserted into the dephosphorylated (calf intestine phosphatase; Boehringer Mannheim) *Sma*I site of the pUC19 plasmid by T4 DNA ligase in the presence of 1 mM hexamminecobolt chloride. Ligated recombinant plasmids were transformed into *Escherichia coli* dH5α by the method of Hanahan (1983). A primer-extended cDNA library was made as described above, except that priming was carried out with a synthetic primer, 5′-GGAGTGGAAGGTCAGCTTG-3′, specific for the 5′ end of clone hsyn4 (bases 719–738 in Fig. 2).

Isolation of Human Syndecan cDNAs—Colony lifts (Colony/Plaque Screen filters, Du Pont-New England Nuclear) were prehybridized in 1 M NaCl, 1% sodium dodecyl sulfate, 10% dextran sulfate, 5 × Denhardt's solution, 100 μg/ml DNA, 50% formamide at 42 °C, followed by hybridization with a 1.0-kb *Hin*dII fragment of pM-4 ^{32}P-labeled by multiprime labeling (Amersham Corp.). The pM-4 *Hin*dII fragment contains sequences coding for both the transmembrane and cytoplasmic domains of mouse syndecan (Saunders *et al.*, 1989). Filters were washed in 2 × SSC, 1% sodium dodecyl sulfate at 60 °C and autoradiographed on Kodak X-Omat or Fuji x-ray film. In the case of the primer-extended library, colony lifts were probed with the multiprime ^{32}P-labeled *Pst*I fragment of hsyn4 from the 5′ region (bases 341–709 in Fig. 2). The clone with the longest insert (hsynpr7) was analyzed further.

Subcloning and DNA Sequencing—Purified pUC19-hsyn4 and pUC19-hsynpr7 plasmids were isolated. Fragments released by restriction endonuclease digestions were size-fractionated by electrophoresis in Bio-Rad low temperature gelling agarose. After excision from the gel, fragments were subcloned to M13mp18 or M13mp19 for sequence analysis. DNA sequencing was performed by the dideoxy chain termination method (Sanger *et al.*, 1977) using a modified T7 DNA polymerase (Sequenase and Sequenase Version 2). Both the universal M13 primer and synthetic oligonucleotide primers were used. Areas with GC compression problems were also sequenced with dITP reactions.

Northern Blots—RNA was isolated as described above from NMuMg, MCF-7, and HBL-100 (American Type Culture Collection) cells and from human fetal skin. Fetal skin was grown in the presence of liquid nitrogen and transferred to GIT buffer. For Northern blot analysis, RNA samples were separated by 1% formaldehyde-agarose gel electrophoresis. The gel was blotted onto a GeneScreen hybridization membrane. Blots were hybridized with the multiprimed hsyn4 or pM-4 1.0-kb *Hin*dII fragment. Hybridization and washes were done as described above, except that the pM-4 labeled Northern blot was washed at 55 °C.

Sequence Analysis and Comparisons—Analysis of sequences was done using the sequence analysis software package of the University of Wisconsin Genetics Computer Group. The mutational difference matrix table of Dayhoff was used in protein comparison to show similar amino acids (Schwartz and Dayhoff, 1979). No significant homologies were found during sequence comparisons with the GenBank™, EMBL, and National Biomedical Research Foundation protein data bases.

RESULTS

cDNA Cloning of Human Syndecan—Our biochemical work with a human breast epithelial cell line (HBL-100) has indicated that these cells contain a cell-surface proteoglycan very similar to murine syndecan.[2] Therefore, we isolated mRNA from HBL-100 cells by oligo(dT)-cellulose chromatography and used this mRNA to construct cDNA by the RNase H method (Cubler and Hoffman, 1983). For the establishment of a cDNA library, cDNA >500 bp was cut out from the agarose gel and subsequently eluted from the gel by isotachophoresis (Öfverstedt *et al.*, 1984). cDNA was then ligated to the *Sma*I site of the pUC19 plasmid, and the recombinant cDNA library was transformed to the dH5α strain of *E. coli* cells. Plates were colony-lifted and hybridized with a multiprimed ^{32}P-labeled 1.0-kb *Hin*dII fragment of pM-4. pM-4 is a partial cDNA clone of murine syndecan (Saunders *et al.*, 1989), and its 1.0-kb *Hin*dII fragment contains sequences coding both the cytoplasmic and transmembrane domains of the syndecan core protein. Because this *Hin*dII fragment of pM-4 gave us the best signal in Northern blots for human RNA, we assumed that these domains may be the most conserved areas between species; and thus, their use could facilitate the isolation of correct clones for human syndecan (see "Northern Blot Analysis" in Fig. 4).

pM-4 positive clones were further isolated; and one of them, hsyn4, repeatedly gave a strong signal and was thus sequenced. hsyn4 is 2099 bp long and contains 3′-untranslated sequences and most of the coding region of human syndecan (Figs. 1 and 2). To get the full-length sequence for human syndecan, we constructed a new cDNA library by priming HBL-100-derived mRNA with a specific 20-mer for the 5′ area of the hsyn4 sequence (bases 719–738 in Fig. 2). Positive colonies were screened with a multiprimed ^{32}P-labeled *Pst*I-digested fragment from the 5′ area of hsyn4 (see Fig. 1). The longest of the positive clones (hsynpr7) contained the rest of the 5′ information for the human syndecan cDNA (Figs. 1 and 2).

Sequence of Human Syndecan—The human syndecan cDNA sequence derived from hsyn4 and hsynpr7 clones introduced a 2430-bp-long sequence, which had an open reading frame of 930 nucleotides, starting at position 206 with ATG (Fig. 2). The 1292-long 3′-untranslated region ended with a 27-residue-long poly(A) tail following the polyadenylation signal AATAA 20 bases upstream (Fig. 2). The human syndecan nucleotide sequence (Fig. 2) revealed 78% similarity with the mouse syndecan sequence (Saunders *et al.*, 1989) with several caps (>20). Most of the caps were located in the 3′-untranslated area. One of the caps in the 3′ end resulted in the formation of a deletion in the human sequence, which caused formation of an interesting 17-residue poly(T) stretch (poly(U) in mRNA). Poly(T) stretches (>10) have been found

[2] K. Elenius, M. Salmivirta, P. Inki, and M. Jalkanen, manuscript in preparation.

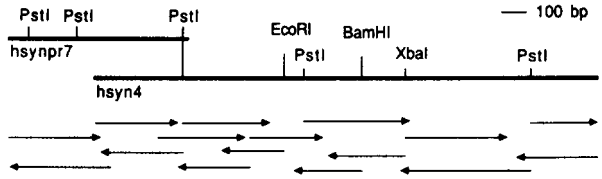

FIG. 1. **Cloning strategy of human syndecan cDNA sequence.** hsyn 4 includes the 3′-untranslated sequence and most of the coding region sequence. hsynpr7 was primed with a synthetic primer just before the *PstI* site of clone hsyn4. hsynpr7 contains the rest of the coding region sequence and the 5′-untranslated sequence. *Arrows* indicate sequencing direction and strategy.

```
1     ggagaggtgcgggccgaatccgagccgagcgagaggaatccggcagtagagagcggactc
      cagccggcggaccctgcagcccctcgcctgggacagcggcgcgctgggcaggcgcccaaga
      gagcatcgagcagcggaacccgcgaagccggcccgcagccgcgacccgcgcagcctgccg
      ctctcccgccgccggtccgggcagcatgaggcgcgcggcgctctggctctggctgcgc
                              M  R  R  A  A  L  W  L  W  L  C  A     12
      gctggcgctgagcctgcagctggccctgccgcaaattgtggctactaatttgccccctga
      L  A  L  S  L  Q  L  A  L  P  Q  I  V  A  T  N  L  P  P  E     32
301   agatcaagatggctctggggatgactctgacaacttctccggctcaggtgcaggtgctttt
      D  Q  D  G  S  G  D  D  S  D  N  F  S  G  S  G  A  G  A  L     52
      gcaagatatcaccttgtcacagcagacccccctccacttggaaggacacgcagctcctgac
      Q  D  I  T  L  S  Q  Q  T  P  S  T  W  K  D  T  Q  L  L  T     72
      ggctattcccacgtctccagaacccaccggcctggaggctacagctgcctccacctccac
      A  I  P  T  S  P  E  P  T  G  L  E  A  T  A  A  S  T  S  T     92
      cctgccggctggagaggggcccaaggagggagaggctgtagtcctgccagaagtggagcc
      L  P  A  G  E  G  P  K  E  G  E  A  V  V  L  P  E  V  E  P    112
      tggcctcaccgcccgggagcaggaggccacccccgcacccagggagaccacacagctccc
      G  L  T  A  R  E  Q  E  A  T  P  R  P  R  E  T  T  Q  L  P    132
601   gaccactcatcaggcctcaacgaccacagccaccacggcccaggagcccgccacctccca
      T  T  H  Q  A  S  T  T  T  A  T  T  A  Q  E  P  A  T  S  H    152
      cccccacagggacatgcagcctggccaccatgagacctcaacccctgcaggacccagcca
      P  H  R  D  M  Q  P  G  H  H  E  T  S  T  P  A  G  P  S  Q    172
      agctgaccttcacactccccacacagaggatggaggtccttctgccaccgagagggctgc
      A  D  L  H  T  P  H  T  E  D  G  G  P  S  A  T  E  R  A  A    192
      tgaggatggagcctccagtcagctcccagcagcagagggctctgggagcaggacttcac
      E  D  G  A  S  S  Q  L  P  A  A  E  G  S  G  E  Q  D  F  T    212
      ctttgaaacctcgggggagaatacggctgtagtggccgtggagcctgaccgccggaacca
      F  E  T  S  G  E  N  T  A  V  V  A  V  E  P  D  R  R  N  Q    232
901   gtccccagtggatcaggggccacgggggcctcacagggcctcctggacaggaaagaggt
      S  P  V  D  Q  G  A  T  G  A  S  Q  G  L  L  D  R  K  E  V    252
      gctgggaggggtcattgccggaggcctcgtggggctcatctttgctgtgtgcctggtggg
      L  G  G  V  I  A  G  G  L  V  G  L  I  F  A  V  C  L  V  G    272
      tttcatgctgtaccgcatgaagaagaaggacgaaggcagctactccttggaggagccgaa
      F  M  L  Y  R  M  K  K  K  D  E  G  S  Y  S  L  E  E  P  K    292
      acaagccaacggcggggcctaccagaagcccaccaaacaggaggaattctatgcctgacg
      Q  A  N  G  G  A  Y  Q  K  P  T  K  Q  E  E  F  Y  A  *       310
      cgggagccatgcgcccctccgccctgccactcactaggcccccacttgcctcttccttg
1201  aagaactgcaggccctggcctccctgccaccaggccacctccccagcattccagccctt
      ctggtcgctcctgcccacggagtcgtgggtgtgctgggagctccactctgcttctctgac
      ttctgcctggagacttagggcaccaggggtttctcgcataggacctttccaccacagcca
      gcacctggcatcgcaccattctgactcggtttctccaaactgaagcagcctctccccagg
      tccagctctggaggggagggggatccgactgctttggacctaaatggcctcatgtggctg
1501  gaagatctgcgggtggggcttggggctcacacacctgtagcacttactggtaggaccaag
      catcttgggggggtggccgctgagtggcagggacaggagtcactttgtttcgtggggagg
      tctaatctagatatcgacttgtttttgcacatgtttcctctagttctttgttcatagccc
      agtagaccttgttacttctgaggtaagttaagtaagttgattcggtatccccccatcttg
      cttccctaatctatggtcgggagacagcatcagggttaagaagacttttttttttttttt
1801  ttaaactaggagaaccaaatctggaagccaaaatgtaggcttagtttgtgtgttgtctct
      tgagtttgtcgctcatgtgtgcaacagggtatggactatctgtctggtggccccgtttct
      ggtggtctgttggcaggctggccagtccaggctgccgtggggccgccgcctctttcaagc
      agtcgtgcctgtgtccatgcgctcagggccatgctgaggcctgggccgctgccacgttgg
      agaagcccgtgtgagaagtgaatgctgggactcagccttcagacagagaggactgtaggg
2101  agggcggcaggggcctggagatcctcctgcagaccacncccgtcctgcctgtgcgccgtc
      tccaggggctgcttcctcctggaaattgacgaggggtgtcttgggcagagctggctctga
      gcgcctccatccaaggccaggttctccgttagctcctgtggcccccaccctgggccctggg
      ctggaatcaggaatatttttccaaagagtgatagtcttttgcttttggcaaaactctactt
      aatccaatgggttttttccctgtacagtagatttccaaatgtaataaactttaatatataa
2401  gtaaaaaaaaaaaaaaaaaaaaaaaaaaaa
```

FIG. 2. **Sequence of human syndecan.** The sequence contains 2430 bp. It has an open reading frame of 930 bp coding a 310-amino acid-long core protein. *Circled* serine residues demonstrate possible glycosaminoglycan side chain attachment sites. *Boldface underlining* indicates the hydrophobic transmembrane domain. *Lightface underlining* shows the aataa polyadenylation signal 14 base pairs before the 28-base poly(A) tail.

```
  1 MRRAALWLWLCALALSLQLALPQIVATNLPPEDQDGSGDDSDNFSGSGAG 50
    öööööööööööööööö.öö ööööööö.ö:ööööööööööööööööööö.ö
  1 MRRAALWLWLCALALRLQPALPQIVAVNVPPEDQDGSGDDSDNFSGSGTG 50

 51 ALQDITLSQQTPSTWKDTQLLTAIPTSPEPTGLEATAASTSTLPAGEGPK 100
    öö.ö ööö.ööööööö. öööö.öö.ööö: :...ö öö.ööööö ö.
 51 ALPD.TLSRQTPSTWKDVWLLTATPTAPEPTSSNTETAFTSVLPAGEKPE 99

101 EGEAVVLPEVEPGLTAR..EQEATPRPRETTQLPTTHQASTTTATTAQEP 148
    ööö:ö: .ö.ööö:ööö ö.ö.ö.ööööö.ööö.ö:.ööö. .ööööö.:
100 EGEPVLHVEAEPGFTARDKEKEVTTRPRETVQLPITQRASTVRVTTAQAA 149

149 ATSHPHRDMQPGHHETSTPAGPSQADLHTPHTEDGGPSATERAAEDGASS 198
    .ööööö :ööööö ööööö.ö.:ö:ö:ö :.ö:.ö:ööö.ö... ..ööö...
150 VTSHPHGGMQPGLHETSAPTAPGQPDHQPPRVEGGGTSVIKEVVEDGTAN 199

199 QLPAAEGSGEQDFTFETSGENTAVVAVEPDRRNQSPVDQGATGASQGLLD 248
    ööööö:ööööööööööööööööööö.ööööö: ööö.ööö:öööööö:ööö
200 QLPAGEGSGEQDFTFETSGENTAVAAVEPGLRNQPPVDEGATGASQSLLD 249

249 RKEVLGGVIAGGLVGLIFAVCLVGFMLYRMKKKDEGSYSLEEPKQANGGA 298
    öööööööööööööööööööööööö:ööööööööööööööööööööööööö
250 RKEVLGGVIAGGLVGLIFAVCLVAFMLYRMKKKDEGSYSLEEPKQANGGA 299

299 YQKPTKQEEFYA 310
    öööööööööööö
300 YQKPTKQEEFYA 311
```

FIG. 3. **Homology between human (*upper*) and mouse (*lower*) syndecan protein sequences.** Most of the identical parts between human and mouse syndecans are the signal peptide, areas around glycosaminoglycan attachment sites, and transmembrane and cytoplasmic domains. Ö, identical amino acids; : and ., similar kinds of amino acids using the comparison table of Dayhoff (Schwartz and Dayhoff, 1979). Possible glycosaminoglycan attachment sites are indicated with *arrowheads*. *Underlining* shows transmembrane domains.

in 3′-untranslated areas of mRNA molecules for several proteins, and it has been proposed that these poly(T) stretches are capable of making loops with the poly(A) tail of mRNA that could participate in the stabilization of mRNA (Hennessy *et al.*, 1989).

The amino acid sequence derived from the coding region of human syndecan cDNA indicated a 310-amino acid-long core protein, 77% identical to the mouse syndecan protein sequence (Fig. 3). The human syndecan core protein sequence, like that of mouse syndecan, had two hydrophobic stretches. The first one was found in the NH_2 terminus, representing the signal peptide sequence. The signal peptidase cleavage site has been proposed to follow residue 17 (Saunders *et al.*, 1989) because cleavage at this site would generate an NH_2-terminal glutamine that could form a blocked NH_2 terminus observed for mouse syndecan (Weitzhandler *et al.*, 1988). The other hydrophobic sequence represented a typical transmembrane domain (the 25 residues *underlined* in Fig. 2), which subsequently divided the core protein into three individual domains: the ectodomain, transmembrane domain, and cytoplasmic domain as described earlier for mouse syndecan (Saunders *et al.*, 1989).

The ectodomain contained five possible glycosaminoglycan attachment sites. These were at positions 37, 45, 47, 206, and 216 (*circled* serine residues in Fig. 2 and *arrowheads* in Fig. 3) and were identical to those of mouse syndecan (see Fig. 3). Two of these sites (at positions 37 and 206) matched the consensus sequence of (E/D)GSG(E/D), which is proposed to be one of the possible glycosaminoglycan attachment sites by Zimmermann and Ruoslahti (1989). Position 216 was similar but had the second glycine changed to threonine. Positions 45 and 47 matched the consensus sequence SG*X*G proposed by Bourdon *et al.* (1987) as the attachment site of chondroitin sulfate. Conservation of peptide sequences around serines proposes a fundamental role of these areas for the successful initiation of glycosaminoglycan synthesis and chain elongation by transferases. It was also interesting to see that other parts of the ectodomain were less conserved than the areas around the GAG attachment sites, suggesting that changes in the ectodomain core protein can occur during evolution but that they are directed to functionally less important areas of syndecan. There was only one possible *N*-glycosylation site in human syndecan 2 amino acids before the serine residue at position 45. This asparagine may function as a regulator for the attachment of glycosaminoglycans to serines 45 and 47 (Saunders *et al.*, 1989). The variation of syndecan glycosylation may be very important for the outcome of syndecan function in various tissues; and therefore, this site has been conserved in both species. Recently, Sanderson and Bernfield (1988) have shown that the length, number, and type of GAG can vary among the syndecans isolated from different tissues.

Thus, it may be that the glycosylation of the asparagine residue could inhibit the function of xylosyltransferase, an initiator of GAG synthesis. Our work[3] with mouse syndecan has indicated that this molecule could exist in the form of plain heparan sulfate when isolated from the condensating dental mesenchyme, where it has been shown to be transiently expressed (Thesleff *et al.*, 1988; Vainio *et al.*, 1989a).

Mouse syndecan is shed by cleavage of its ectodomain from the membrane-associated domain, and the shed molecule is indistinguishable from the syndecan ectodomain released from the cell surface by mild trypsin treatment (Jalkanen *et al.*, 1987; Weitzhandler *et al.*, 1988). Both mouse and human syndecans contain the dibasic sequence Arg-Lys in the ectodomain adjacent to the transmembrane domain. Our biochemical evidence[2] with the HBL-100 cell line also indicates that these cells can release heparan sulfate-rich proteoglycan from cell surfaces into the medium, similar to NMuMG cells. This suggests that the RK site is functional also in human cells. The ectodomain of human syndecan also contained another dibasic RR sequence just 18 residues apart from the first one. Proteolytic cleavage at this site would also release the ectodomain with all GAG chains intact (Figs. 2 and 3).

Human syndecan had a 25-amino acid-long hydrophobic transmembrane domain which was identical to the mouse syndecan transmembrane domain, except for the change of glycine 272 to alanine. The cytoplasmic domain of human syndecan was 34 amino acids long and was fully conserved between human and mouse. In both species, these domains contained 4 tyrosine residues, 1 at the end of the transmembrane domain and 3 others in the cytoplasmic domain in identical locations (Fig. 3). Evolutionally, extremely conserved transmembrane and cytoplasmic domains may indicate a very crucial role for these domains in the so far largely unknown function of syndecan as a matrix receptor.

Northern Blot Analysis—Equal amounts (10 μg) of total RNA from normal murine (NMuMg) and human (HBL-100) mammary epithelial cells, mammary carcinoma cells (MCF-7), and human fetal skin were submitted to Northern blot analysis under high stringency conditions with multiprimed ^{32}P-labeled hsyn4. This analysis revealed two mRNA species in all samples (Fig. 4). Their sizes were slightly smaller than the corresponding mouse species (2.6 and 3.4 kb). Also, the relative abundances of human syndecan mRNAs were different (about 1:1 in all cell lines); but in skin, the larger band was more abundant. This has been reported to be the opposite for mouse skin (Saunders *et al.*, 1989). The signal for NMuMG cell-derived RNA with hsyn4 as a probe was much weaker than that for human cell line-derived RNAs because identical RNA loadings showed only the 2.6-kb mRNA (Fig. 4). However, with longer exposures, also the larger 3.4-kb mRNA was detectable (data not shown).

Species cross-reactivity was tested further by probing RNA samples from both species with mouse syndecan. For this, 5 μg of RNA from NMuMG cells and 15 μg from HBL-100 cells were size-fractionated, blotted onto a GeneScreen membrane, and hybridized with a 1.0-kb *Hin*dII fragment of pM-4. This fragment contains sequences coding for the transmembrane and cytoplasmic domains, the most homologous parts between mouse and human syndecans (Fig. 3). The filters were washed under low stringency. The results were very similar; the signal from HBL-100 RNA was less intense than that from NMuMg RNA (Fig. 4), although the loading of HBL-100 RNA was three times higher. When Northern blots were done with whole pM-4, only a very weak signal from human cells was visible because relative nonhomologous parts of pM-4 reduced the signal with HBL-100 RNA even more (data not shown). Sizes of mRNA species in HBL-100 and NMuMG cells in Northern blot analysis hybridized with the pM-4 1.0-kb *Hin*dII fragment were the same as sizes with hsyn4. The reason why syndecan has two different mRNA sizes is not

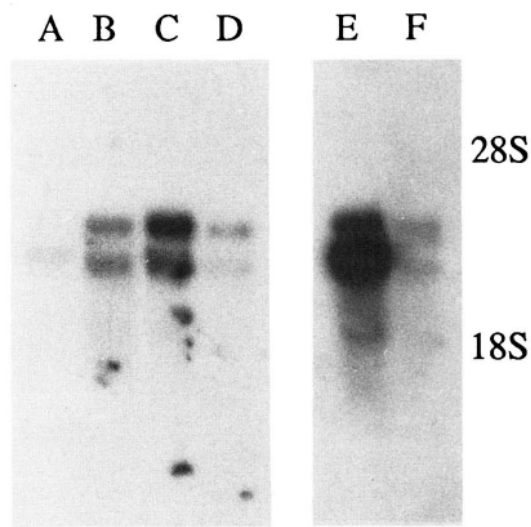

FIG. 4. **Northern blot analysis of mouse and human RNAs with hsyn4 and pM-4.** Total RNA (10 μg) from NMuMG cells (*lane A*), MCF-7 cells (*lane B*), HBL-100 cells (*lane C*), and human fetal skin (*lane D*) were size-fractionated and blotted onto a GeneScreen membrane. The membrane was hybridized with multiprimed labeled hsyn4 and washed under high stringency conditions. In *lane E*, only 5 μg of RNA from NMuMG cells and in *lane F*, 15 μg of RNA from HBL-100 cells were hybridized with the 0-kb *Hin*dII fragment of pM-4 (mouse syndecan partial cDNA clone). The filter was washed under low stringency conditions. hsyn 4 shows the intense signal from human cells, but not from mouse cells; pM-4 shows the reverse. Both probes gave identical mRNA species in human and mouse. In human, sizes of mRNA species are a little smaller than those in mouse (about 2.6 and 3.4 kb).

```
252 VLGGVIAGGLVGLIFAVCLVGFMLYRMKKKDEGSYSLEEPKQANGGAYQK 301
    öö::ööööö::ö::öö: ö: :::öööö:ööööööö.ö:ö.ö :.::ööööö
341 VLAAVIAGGVIGFLFAIFLILLLVYRMRKKDEGSYDLGERK.PSSAAYQK 389

302 PTKQEEFYA 310
    :... öööö
390 APTK.EFYA 397
```

FIG. 5. **Homology between protein sequences of transmembrane and cytoplasmic domains in human syndecan (*upper*) and human lung fibroblast heparan sulfate core protein 48K5 (*lower*).** Similarity between these domains is 86%, and identity is 56%. Conserved tyrosines are indicated with *arrowheads*, and the transmembrane domain is *underlined*. See the legend to Fig. 3 for definition of the symbols.

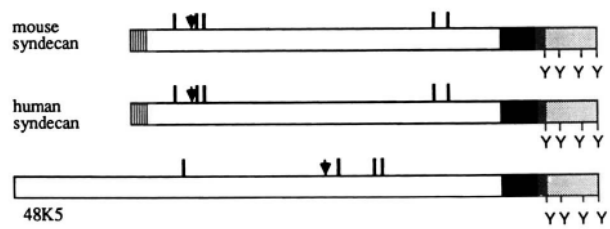

FIG. 6. **Domain structure of members including integral membrane proteoglycan family.** The signal peptide is *striped*, the ectodomain is *open*, the transmembrane domain is *cross-hatched*, and the cytoplasmic domain is *shaded*. Tyrosines in the transmembrane and cytoplasmic domains are shown (*Y*). Possible glycosaminoglycan attachment sites are indicated with *bars*, and possible *N*-glycosylation sites with *arrows*.

[3] M. Salmivirta, U. Hofer, S. Vainio, K. Elenius, R. Chiquet-Ehrismann, I. Thesleff, and M. Jalkanen, manuscript in preparation.

known. Northern blot analysis with the 48K5 probe has also revealed two classes of mRNAs (2.9 and 4.2 kb); and in this case, the sequence of the 3′ prime end suggests that these mRNAs are generated by the use of alternative polyadenylation signals (Marynen *et al.*, 1989).

DISCUSSION

Although cell-surface proteoglycans are found on all adhesive cells, only a few of them have been characterized today. Some of these are so-called "part-time proteoglycans," which appear both with and without GAG chains (see Ruoslahti, 1989). None of these, however, is related in structure to syndecan, a heparan/chondroitin sulfate-rich proteoglycan (Saunders *et al.*, 1989) found mainly on the surface of epithelial cells in mature tissues (Hayashi *et al.*, 1987). Instead, a cell-surface proteoglycan (48K5) from lung fibroblasts (Marynen *et al.*, 1989) shows intriguing similarity to syndecan (see Figs. 5 and 6). We present in this paper the full-length cDNA for human syndecan, which allowed us to compare this sequence to those of murine syndecan and the 48K5 proteoglycan from human lung fibroblasts. This comparison revealed the most conserved areas of syndecans in both species, suggesting their functional importance, but also indicated that the 48K5 proteoglycan has similar transmembrane and cytoplasmic domains as syndecan, suggesting that these molecules belong to the same novel gene family of integral membrane proteoglycans, the syndecans.

Conservation of Syndecan Structure—Comparison of human and mouse syndecans revealed several conserved areas. The most conserved were the cytoplasmic and transmembrane domains. In a total of 59 amino acid residues, there was only one nonsignificant change (alanine to glycine; see Fig. 3). Similar conservation has also been found for the cytoplasmic domain of the integrin β-subunit (Marcantonio and Hynes, 1988). One dominating feature in both syndecans was the identical location of 4 tyrosine residues, 1 at the end of the transmembrane domain and 3 others in the cytoplasmic domain (see Figs. 3 and 6). Although no information is available for the function of these residues, they could participate in the transportation of these molecules into the correct location within cells, as described for the mannose 6-phosphate receptor (Lobel *et al.*, 1989); or the covalent modification of tyrosine residues could also regulate the adhesion of these molecules to cytoplasmic components, as proposed for integrin β-subunits (Freed *et al.*, 1989).

The ectodomains of mouse and human syndecans showed less homology than the transmembrane and cytoplasmic domains (see Fig. 3). Both ectodomains contained, however, well-conserved areas, like GAG attachment sites and the single *N*-glycosylation site as well as the dibasic sequence site adjacent to the extracellular face of the transmembrane domain, potentially serving as the proteinase-sensitive site involved in release of the ectodomain from the cell surface (Jalkanen *et al.*, 1987).

Syndecans: A Novel Gene Family of Integral Membrane Proteoglycans—A high degree of similarity was also observed between the transmembrane and cytoplasmic domains of syndecan and the cell-surface proteoglycans from human lung fibroblasts (see Figs. 5 and 6). The striking similarity was the identical location of 4 tyrosine residues in both molecules (Fig. 5). The ectodomain of fibroblast syndecan (48K5) is completely nonhomologous (Marynen *et al.*, 1989), but is still structurally related. For example, it also contains dibasic sequences next to the transmembrane domain (one KR sequence and one RK sequence 2 and 24 residues from the start of the transmembrane domain, respectively) which may be proteinase-susceptible and explain at least part of that heterogeneity among the core proteins from lung fibroblasts (Lories *et al.*, 1989). One of the GAG attachment sites in 48K5 belongs to the consensus sequence SG*X*G, but two of them are only partial; and none belong to the consensus sequence (E/D)GSG(E/D) (Marynen *et al.*, 1989). This is surprising because the 48K5 proteoglycan is supposedly a heparan sulfate proteoglycan (Lories *et al.*, 1989; Marynen *et al.*, 1989). Therefore, it could be that other parts on the core protein can also influence what kind of glycosylation each core receives. Only a partial sequence for this proteoglycan has been described so far, and the complete sequence of this molecule could reveal other GAG attachment sites more closely related to the GAG attachment sites of syndecan. 48K5 has similar transmembrane and cytoplasmic domains as syndecan, and it could be that these structures favor transportation of both core proteins to cellular compartments that results in the addition of heparan sulfate to the core. The dissimilarity of ectodomains, however, may indicate that syndecans represent a collection of receptors that differ in their interactions. Heterogeneity has been shown to be evident for mouse epithelial syndecan because it can be isolated from different tissues with different glycosylation patterns; in simple epithelia, syndecan has more and larger heparan sulfate and chondroitin sulfate chains than the proteoglycan from stratified epithelia (Sanderson and Bernfield, 1988). In the 48K5 molecule, the ectodomain is completely different, but the conserved transmembrane and cytoplasmic structures imply that this molecule can have specific ligand-binding properties but that the conserved domains serve common roles from one molecule to another that are essential to syndecan function.

Distribution of Syndecans—In adult tissues, syndecan is mainly limited to epithelial tissues (Hayashi *et al.*, 1987). However, embryonic mesenchyme also expresses syndecan (Thesleff *et al.*, 1988; Vainio *et al.*, 1989a, 1989b), which is also evident at the mRNA level.[3] The distribution of the 48K5 proteoglycan is not known; and therefore, it is difficult to limit it only to fibroblasts. Thus, it is too early to talk of syndecans of epithelial or fibroblastic origin before more information about their expression has been obtained. Furthermore, syndecan is not restricted to epithelial and mesenchymal cells. Recently, it was shown to be expressed by pre-B and plasma cells, but not by B cells, indicating that it may play an important role in the development of B cells, which evidently interact with matrices during their development (Sanderson *et al.*, 1989). Mouse syndecan cDNA detects a 4.5-kb mRNA from brain tissues (Saunders *et al.*, 1989), and we have also obtained similar results with human brain tissues.[4] This molecule has not been cDNA-cloned yet, but it may be closely related to syndecan because Northern blot hybridizations were carried out under high stringency conditions. Syndecan molecules have not been searched yet, *e.g.* from endothelial or smooth muscle cells, but these cells may express syndecan-like molecules which are related to the specialized morphology and developmental history of these cells.

In conclusion, we have cDNA-cloned the full-length cDNA for the cell-surface proteoglycan from the HBL-100 cell line that is a human equivalent to mouse syndecan. We would like to propose that proteoglycans containing similar transmembrane and cytoplasmic domains should be included as members of a novel gene family of integral membrane proteoglycans, the syndecans. The members of this family may represent a collection of receptors that differ in their interactions due to variation of the ectodomain in length, size, and charge.

[4] M. Mali, H. Hirvonen, and M. Jalkanen, unpublished data.

The conserved similarity of their transmembrane and cytoplasmic domains, on the other hand, may serve common roles from one molecule to another that are essential to syndecan function during the development of organs or in the stabilization of cell morphology.

Acknowledgments—We wish to thank Drs. Merton Bernfield and Scott Saunders for discussion, Drs. Kati Elima, Jyrki Mäkelä, Arto Määttä, and Tapani Vihinen for technical advice, and Taina Kalevo and Tuula Oivanen for excellent technical assistance.

REFERENCES

Argraves, W. S., Suzuki, S., Arai, H., Thompson, K., Pierschbacher, M. D., and Ruoslahti, E. (1987) *J. Cell Biol.* **105,** 1183–1190

Argraves, W. S., Dickerson, K., Burgress, W. H., and Ruoslahti, E. (1989) *Cell* **58,** 623–629

Bourdon, M. A., Krusius, T., Campbell, S., Schwartz, N. B., and Ruoslahti, E. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84,** 3194–3198

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299

Cubler, U., and Hoffman, B. J. (1983) *Gene (Amst.)* **25,** 263–269

D'Souza, S. E., Ginsberg, M. H., Burke, T. A., Lam, S. C.-T., and Plow, E. F. (1988) *Science* **242,** 91–93

Freed, E., Gailit, J., van der Geer, P., Ruoslahti, E., and Hunter, T. (1989) *EMBO J.* **8,** 2955–2965

Hanahan, D. (1983) *J. Mol. Biol.* **166,** 557–580

Hayashi, K., Hayashi, M., Jalkanen, M., Firestone, J. H., Trelstad, R., and Bernfield, M. (1987) *J. Histochem. Cytochem.* **35,** 1079–1088

Hennessy, S. W., Frazier, B. A., Kim, D. D., Deckwerth, T. L., Baumgartel, D. M., Rotwein, P., and Frazier, W. A. (1989) *J. Cell Biol.* **108,** 729–736

Hirst, R., Horwitz, A., Buck, C., and Rohrschneider, L. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83,** 6470–6474

Hynes, R. O. (1987) *Cell* **48,** 549–554

Jalkanen, M., Nguyen, H., Rapraeger, A., Kurn, N., and Bernfield, M. (1985) *J. Cell Biol.* **101,** 976–984

Jalkanen, M., Rapraeger, A., Saunders, S., and Bernfield, M. (1987) *J. Cell Biol.* **105,** 3087–3096

Koda, J. E., Rapraeger, A., and Bernfield, M. (1985) *J. Biol. Chem.* **260,** 8157–8162

Lobel, P., Fujimoto, K., Ye, R. D., Griffiths, G., and Kornfeld, S. (1989) *Cell* **57,** 787–796

Lories, V., Cassiman, J.-J., Van den Berghe, H., and David, G. (1989) *J. Biol. Chem.* **264,** 7009–7016

Marcantonio, E. E., and Hynes, R. O. (1988) *J. Cell Biol.* **106,** 1765–1772

Marynen, P., Zhang, J., Cassiman, J., Van den Berghe, H., and David, C. (1989) *J. Biol. Chem.* **264,** 7017–7024

Öfverstedt, L.-G., Hammarström, K., Balgobin, N., Hjertén, S., Petterson, U., and Chattopadhyaya, J. (1984) *Biochim. Biophys. Acta* **782,** 120–126

Pytela, R., Pierschbacher, M. D., and Ruoslahti, E. (1985) *Cell* **40,** 191–198

Rapraeger, A., Jalkanen, M., Endo, E., Koda, J., and Bernfield, M. (1985) *J. Biol. Chem.* **260,** 11046–11052

Rapraeger, A., Jalkanen, M., and Bernfield, M. (1986) *J. Cell Biol.* **103,** 2683–2696

Ruoslahti, E. (1989) *J. Biol. Chem.* **264,** 13369–13372, and references therein

Ruoslahti, E., and Pierschbacher, M. D. (1987) *Science* **238,** 491–497

Sanderson, R. D., Lalor, P., and Bernfield, M. (1989) *Cell Regul.* **1,** 27–35

Sanderson, R. D., and Bernfield, M. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 9562–9566

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467

Saunders, S., and Bernfield, M. (1988) *J. Cell Biol.* **106,** 423–430

Saunders, S., Jalkanen, M., O'Farrell, S., and Bernfield, M. (1989) *J. Cell Biol.* **108,** 1547–1556

Schwartz, R. M., and Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, pp. 353–358, National Biomedical Research Foundation, Washington, D. C.

Sun, X., Mosher, D. F., and Rapraeger, A. (1989) *J. Biol. Chem.* **264,** 2885–2889

Thesleff, I., Jalkanen, M., Vainio, S., and Bernfield, M. (1988) *Dev. Biol.* **129,** 565–572

Vainio, S., Jalkanen, M., and Thesleff, I. (1989a) *J. Cell Biol.* **108,** 1945–1954

Vainio, S., Lehtonen, E., Jalkanen, M., Bernfield, M., and Saxén, L. (1989b) *Dev. Biol.* **134,** 382–391

Weitzhandler, M., Streeter, H. B., Henzel, W. J., and Bernfield, M. (1988) *J. Biol. Chem.* **263,** 6949–6952

Zimmermann, D. R., and Ruoslahti, E. (1989) *EMBO J.* **8,** 2975–2981