Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 10404–10408, November 1993
Biochemistry

Structural characterization of the complete human perlecan gene and its promoter

(heparan sulfate/proteoglycan/basement membrane)

ISABELLE R. COHEN, SUSANNE GRÄSSEL, ALAN D. MURDOCH, AND RENATO V. IOZZO*

Department of Pathology and Cell Biology and the Jefferson Cancer Institute, Thomas Jefferson University, Room 249, Jefferson Alumni Hall, Philadelphia, PA 19107

Communicated by Elizabeth D. Hay, August 2, 1993 (received for review June 25, 1993)

ABSTRACT **The complete intron–exon organization of the gene encoding human perlecan (*HSPG2*), the major heparan sulfate proteoglycan of basement membranes, has been elucidated, and specific exons have been assigned to coding sequences for the modular domains of the protein core. The gene was composed of 94 exons, spanning >120 kbp of genomic DNA. The exon arrangement was analyzed vis-à-vis the modular structure of the perlecan, which harbors protein domains homologous to the low density lipoprotein receptor, laminin, epidermal growth factor, and neural cell adhesion molecule. The exon size and the intron phases were highly conserved when compared to the corresponding domains of the homologous genes, suggesting that most of this modular proteoglycan has evolved from a common ancestor by gene duplication or exon shuffling. The 5′ flanking region revealed a structural organization characteristic of housekeeping and growth control-related genes. It lacked canonical TATA or CAAT boxes, but it contained several GC boxes with binding sites for the transcription factors SP1 and ETF. Consistent with the lack of a TATA element, the perlecan gene contained multiple transcription initiation sites distributed over 80 bp of genomic DNA. These results offer insights into the evolution of this chimeric molecule and provide the molecular basis for understanding the transcriptional control of this important gene.**

In the past few years proteoglycans have assumed a pivotal role in diverse areas of human biology not only because of their physicochemical attributes but also because of their involvement in regulating cellular growth and differentiation (1, 2). A key player is perlecan (*HSPG2*), the major heparan sulfate proteoglycan of basement membranes and extracellular matrices (2–7). Complete cDNA cloning of the human species (6, 7) predicts a protein core of ≈467 kDa excluding any posttranslational modification, thus making perlecan one of the largest gene products of the human body. It is now apparent that the heparan sulfate proteoglycan originally isolated from the Engelbreth-Holm–Swarm (EHS) tumor (8) is identical to that found in the pericellular matrices of human colon carcinoma cells (9, 10), human lung fibroblasts (11), bovine endothelial cells (12), and mouse mammary epithelial (13) cells. The protein core of perlecan has undoubtedly descended from the use of protein modules previously identified in other extracellular matrix and ligand molecules. It comprises five distinct domains with only the first domain, the heparan sulfate-binding region, unique to perlecan (6). The other four domains exhibit homology to the low density lipoprotein (LDL) receptor, the N-terminal region of laminin A and B short arms, the neural cell adhesion molecule (N-CAM), and the globular C terminus of the laminin A chain, respectively (5–7). Because of its complex molecular organization, strategic topology, and widespread distribution (4), we predict that perlecan plays a crucial role not only in participating in the orderly assembly of extracellular matrices but also in interfering with the binding/delivery of nutrients and growth factors to target cells.

To gain insights into the regulation of perlecan gene expression, we elucidated the complete genomic organization of the human perlecan gene including the complete intron/exon boundaries and the 5′ flanking region.† The results revealed a complex gene comprising 94 exons and spanning >120 kbp of DNA. Both the exon sizes and phases were remarkably conserved in the various domains vis-à-vis the corresponding domains of the homologous genes. The putative promoter region was located within a typical CpG island, lacked a canonical TATA box, and contained several cis-acting elements characteristic of housekeeping and growth-control related genes. Primer extension and S1 nuclease protection assays revealed multiple transcription initiation sites, suggesting that control of the perlecan gene expression is complex.

A preliminary account of this work has appeared in abstract form (14).

MATERIALS AND METHODS

Materials. All reagents were of molecular biology grade. Radionucleotides [α-^{32}P]dCTP and [γ-^{32}P]ATP (≈3000 Ci/mmol; 1 Ci = 37 GBq) and deoxyadenosine 5′ [α-[^{35}S]thio]triphosphate (≈1000 Ci/mmol) were obtained from Amersham.

Isolation and Characterization of Genomic Clones. To isolate the entire human perlecan gene, we screened seven different genomic libraries including two *Mbo* I-generated cosmid libraries and three *Mbo* I-generated phage libraries in λFIX, λFIXII, and λDASH vectors (Clontech), respectively. In addition, because human perlecan is located on chromosome 1 (15), we screened two chromosome 1-specific Charon 21A libraries prepared from complete DNA digestion with either *Eco*RI (ATCC 57738) or *Hin*dIII (ATCC 57754). Probes were generated by PCR or restriction enzyme digestion of the previously isolated cDNAs (6) and were labeled by random priming (16). At least 10^6 recombinant clones (10^4 for the chromosome 1-specific library) were screened. Genomic clones positive on quaternary plaque screening were analyzed by Southern blotting and the appropriate fragments were subcloned into pBluescript (Stratagene) (17). DNA was sequenced by a modified dideoxynucleotide chain-termination method (18) or by an automated sequencing system (Applied Biosystems) using primers based on either the T3 or T7 polylinker sequences of pBluescript or synthetic oligonucleotides. The G+C-rich region of the 5′ flanking region

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.

Abbreviations: LDL, low density lipoprotein; N-CAM, neural cell adhesion molecule.

*To whom reprint requests should be addressed.

†The sequence reported in this paper has been deposited in the GenBank data base (accession no. L22078).

proved very difficult to sequence and was resolved by generating multiple deletions and using several primers or sequencing reactions at higher temperature. Computer analyses were performed using the GCG or PC/GENE package programs as described (17).

S1 Nuclease Protection and Primer-Extension Assays. For S1 nuclease protection, we used ≈50 μg of total RNA isolated from human colon carcinoma cells (15) and a 5′-labeled *Nco* I fragment containing part of exon 1 and 2.5 kb of 5′ flanking region. The RNA was hybridized to 5×10^6 cpm of the 5′-labeled fragment for 3 hr at 37°C in 40 mM Hepes, pH 6.5/0.4 M NaCl/1 mM EDTA/80% formamide. After hybridization, the samples were incubated with S1 nuclease (500 units/ml) at 37°C for 0–30 min. The samples were precipitated with ethanol, resuspended in formamide-containing buffer, and analyzed on a 6% polyacrylamide/urea sequencing gel. For primer extension, ≈3 μg of poly(A)$^+$ RNA from the same cells was annealed for 1 hr at 58°C with a 5′-end-labeled 25-mer oligonucleotide complementary to the sequence starting at position 168 of the cDNA (6). The samples were incubated with reverse transcriptase (12.5 units; 37°C for 1 hr), purified by column chromatography, and analyzed on a 6% sequencing gel as described above.

RESULTS AND DISCUSSION

Intron–Exon Organization of the Human Perlecan Gene. The human perlecan gene comprised 94 exons and spanned at least 120 kbp of genomic DNA, a conservative estimate since several genomic clones did not overlap with each other (Fig. 1). The exons of perlecan varied in size from 45 bp (exon 3) to 1.2 kbp (exon 94), with a mean length of 150 bp. All the exon–intron junctions followed the GT/AG rule, except for exons 15 and 27, which contained GA as 5′ splice donor (Fig. 2). Detailed analysis of the gene structure revealed a remark-

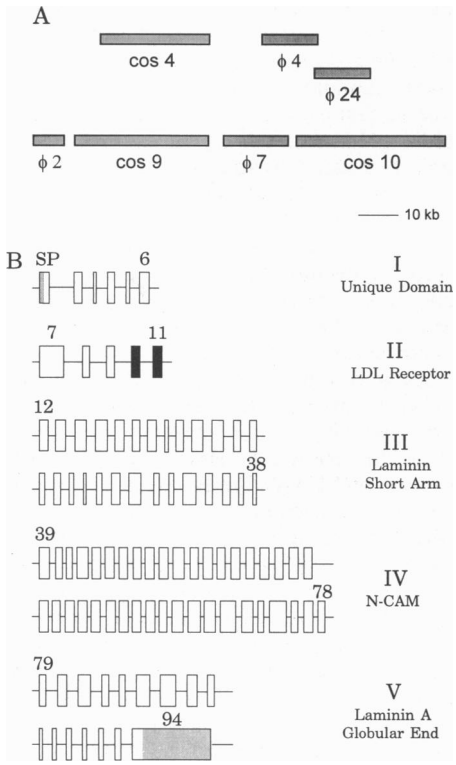

FIG. 1. Structure of the human perlecan (*HSPG2*) gene. (*A*) Schematic representation of the human perlecan genomic clones derived from either cosmid (Cos) or λ phage (ϕ) libraries. (*B*) Each vertical rectangle corresponds to one exon in scale with each other. Introns and flanking regions (solid lines) are not in scale (cf. Fig. 2 for additional details of intron size). The 94 exons are arranged in accordance with their specific domains as depicted on the right. The first exon contains the signal peptide (SP). The 5′ and 3′ untranslated regions are represented by shaded rectangles. Exons 10 and 11, encoding domain IIa, are represented by solid rectangles.

Exon No.	Size	5′ Splice Donor	Intron Length	3′ Splice Acceptor	Codon Phase	Amino Acid	
1	143	CTGGCG gtgag	>9000	tacag GTGACC	0	21-Ala/Val	SP
2	136	CAGGAG gtgag	208	tgtag ACGACC	I	67-Asp	
3	45	AGATGG gtaag	>2000	tccag TTTATT	I	82-Val	
4	110	GACACG gtgag	97	cccag CTGGAG	0	118-Thr/Leu	I
5	59	CATCAA gtgag	>283	cccag GGAGCT	II	138-Lys	
6	161	GCACAG gtgag	>549	tgcag TGCCCC	I	192-Val	
7	384	ACTGTG gtgcg	>950	aacag GCCCCC	I	320-Gly	
8	120	ACTGCC gtgag	>1000	cccag CCACCA	I	360-Pro	II
9	132	GCTGCA gtgag	162	ctcag TGACCC	I	404-Met	
10	145	TCCCAG gtgcg	>630	accag GGTGAC	II	452-Arg	IIa
11	152	AACGAG gtcac	93	agcag GCCCCT	I	503-Gly	
12	147	TCAAGG gtgcg	>6000	cacag GTGTGA	I	552-Gly	
13	164	AACAAG gtgag	>1000	cctag GTGGAC	0	606-Lys/Val	
14	180	TCTGAG gtggt	93	cccag GAGCAC	0	666-Glu/Glu	
15	197	GTGCAG gagcc	108	cacag ATGCCC	II	732-Arg	
16	148	TGCCTG gtgag	>273	tgcag AATTGC	0	781-Leu/Asn	
17	128	CCGCAG gtcag	272	cgcag ATTCTC	II	824-Arg	
18	146	CCGTCA gtagt	>413	tccag ACCAGG	I	873-Asn	
19	68	TGTAAG gtacg	>363	tgcag AACAAT	0	895-Lys/Asn	
20	141	GCCCAG gtacc	104	cccag TTGCAT	0	942-Glu/Leu	
21	177	GACAAG gtggg	>500	cccag GTGACC	0	1001-Lys/Val	
22	180	CGGGAG gtgag	>128	ggcag CAAGCA	0	1061-Glu/Gln	
23	119	GAGCAG gtgtc	>332	cccag GGTCTC	II	1101-Arg	
24	112	TGCCAG gtgag	83	tccag GACTGT	0	1138-Gln/Asp	III
25	114	TGCCAG gtaag	82	tccag GGCTGC	0	1176-Gln/Gly	
26	129	GGCCAG gtgcg	117	ctcag GCTGCC	0	1219-Gln/Ala	
27	86	TGAGAG gagag	260	tgcag GTGCGC	II	1248-Arg	
28	50	GCCAGA gtgag	>200	cagag GAGACA	I	1265-Arg	
29	95	TGCAAG gtcag	237	cccag GCCCAG	0	1296-Lys/Ala	
30	141	CACCTG gtgag	238	gccag ATCTCC	0	1343-Leu/Ile	
31	192	GACAAG gtggg	>1000	tgcag GTGGCG	0	1407-Lys/Val	
32	93	ATCACG gtgag	>2000	tccag GGCAAC	0	1438-Thr/Gly	
33	81	CGAGAG gtaag	210	cccag GAATTC	0	1465-Glu/Glu	
34	231	TGCCAG gtgag	615	cctag GACTGT	0	1542-Gln/Asp	
35	114	TGCTCG gtgac	>472	tgcag CAATGC	0	1580-Ser/Gln	
36	128	GAACAT gtggg	144	cccag GTTTTC	II	1623-Met	
37	87	TGAGCA gtaag	>394	tacag GTGTGG	II	1652-Gln	
38	59	CAGAGA gtaag	183	cccag CAAACC	I	1672-Thr	
39	167	CATCAA gtacc	>50	tgcag GGCTCC	0	1727-Gln/Gly	
40	112	TCACTG gtgag	>948	tgcag AGGCTC	I	1765-Glu	
41	100	AGCAAG gtgtg	>206	tgaag TCCCCA	0	1798-Lys/Ser	
42	180	TGCAGG gtaca	>1100	gtcag CCTCGG	I	1859-Ala	
43	126	GGACAG gtgag	131	gacag GGGGCC	I	1901-Gly	
44	153	TGCATG gtgag	86	tgcag GGGGCG	I	1952-Gly	
45	143	CCACAG gtgag	>526	ggcag GCCCGG	0	1999-Gln/Ala	
46	136	TTTCAG gtacg	82	cccag CCTCAG	I	2045-Ala	
47	155	ACCCAG gtatt	>431	caaag GTGCAC	0	2096-Gln/Val	
48	151	CCCCAG gtgag	>680	gtcag TGCCCG	I	2147-Val	
49	152	CACCAG gtatg	86	cacag ACCCAC	0	2197-Gln/Thr	
50	133	TCCCTG gtgag	>165	tgcag GACCCA	I	2242-Gly	
51	146	CACCAG gtaca	164	cccag GTTCGT	0	2290-Gln/Val	
52	136	CCTACC gtgag	92	ctcag CTGCCG	I	2336-Pro	
53	152	CACCAG gtatg	>4000	gccag ACCCAC	0	2386-Gln/Thr	
54	136	TGCCTG gtgag	171	cccag CACTTG	I	2432-Ala	
55	152	CACCAG gtagg	>200	cccag GTGCAT	0	2482-Gln/Val	
56	139	CCCACT gtgag	105	cccag CCCAGG	I	2529-Ser	
57	152	CACCAG gtaca	>260	tgcag ATCGTG	0	2579-Gln/Ile	
58	136	CCCACG gtgag	>80	tncag TGCCCA	I	2625-Val	
59	152	CACCAG gtaca	>200	tccag ACCCAT	0	2675-Gln/Thr	IV
60	139	CCTCCG gtgag	210	cnnag CCCCTG	I	2722-Ala	
61	152	CATCAG gtatg	>210	cccag ACCCGC	0	2772-Gln/Thr	
62	149	TCCCCG gtgag	>167	cccag CCCCAG	I	2822-Ala	
63	152	CACCAG gttaa	>300	cgcag GTCCAC	0	2872-Gln/Val	
64	142	TTCCTG gtgag	385	cacag CTCCAG	I	2920-Ala	
65	152	CACCAG gtaca	359	gacag ACCCAT	0	2970-Gln/Thr	
66	142	CCTACC gtgag	96	cccag GCCTTA	I	3018-Arg	
67	141	TGGAGG gttga	332	cacag ACAACG	I	3065-Asp	
68	135	TGCACG gtgag	>67	cacag GGCCCC	I	3110-Gly	
69	184	GCTGCA gtctg	337	cccag GATTTC	II	3171-Gln	
70	197	CCACAG gtgag	>214	cacag GCAGCC	I	3237-Gly	
71	180	TGGAGA gtaag	>658	gtcag GCCCAC	I	3297-Ser	
72	261	TCCAAG gtgag	>800	cgcag GCCCTC	I	3384-Gly	
73	205	GCTCCG gtgag	>610	tatag AATCCA	II	3452-Arg	
74	98	TCCAAG gtagg	>400	gccag CCCTGC	I	3485-Ala	
75	267	TGCAAG gtgca	>400	ctcag CCTTGC	I	3574-Ala	
76	110	AGCAAG gtaag	199	tgcag CTGGAT	0	3610-Lys/Leu	
77	148	TGCCAG gtaag	95	ggcag AGCGGG	I	3660-Glu	
78	117	CCGATG gtgag	>300	cccag GGATGC	I	3699-Gly	
79	112	GTTCCG gtgag	388	aggag GTTCGA	II	3736-Arg	
80	145	TCCCAG gtgag	>200	tgcag GGCAAG	0	3784-Gln/Gly	
81	210	TGCCAG gtgat	124	tgcag AATGGC	0	3854-Gln/Asn	
82	109	ATCCAG gtacg	>700	cccag AGGCCT	I	3891-Glu	
83	99	AGGAAG gtagg	>500	tgcag GTGTGA	I	3924-Gly	
84	222	GGTCAG gtaag	>300	tgcag GGCTGG	I	3998-Gly	V
85	245	GGCGAG gtgag	>900	cacag GTGTCA	0	4079-Glu/Val	
86	172	TCAAAG gtggc	106	cacag GAGACC	I	4137-Gly	
87	123	AACAAG gtacc	110	catag GCTCTG	I	4178-Gly	
88	57	GCAATG gtgag	>650	cacag ATGCCC	I	4197-Asp	
89	76	CAGGAG gtgag	86	cgtag CCTGCC	II	4222-Ser	
90	79	GGTGTG gtgag	161	ggcag GAGGTG	0	4248-Val/Glu	
91	71	CTTCAG gtagg	97	tgcag GTACCA	II	4272-Arg	
92	84	ACTGCG gtaag	>600	tgcag GGAGGG	II	4300-Arg	
93	104	ACATCG gtaag	127	ggcag GCGGAG	I	4335-Gly	
94	1244						

FIG. 2. Intron/exon organization of the human perlecan gene. Exon sequences are in capital letters; intron sequences are in lowercase letters. Introns that do not split codon triplets are indicated by phase 0, interruption after the first nucleotide is indicated by phase I, and interruption after the second nucleotide is indicated by phase II. Amino acids encoded at the splice site are indicated with the numbers based on the human perlecan cDNA.

able conservation of the modular domains with those from homologous genes. These modules were often composed of exons flanked at either side by introns in the same phase. It has been proposed that this homogeneity in codon phasing may have promoted the coupling of functional domains through the process of intronic recombination by preserving a continuous open reading frame (19). Therefore, we present below the exonic organization of human perlecan gene vis-à-vis that of the homologous protein domains.

Leader Exon. The first exon coded for the 5′ untranslated region and the signal peptide (Fig. 2). Primer extension and S1 nuclease mapping confirmed the length of exon 1 (see below) in contrast to the mouse gene, which harbors a relatively large 5′ untranslated region (5).

Domain I. This domain was the only region unique to perlecan inasmuch as it lacked homology to any other protein (5–7). It contained three SGD sequences, the attachment sites for the heparan sulfate side chains (6), and was encoded by five distinct exons (exons 2–6) (Fig. 2). The proximal SGD triplet resided between exons 2 and 3, with the junction splicing the Asp codon in phase I, while the two distal SGD sequences were contained in exon 3, which also ended in phase I. Therefore, it is plausible that if exon 3 is lost by alternative splicing, all three heparan sulfate attachment sites would be absent, thereby generating a species of perlecan lacking glycosaminoglycan chains. Supporting this notion is the observation that the newly synthesized perlecan in colon carcinoma cells is partially secreted into the medium without any heparan sulfate chains (10). In addition, the nematode homologue of perlecan does not carry any SGD in domain I (20).

Taken together, these data suggest that perlecan may be occasionally secreted as a protein rather than a proteoglycan and raise the possibility of a cell- or tissue-specific control of this process.

Domain II. This domain, homologous to the well characterized LDL receptor (21), was encoded by three distinct exons. The first two repeats of perlecan, including the spacer region, were encoded by a single exon, exon 7. Similarly, in the LDL-receptor gene (21), the corresponding regions (repeats III–V) are also encoded by one exon of remarkably similar size (384 bp for perlecan and 381 bp for LDL receptor). The distal two repeats of perlecan were both encoded by a single, highly conserved exon when compared to the ligand region of the LDL receptor. In addition to the similarity in exon length, the intron phase was identical in both molecules, with all the exons being interrupted in phase I (Fig. 2). Consequently, perlecan and the ligand region of the LDL receptor possess similar features, with the binding site for apolipoproteins (DGSDGE) situated within the largest of the three exons. It has been suggested that the exon arrangement of the LDL-receptor ligand binding domain could generate via alternative splicing different affinity receptors for apolipoproteins (21). Heparan sulfate proteoglycans have recently been implicated in the binding of apolipoprotein B via the heparan sulfate chains (22). Therefore, perlecan protein core, through its LDL binding site, could play a role in the regulation of LDL by binding or storing apolipoproteins in the subendothelial basement membrane or in the extracellular matrix. A lipid-binding activity for perlecan could also be invoked in the hepatic perisinusoidal region inasmuch as perlecan is present in large amounts in the space of Disse (4).

Domain IIa contains the first IgG repeat and is encoded by exons 10 and 11 (see below).

Domain III. This domain, homologous to the N-terminal region of both laminin A and B short arms, was encoded by 27 separate exons of various sizes (50–231 bp). Interestingly, no correlation was detected between exon arrangement and either nominal domain or cysteine-repeat boundaries. A similar comparison between the laminin B1 (23) and B2 (24) genes showed considerable divergence between these molecules and no conservation of exon structure and domain location. We can therefore assume that the laminin portion of the perlecan gene must have evolved from an ancestral gene and has undergone extensive rearrangements. In addition, we found an exact duplication of the distal part of exon 30 in a genomic region adjacent to exon 23 (data not shown). However, this duplicated exon contained a different 5′ splice site. This could generate a shorter form of the laminin-like domain of perlecan, further accentuating the divergence of this module from its laminin homologue.

Domain IV. This extended domain, which harbors the longest array of IgG-like repeats, was encoded by 40 separate exons whose organization showed striking similarity to that of the N-CAM gene (25). Apart from IgG repeats 17–21, the remaining N-CAM-like repeats showed a one repeat/two exon structure, including the IgG repeat located in domain IIa. The sequences that had the strongest internal homology followed a strict rule: the first half of the repeat was always interrupted after the doublet His-Glu in phase 0 and the other half of the repeat invariably ended in phase I. The same pattern of exon organization has been detected in both the N-CAM (25) and the lymphocyte T4 protein (26). This arrangement appeared to be specific for perlecan, as other proteoglycans with IgG folds, such as aggrecan, display a one repeat/one exon pattern (27). Such strict adherence to intron–exon junction phase makes this region a likely candidate for differential splicing of the repeats to generate variant forms of perlecan. Evidence for alternative splicing has been presented for mouse perlecan, which contains seven fewer repeats than the human molecule (3, 5). The first half of the mouse IgG repeat 5 aligns with human repeat 5, whereas the second half aligns with human repeat 12, thus supporting the existence of alternatively spliced transcripts in this region.

Domain V. This domain, homologous to the globular C terminus of the laminin A chain, was encoded by 16 exons, including the largest exon of 1.2 kbp, harboring the entire 3′ untranslated region. Sequencing an additional 500 bp of 3′ flanking region revealed no additional polyadenylylation signals (data not shown), confirming our previous data from cDNA cloning (6). This finding further stresses the concept that various transcripts previously detected for perlecan (3) are due not to differential usage of polyadenylylation sites but rather to alternative splicing of internal exons. As in the case of the other laminin-like region, the organization of domain V also lacked correlation between domain boundaries and exon structure. An additional feature of this domain was the presence of very small consecutive exons at its 3′ end, the significance of which is not yet apparent.

The 5′ Flanking Region of Human Perlecan Gene Lacks Canonical TATA and CAAT Boxes and Is Contained in a CpG Island. The putative promoter and exon 1 were isolated by screening the chromosome 1-specific library with the first 143 bp of the perlecan cDNA; >3.5 kbp of 5′ flanking region was sequenced and ≈1 kbp is presented in Fig. 3*A*. Of note, the perlecan gene promoter lacked canonical TATA or CAAT boxes, but it contained four GC boxes and three GGGCGG hexanucleotides, which are binding sites for the zinc finger transcription factor SP1 (28). In addition, 5 GGGCGG hexanucleotides and a GC box were found in the first exon and the first intron, respectively (Fig. 3*A*). The perlecan promoter exhibited a high G+C content, with >80% GC in the 500 bp immediately upstream to exon 1 (Fig. 3*B*) and a CpG/GpC ratio of ≈0.9, thus indicating the absence of CpG suppression (29). In addition, this region contained 21 CCGG tetranucleotides, the recognition site for the methylation-sensitive *Hpa* II enzyme. Collectively, these features indicate that the 5′ end and promoter region of the human perlecan gene are found in a CpG island, also referred to as *Hpa* II tiny fragment (HTF) island (29), a portion of DNA that contains clusters of unmethylated CpG dinucleotides in a microenvironment in which

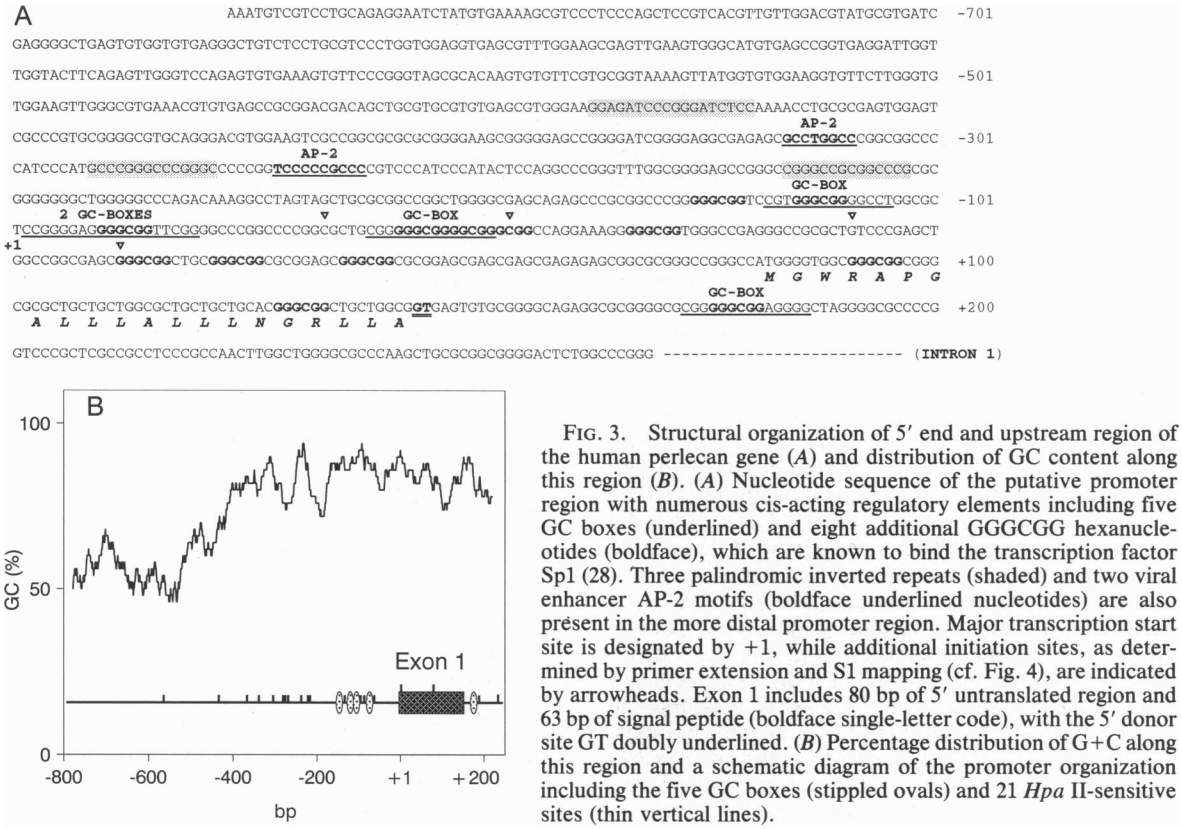

FIG. 3. Structural organization of 5′ end and upstream region of the human perlecan gene (*A*) and distribution of GC content along this region (*B*). (*A*) Nucleotide sequence of the putative promoter region with numerous cis-acting regulatory elements including five GC boxes (underlined) and eight additional GGGCGG hexanucleotides (boldface), which are known to bind the transcription factor Sp1 (28). Three palindromic inverted repeats (shaded) and two viral enhancer AP-2 motifs (boldface underlined nucleotides) are also present in the more distal promoter region. Major transcription start site is designated by +1, while additional initiation sites, as determined by primer extension and S1 mapping (cf. Fig. 4), are indicated by arrowheads. Exon 1 includes 80 bp of 5′ untranslated region and 63 bp of signal peptide (boldface single-letter code), with the 5′ donor site GT doubly underlined. (*B*) Percentage distribution of G+C along this region and a schematic diagram of the promoter organization including the five GC boxes (stippled ovals) and 21 *Hpa* II-sensitive sites (thin vertical lines).

the number of CpG approximately equals the number of GpC (29). These HTF islands have been correlated with transcriptional control regions and are typically observed in genes that encode oncoproteins, growth factors, transcription factors, and housekeeping proteins (30). Similar promoter features are observed in several extracellular matrix genes, including those encoding α1 and α2 chains of human type IV collagen (31), chicken α2 (VI) collagen (32), human laminin B1 (23) and B2 (24) chains, human elastin gene (33), and α5 integrin subunit (34). The more distal part of the perlecan promoter contained two viral enhancer AP-2 motifs and three short palindromic direct and indirect repeats, which, by forming secondary structure, could influence the regulation of perlecan gene expression. Additional features of perlecan promoter predict that the G+C-rich regions, particularly the 5′-CCCC-3′ motifs, could bind the transcription factor ETF, which stimulates transcription of promoters lacking TATA boxes but enriched in these polycytosine stretches (35).

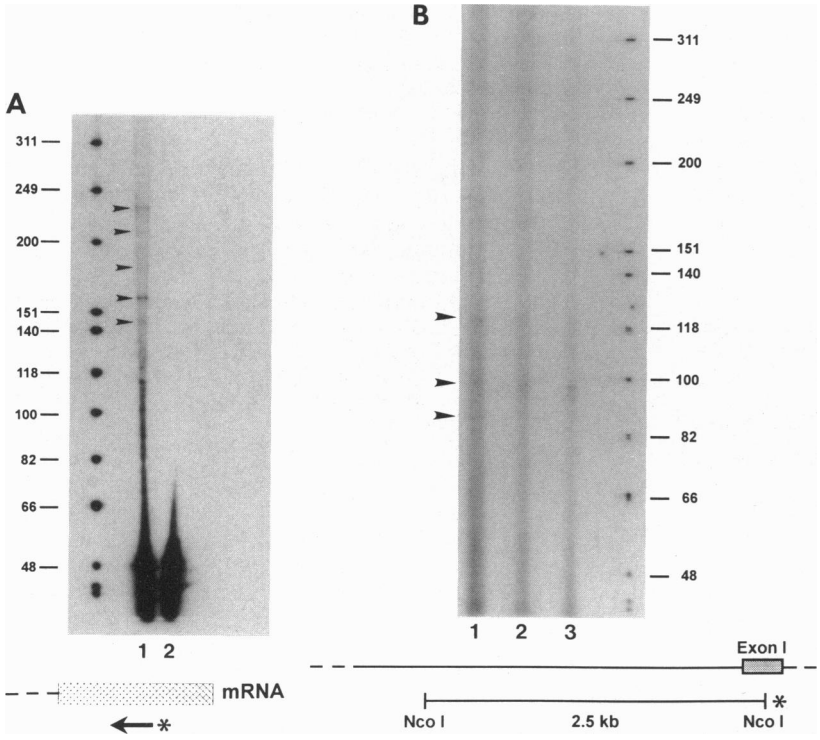

FIG. 4. Multiple transcription start sites for the perlecan message as determined by primer extension (*A*) or S1 nuclease mapping (*B*) assays. (*A*) (*Lower*) Schematic representation of the 25-mer oligonucleotide (arrow) used for priming with the labeling position designated by an asterisk. (*Upper*) Lane 1, primer-extended product using poly(A)$^+$ RNA from human colon carcinoma cells. Lane 2, primer extension using tRNA as template. Multiple transcription initiation sites are indicated by arrowheads. (*B*) For S1 nuclease mapping, a 2.5-kbp *Nco* I fragment containing 84 bp of exon 1 and 5′ flanking region (shown schematically below) was end-labeled and hybridized to total RNA from human colon carcinoma cells. S1 nuclease digestion (500 units/ml) was performed for 5 (lane 1), 15 (lane 2), and 30 (lane 3) min. When tRNA was used as template, no protected bands were seen (not shown). Numbers on left in *A* and on right in *B* designate migration of size markers in bp derived from ^{32}P-labeled *Hin*fI digests of ΦX174 viral DNA.

Multiple Transcription Start Sites for the Human Perlecan Message. The absence of a canonical TATA box suggested the possibility that there might be multiple sites of transcription initiation. To test this possibility and to provide support for the notion that the 5′ flanking region contained the promoter of the perlecan gene, primer extension and S1 nuclease protection assays were performed (Fig. 4) with RNA from colon carcinoma cells, which express perlecan message at high levels (15). Because of the high G+C content of the first exon, primer extension was performed with a primer located at position +168 of the cDNA. These experiments detected five major bands (Fig. 4*A*), mapping the transcription start sites at −67, −47, −10, +1, and +12 bp, respectively. The S1 nuclease protection assay, using as a probe a 2.5-kbp *Nco* I fragment containing the first 84 bp of exon 1, revealed three distinct fragments of 84, 95, and 149 bp, respectively (Fig. 4*B*). These three fragments corresponded to the putative start sites at positions −65, −10, and +1 as detected above. The two additional bands detected by primer extension could represent premature termination by the reverse transcriptase. The consistency of these two independent analyses clearly indicates that human perlecan mRNA transcription is initiated at multiple sites. Based on the intensity of the bands, we estimated the site at +1 to be the predominant start site of transcription.

Conclusions. The present communication has elucidated the genomic organization of the perlecan gene encoding the major heparan sulfate proteoglycan of basement membranes. This multidomain molecule exhibits features that are remarkable not only in terms of exonic organization but also in terms of conservation of each discrete functional unit. The 94 exons are organized in a domain-specific manner and their clustering provides insights into the evolutionary relationship inferred from the cDNAs of various species. The modules are segregated into separate exonic groupings that reflect the functional domains of the molecule. Strikingly, for both domains II and IV, the exon organization is nearly identical to that followed by the LDL receptor and the N-CAM genes, respectively. Both the exon size and the exon–intron phases are invariably conserved between perlecan and these two genes, suggesting that perlecan has evolved by gene duplication or exon shuffling (19). Another interesting feature of the perlecan gene is the similarity in the organization of the two laminin-like regions (domains III and V, respectively) with those in laminin B1 and B2 chains (23, 24). In synchrony with these two related genes, perlecan also shows no significant correlation between the exonic organization and the location of these domains with respect to their internal repeats. Finally, our study predicts the generation of alternatively spliced variants of perlecan in both domains I and IV. These differentially spliced forms could lead to the identification of tissue-specific or developmentally regulated perlecan variants.

This work was supported by National Institutes of Health Grants CA-39481 and CA-47282 and by a Faculty Research Award (FRA-376) from the American Cancer Society (to R.V.I.).

1. Hay, E. D. (1991) *Cell Biology of the Extracellular Matrix* (Plenum, New York).
2. Timpl, R. (1993) *Experientia* **49**, 417–428.
3. Noonan, D. M. & Hassell, J. R. (1993) *Kidney Int.* **43**, 53–60.
4. Murdoch, A. D. & Iozzo, R. V. (1993) *Virchow Arch. A Pathol. Anat. Histopathol.*, in press.
5. Noonan, D. M., Fulle, A., Valente, P., Cai, S., Horigan, E., Sasaki, M., Yamada, Y. & Hassell, J. R. (1991) *J. Biol. Chem.* **266**, 22939–22947.
6. Murdoch, A. D., Dodge, G. R., Cohen, I., Tuan, R. S. & Iozzo, R. V. (1992) *J. Biol. Chem.* **267**, 8544–8557.
7. Kallunki, P. & Tryggvason, K. (1992) *J. Cell Biol.* **116**, 559–571.
8. Hassell, J. R., Gehron-Robey, P., Barrach, H.-J., Wilczek, J., Rennard, S. I. & Martin, G. R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4494–4498.
9. Iozzo, R. V. (1984) *J. Cell Biol.* **99**, 403–417.
10. Iozzo, R. V. & Hassell, J. R. (1989) *Arch. Biochem. Biophys.* **269**, 239–249.
11. Heremans, A., Van Der Schueren, B., De Cock, B., Paulsson, M., Cassiman, J.-J., Van Den Berghe, H. & David, G. (1989) *J. Cell Biol.* **109**, 3199–3211.
12. Saku, T. & Furthmayr, H. (1989) *J. Biol. Chem.* **264**, 3514–3523.
13. Jalkanen, M., Rapraeger, A. & Bernfield, M. (1988) *J. Cell Biol.* **106**, 953–962.
14. Cohen, I. R., Grässel, S., Murdoch, A. D. & Iozzo, R. V. (1993) *FASEB J.* **7**, A836 (abstr.).
15. Dodge, G. R., Kovalszky, I., Chu, M. L., Hassell, J. R., McBride, O. W., Yi, H. F. & Iozzo, R. V. (1991) *Genomics* **10**, 673–680.
16. Feinberg, A. P. & Vogelstein, B. (1984) *Anal. Biochem.* **137**, 266–267.
17. Danielson, K. G., Fazzio, A., Cohen, I., Cannizzaro, L. A., Eichstetter, I. & Iozzo, R. V. (1993) *Genomics* **15**, 146–160.
18. Chen, E. Y. & Seeburg, P. H. (1985) *DNA* **4**, 165–170.
19. Patthy, L. (1991) *Curr. Opin. Struct. Biol.* **1**, 351–361.
20. Rogalski, T. M., Williams, B. D., Mullen, G. P. & Moerman, D. G. (1993) *Genes Dev.* **7**, 1471–1484.
21. Südhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.
22. Williams, K. J., Fless, G. M., Petrie, K. A., Snyder, M. L., Brocia, R. W. & Swenson, T. L. (1992) *J. Biol. Chem.* **267**, 13284–13292.
23. Vuolteenaho, R., Chow, L. T. & Tryggvason, K. (1990) *J. Biol. Chem.* **265**, 15611–15616.
24. Kallunki, T., Ikonen, J., Chow, L. T., Kallunki, P. & Tryggvason, K. (1991) *J. Biol. Chem.* **266**, 221–228.
25. Cunningham, B. A., Hemperly, J. J., Murray, B. A., Prediger, E. A., Brackenbury, R. & Edelman, G. M. (1987) *Science* **230**, 799–806.
26. Littman, D. R. & Gettner, S. N. (1987) *Nature (London)* **325**, 453–455.
27. Doege, K. J., Sasaki, M. & Yamada, Y. (1990) *Biochem. Soc. Trans.* **18**, 200–202.
28. Kadonaga, J. T., Jones, K. A. & Tjian, R. (1986) *Trends Biochem. Sci.* **11**, 20–23.
29. Bird, A. P. (1986) *Nature (London)* **321**, 209–213.
30. Kozak, M. (1992) *Annu. Rev. Cell Biol.* **8**, 197–225.
31. Soininen, R., Huotari, M., Ganguly, A., Prockop, D. J. & Tryggvason, K. (1989) *J. Biol. Chem.* **264**, 13565–13571.
32. Koller, E., Hayman, A. R. & Trueb, B. (1991) *Nucleic Acids Res.* **19**, 485–491.
33. Bashir, M. M., Indik, Z., Yeh, H., Ornestein-Goldstein, N., Rosenbloom, J. C., Abrams, W., Fazio, M., Uitto, J. & Rosenbloom, J. (1989) *J. Biol. Chem.* **264**, 8887–8891.
34. Birkenmeier, T. M., McQuillan, J. J., Boedeker, E. D., Argraves, W. S., Ruoslahti, E. & Dean, D. C. (1991) *J. Biol. Chem.* **266**, 20544–20549.
35. Kageyama, R., Merlino, G. T. & Pastan, I. (1988) *J. Biol. Chem.* **264**, 15508–15514.