Mouse aggrecan, a large cartilage proteoglycan: protein sequence, gene structure and promoter sequence

Hideto WATANABE,* Luoyi GAO,* Satoru SUGIYAMA,* Kurt DOEGE,† Koji KIMATA‡ and Yoshihiko YAMADA*§

*Laboratory of Developmental Biology, National Institute of Dental Research, National Institutes of Health, Bethesda, MD 20892, U.S.A.,
†Department of Biochemistry and Molecular Biology, Oregon Health Science University, Portland, OR 97201, U.S.A., and
‡Institute for Molecular Science of Medicine, Aichi Medical University, Nagakute, Aichi 480-11, Japan

Seven genomic clones for mouse aggrecan core protein have been isolated including 3 kb of 5′- and 7 kb of 3′-flanking sequences. All exon sequences and their intron boundary sequences in these clones were identified and mapped by DNA sequencing. The gene spans at least 61 kb and contains 18 exons. Exon 1 encodes 5′-untranslated sequence and exon 2 contains a translation start codon, methionine. The coding sequence is 6545 bp for a 2132-amino-acid protein with calculated M_r = 259131 including an 18-amino-acid signal peptide. There is a strong correlation between structural domains and exons. Notably, the chondroitin sulphate domain consisting of 1161 amino acids is encoded by a single exon of 3.6 kb. Although link protein has similar structural domains and subdomains, the sequence identity and the organization of exons encoding the subdomains B and B′ of G1 and G2 domains revealed a strong similarity of mouse aggrecan to both human versican and rat neurocan. Primer extension analysis identified four transcription start sites which are close together. The promoter sequence showed high G/C content (65%) and contained several consensus binding motifs for transcription factors including Sp-1 and the glucocorticoid receptor. There are stretches of sequences similar to the promoter region of both the type-II collagen and link protein genes. These sequences may be important for cartilage gene expression.

INTRODUCTION

Aggrecan, a large aggregating proteoglycan, is one of the major structural constituents of cartilage. Aggrecan forms aggregates with hyaluronic acid. Link protein, a small glycoprotein which shares structural similarity with the N-terminal region of aggrecan, stabilizes the aggregates by interacting with both hyaluronic acid and aggrecan. The aggregates have a unique gel-like property and function to resist compression and deformation (for review, see [1,2]). A deficiency of aggrecan has been described in cartilage of *cmd* (cartilage matrix deficiency) mice which are characterized by a cleft palate and short trunk, limbs and snout [3]. Similar defects are also found in chick nanomelia [4].

Aggrecan consists of a protein core of approximately 220000-M_r which is extensively modified by covalently attached glycosaminoglycan side-chains and oligosaccharides. Biochemical studies have demonstrated many of the properties of this proteoglycan, including its binding to hyaluronic acid via an N-terminal globular region [5] and its extensive modification with chondroitin sulphate (CS) (up to 100 chains per monomer), keratan sulphate (KS) and other oligosaccharides [6]. cDNA cloning and sequencing studies have yielded a complete deduced primary structure for aggrecan from various species such as rat [7], human [8], chicken [4,9,10] and a partial one for mouse [11] which provide a structural model for aggrecan (for review, see [12]). Aggrecan has three globular domains (G1, G2, G3), and two glycosaminoglycan attachment domains (KS and CS domains) located between G2 and G3. The two globular domains, G1 and G2, comprise the N-terminus of the proteoglycan, while G3 makes up the C-terminus. G1 has hyaluronic acid-binding activity and interacts with link protein. G2 is similar to both G1 and link protein, but its function is still unknown. G3 is a complex structure containing epidermal growth factor (EGF)-like, lectin-like and complement regulatory protein (CRP)-like domains. The CS domain provides a number of attachment sites for CS, which enables retention of water in cartilage tissue. A short proline-rich sequence proximal to the CS domain has been shown to be the region most densely substituted with KS and is designated the KS domain. Similar domain structures are found in other proteoglycans such as versican [13,14] and neurocan [15,16]. Whereas cDNA data are available from different species, only partial structures of the aggrecan gene have been reported for rat [17] and chicken [18].

In the present study, we isolated the mouse aggrecan gene from genomic libraries and determined exon sequences and exon–intron junctions. We also identified the transcription start sites by primer extension and characterized the promoter region of the aggrecan gene by DNA sequencing. This information will be useful for the study of the regulation of the aggrecan gene and of the function of aggrecan in development, especially using a transgenic mouse model.

MATERIALS AND METHODS

Genomic libraries

Two mouse genomic libraries, 129 SvJ in the λ Fix II vector (Stratagene, La Jolla, CA, U.S.A.), and Balb/c in the EMBL3/SP6/T7 vector (Clontech, Palo Alto, CA, U.S.A.) were screened with ^{32}P-labelled mouse cDNA (pMAG8-115) [19] and rat cDNA [7] for aggrecan. The hybridization was performed in 6 × SSC

Abbreviations used: KS, keratan sulphate; CS, chondroitin sulphate; *cmd*, cartilage matrix deficiency; EGF, epidermal growth factor; CRP, complement regulatory protein; 10 × SSC, 1.5 M NaCl/0.15 M sodium citrate; 50 × Denhardt's; 1% Ficoll/1% polyvinylpyrrolidone/1% BSA; IGD, interglobular domain.

§To whom correspondence should be addressed.

The nucleotide sequence reported in this paper has been submitted to the GenBank®/EMBL Data Bank with accession number U22901.

(SSC: 0.15 M NaCl/0.015 M sodium citrate)/10 × Denhardt's solution/0.1 % SDS/100 μg/ml denatured salmon sperm DNA at 62 °C overnight. The filters were washed twice in 2 × SSC/0.1 % SDS at room temperature for 5 min, and in 0.1 × SSC/0.1 % SDS at 62 °C for the mouse probe or at room temperature for the rat probe, respectively, for 20 min twice. Positive phage clones were re-screened until pure and grown in liquid culture. Phage were purified by CsCl density gradient and DNA was extracted as previously described [20]. Restriction sites of these phage DNA were mapped by partial digestion with various restriction enzymes and by Southern hybridization with oligonucleotides as probes using Flash Non-radioactive Gene Mapping Kit (Stratagene).

DNA sequencing

For sequencing purposes, Southern blots of restriction enzyme digests of genomic clones were hybridized with cDNA probes and positive DNA fragments were subcloned into pBluescript II SK(−). In addition to oligonucleotide primers corresponding to the T3 and T7 recognition sites of the vector, 90 specific synthetic oligonucleotides were prepared and used as sequencing primers. Sequencing was performed on the double-stranded DNA using [^{35}S]dATP and Sequenase (USB, Cleveland, OH, U.S.A.). In some cases, cycle DNA sequencing was performed in a GenAmp PCR System 9600 (Perkin-Elmer, Norwalk, CT, U.S.A.) for 30 cycles of 96 °C, 30 s; 50 °C, 15 s; 72 °C, 4 min with the PRISM ready reaction kit (Applied Biosystems, Foster City, CA, U.S.A.). Sequence reactions were analysed by an automated DNA sequencer, ABI model 373A (Applied Biosystems). The sequences were analysed using GCG software.

Primer extension

The start site of transcription of the mouse aggrecan gene was analysed by primer extension. Total RNA was isolated from the sternal cartilage of newborn mice using the guanidium isothiocyanate method [21]. A specific oligonucleotide (5′-GAGAGAGAGTGAGGGACCTTGAGCCGTTCC-3′) labelled with [γ-^{32}P]ATP at the 5′-end and 50 μg of total RNA were annealed and cDNA was extended by reverse transcriptase (M-MLV, Stratagene, La Jolla, CA, U.S.A.). The primer-extended product was analysed on a 6 % polyacrylamide sequencing gel in parallel with dideoxy sequencing reactions primed with the same oligonucleotide.

RESULTS AND DISCUSSION

Genomic clones for mouse aggrecan

Screening of the mouse genomic (129 SvJ) library was performed under high stringency using a mouse aggrecan cDNA, pMAG8-115 [19], as a probe and yielded a clone, λmag-1, containing the 3′ portion of the aggrecan gene. Five additional clones, λmag-2, λmag-3, λmag-4, λmag-9 and λmag-10, were obtained by screening with cDNAs encoding various portions of the rat aggrecan gene [7]. Restriction maps of these clones showed considerable overlaps, indicating that they represent the same gene. Clone λmag-11 was isolated from a mouse genomic (Balb/c, EMBL3/SP6/T7) library by screening with a PCR product containing nucleotide residues 6–121 of the rat aggrecan cDNA sequence as a probe. Although λmag-11 did not overlap with other clones, it contained the most 5′ end of the gene. There is also a gap between clones λmag-1 and λmag-3. These seven clones and their relative positions are shown in Figure 1.

Exon structure

All exon sequences including exon–intron boundaries in the seven clones were determined by DNA sequencing. The gene for mouse aggrecan spans more than 61 kb from the transcriptional start site to the polyadenylation site and contains 18 exons (Figure 1). Exon sizes and exon–intron boundary sequences are shown in Table 1. Exon 1 codes for the 5′-untranslated sequence and the translation starts in exon 2. The coding sequence contains 6545 bases for a core protein of 2132 amino acids with a calculated M_r of 259131 including an 18-amino-acid signal peptide (Figure 2).

The sequences at the intron–exon boundaries are in agreement with general consensus splice sequences [22]. The consensus sequence for the 3′ ends of the introns of the aggrecan gene was as follows:

$$\begin{matrix} c_{50}\ a_{56} \\ \qquad c_{100}a_{100}g_{100}/splice \\ t_{38}\ c_{31} \end{matrix}$$

The subscript numbers denote the frequency of the most common nucleotides in percentage terms. The consensus sequence for the 5′ end of the introns was as follows:

$$\begin{matrix} \qquad g_{56}\ a_{63}\ g_{63} \\ splice/g_{100}t_{100} \\ \qquad a_{44}\ g_{13}\ a_{25} \end{matrix}$$

Most of the 18 exons begin with a split codon. These exons include 3–12, 15, 17 and 18. The introns vary in size. The longest is intron 1 which spans more than 21 kb. Intron 8 is the shortest with an approximate size of 190 bp. Sequence comparisons of the mouse aggrecan with the corresponding rat, human and chicken sequences are summarized in Table 2. The overall nucleotide identity of the coding sequence between mouse and rat is 92.7 %, between mouse and human 75.9 %, and between mouse and chicken 65.3 %. The differences occur most frequently in the third positions. Overall amino acid identities of aggrecan core protein are 95.0 % for mouse and rat, 85.9 % for mouse and human and 67.9 % for mouse and chicken. The conservation of nucleotide and amino acid sequences varies, however, between the different domains (Table 2). Significant correspondence between exon and structural domains is observed in the aggrecan gene and these features are described in detail below.

Signal peptide and N-terminal globular domains G1 and G2

The first methionine codon is followed by a presumptive signal peptide of 18 amino acids with 100 % and 75 % sequence identity to rat and human molecules respectively. The two N-terminal globular domains, G1 and G2, show disulphide-bonded structural motifs. G1 consists of three loop-like subdomains, loops A, B and B′, whose structure is similar to link protein. The B and B′ loops form a tandem homologous repeat, a critical structure for hyaluronic acid-binding activity [9]. The A loop shares structural similarity with an immunoglobulin fold and interacts with the A loop of link protein, which stabilizes the interaction of aggrecan and hyaluronic acid [23]. G2 does not have an A subdomain and consists of B and B′ subdomains. G2 lacks hyaluronic acid-binding activity [5]. G1 and G2 are highly conserved among species in both nucleotide and protein sequences (Table 2). In particular, each subdomain of the mouse G1 domain shows 96–100 % sequence similarity to that of rat and human molecules. Similar structural motifs occur in other hyaluronic acid-binding proteins such as human versican [13,14], rat neurocan [15,16], and the lymphocyte homing receptor (CDw44 [24]). The G1

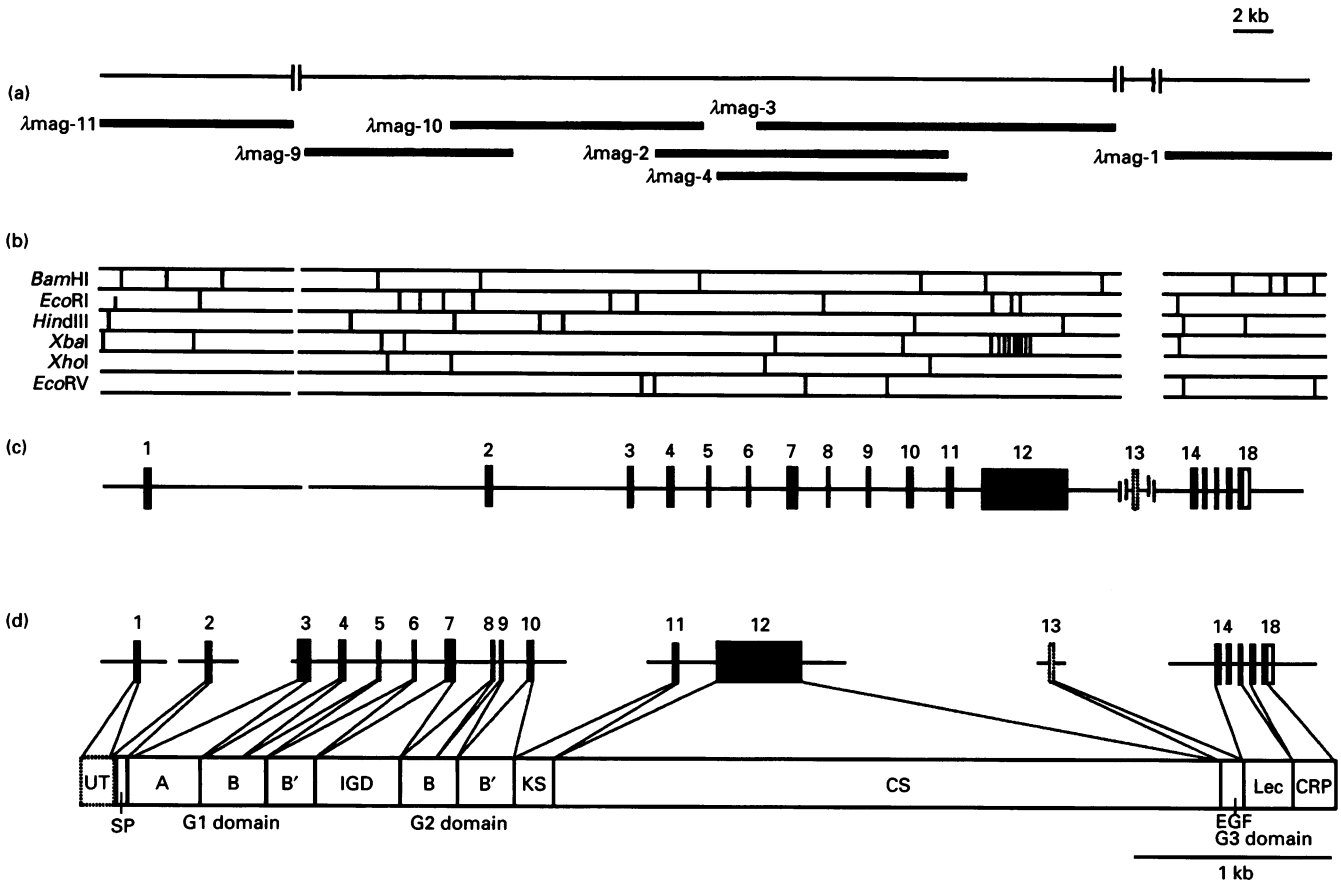
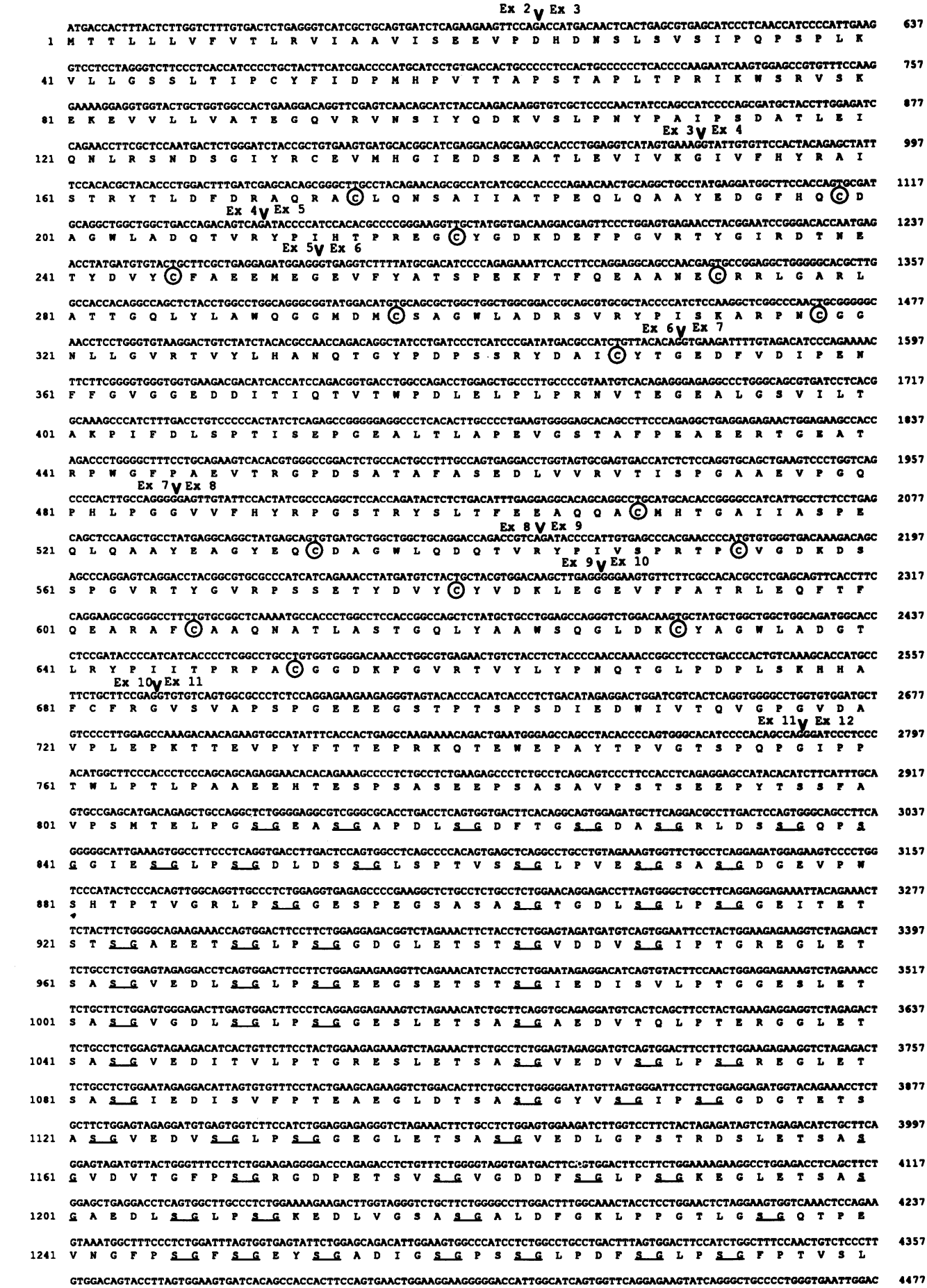


Figure 1 The exon–intron organization and restriction map of the mouse aggrecan gene, and its relationship to structural domains

(**a**) Relative locations of the aggrecan genomic clones. (**b**) Restriction map of the clones. (**c**) A schematic presentation of the exon–intron organization. Exons are numbered. The dashed box shows a potential alternatively spliced exon. (**d**) Relationship of exon and structural domains (UT, untranslated region; SP, signal peptide; KS, keratan sulphate domain; CS, chondroitin sulphate domain; EGF, epidermal growth factor-like domain; Lec, lectin-like domain; CRP, complement recognition peptide-like domain).

Table 1 Exon and intron boundary sequences of the mouse aggrecan gene

Exon	Domain	Exon size (bp)	Acceptor sequence	Exon sequence		Donor sequence	Intron size (kb)
1	5′-UT	509		CCT	AAG	gtaaagaaa	> 21.0
2	SP	77	ctcttccag	CTG	CAG	gtgagaaca	8.5
3	G1-A	384	ctttcacag	ACC	AAG	gtgaagggc	1.8
4	G1-B	175	gtcccacag	GTA	CAG	gtgagactc	2.0
5	G1-B	128	tcccaacag	ATA	AGG	gtgagaaag	1.6
6	G1-B′	294	ccttcacag	GTG	CAG	gtgggactg	1.4
7	IGD	405	ctgttccag	GTG	GGG	gtaagtaca	1.4
8	G2-B	174	tcccgacag	GAG	TCA	gtaaaaaac	0.19
9	G2-B	129	ctttcacag	GAT	AGG	gtacaggcc	1.5
10	G2-B′	294	gttccccag	GGG	GAG	gtactgtag	3.0
11	KS	216	cgtccacag	GTG	CAG	gtggtatag	1.2
12	CS	3482	ttcctccag	GGa	CAG	gtatggagt	?
13	G3-EGF	?	?	?		?	?
14	G3-Lec	158	gtgttacag	ACC	ACA	gtgagtgtg	0.55
15	G3-Lec	83	ccattccag	AAA	CTG	gtgagttcc	0.7
16	G3-Lec	145	tgtccacag	CAA	CCG	gtaagagag	0.9
17	G3-CRP	183	gtccctcag	TGG	ACC	gtgagcatt	0.6
18	G3-3′	74+/	tcgctgcag	CCA			


Ex 2 Ex 3
Ex 3 Ex 4
Ex 4 Ex 5
Ex 5 Ex 6
Ex 6 Ex 7
Ex 7 Ex 8
Ex 8 Ex 9
Ex 9 Ex 10
Ex 10 Ex 11
Ex 11 Ex 12

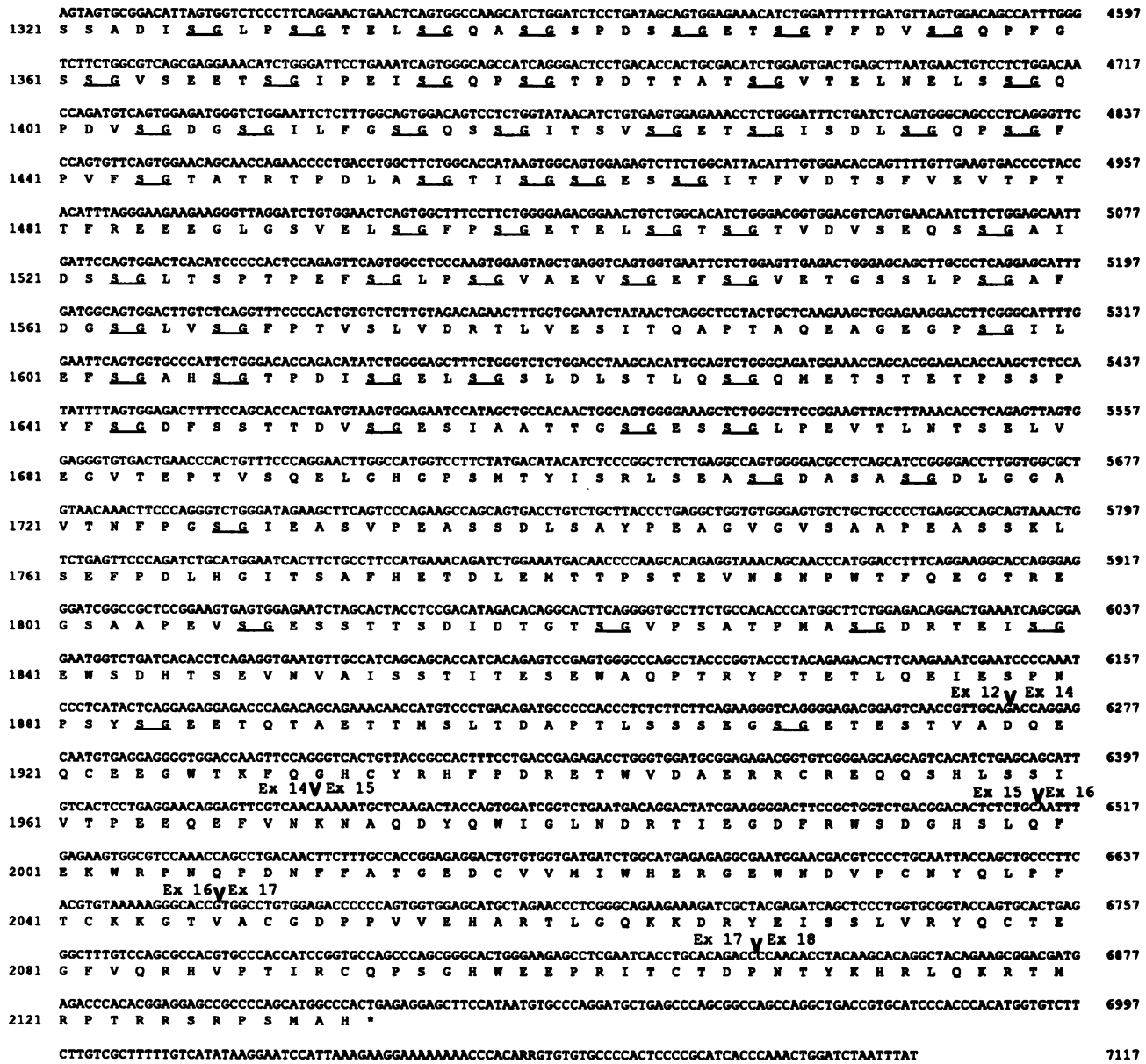


Figure 2 The coding sequence and deduced amino acid sequence of the mouse aggrecan gene

The sequence begins with the translation start site and shows all exons except for the alternatively spliced exon 13. The amino acid numbers on the left and the nucleotide numbers on the right are shown. ▼ shows the exon boundaries. Underlining highlights the Ser-Gly repeats in the CS domain. The cysteine residues in the B and B′ tandem repeats are circled. When we entered the exon sequencing in GenBank, we found the cDNA sequence for mouse aggrecan submitted by another group. There are two differences in the exon sequence: C at 1834 to G and AG at 5316–7 to GA compared with the sequence of the mouse aggrecan cDNA.

region of mouse aggrecan shows 54.8% and 54.3% amino acid sequence identity to human versican and rat neurocan respectively. Human CDw44, which has only one B subdomain, shows 20.6% amino acid sequence identity to the B subdomain of the G1 region of mouse aggrecan. Figure 3(a) shows a pairwise alignment of the homologous B and B′ sequences of mouse aggrecan G1 and G2 domains, mouse link protein, and also human versican. Within the two subgroups, the sequence of the B subdomain is more highly conserved than that of the B′ subdomain. On the B and B′ subdomains, G1 of mouse aggrecan is more similar to human versican and to rat neurocan than to mouse link protein, suggesting that aggrecan is evolutionarily closer to the two proteoglycans than to link protein. Analysis of gene structure revealed that the B loops of the G1 and G2 domains of aggrecan are encoded by two exons: B of the G1 domain by exons 4 and 5, and B of the G2 domain by exons 8 and 9. In contrast, the B′ loops are encoded by a single exon: B′ of the G1 domain by exon 6 and that of the G2 domain by exon 10. It is interesting to note that each of B and B′ loops of link protein is coded for by a single exon (H. Watanabe, unpublished work). By analysis of intron–exon boundaries, the patterns of split of codons are the same among the B and B′ subdomains, supporting the finding of the high conservation of the B and B′ subdomains in aggrecan, link protein and versican (Figure 3b).

Table 2 Comparisons of amino acid and nucleotide sequence identities between mouse, rat, human and chick aggrecan sequences

Abbreviations: aa, amino acid sequence; nt, nucleotide sequence.

	Mouse/rat[a]		Mouse/human		Mouse/chick	
Domain	aa	nt	aa	nt	aa	nt
G1	99.3[b]	95.5	96.4	87.1	77.6	74.5
G1A	100.0	95.9	93.3	86.3	64.0	67.0
G1B	100.0	94.7	100.0	86.8	84.4	80.7
G1B′	97.9	96.0	95.8	88.4	86.5	78.3
IGD	88.1	90.4	77.8	80.1	47.4	65.9
G2	95.5	93.1	89.9	86.7	74.1	75.1
G2B	95.8	92.9	90.8	86.1	72.6	74.8
G2B′	94.8	92.7	88.5	87.3	77.3	75.5
KS	88.5	92.4	68.2	81.6	26.9	61.2
CS	88.3	91.7	70.8	79.5	37.2	63.3
G3[c]	98.1	94.6	88.3	88.6	78.9	78.1

[a]The data for rat, human and chick sequences are from refs. [5], [6] and [8] respectively.
[b]Sequence identity in percentages.
[c]G3 domain without EGF-like domain.

(a)

	G1B′	G2B	G2B′	LP-B	LP-B′	Ver-B	Ver-B′	Neu-B	Neu-B′
G1B	41.4	**68.8**	48.3	47.8	46.6	**72.9**	43.2	**64.5**	46.6
G1B′		42.0	**60.0**	46.1	42.0	41.6	**58.3**	39.3	**53.7**
G2B			47.2	47.8	45.5	**67.7**	37.4	**58.1**	47.7
G2B′				48.3	43.2	41.5	47.9	42.2	**55.2**
LP-B					42.4	44.3	41.6	45.4	46.1
LP-B′						45.1	40.4	41.6	47.7
Ver-B							41.8	**60.6**	41.8
Ver-B′								40.2	**53.1**
Neu-B									38.9

(b)

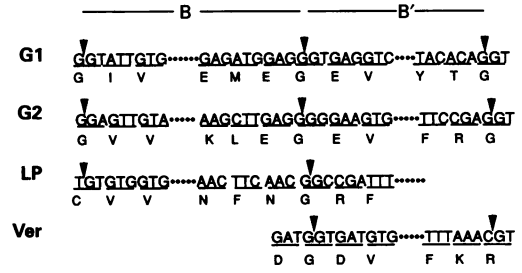


Figure 3 Similarity of sequences encoding the tandem repeat B and B′ subdomains

(**a**) Each pairwise alignment of the amino acid sequences of homologous B and B′ subdomains of mouse aggrecan (G), mouse link protein (LP), human versican (Ver) and rat neurocan (Neu) has been analysed. The numbers indicate the percentage sequence identities. The sequence of mouse link protein is not published (H. Watanabe, unpublished work). (**b**) The split patterns of exons encoding B and B′ subdomains are shown. Arrowheads indicate the split of the codons.

Interglobular domain (IGD)

The IGD is located between the two globular domains. This domain is coded for by exon 7. IGD shows less homology among species. For example, chick has an additional amino acid residue in the carboxyl portion of the domain. IGD is susceptible to several proteinases including collagenases, gelatinases, stromelysin, putative metalloproteinase (PUMP), cathepsin B [25–27], aggrecanase [28] and leucocyte elastase [29]. A consensus sequence for the aggrecanase cleavage site is conserved among species, whereas the metalloproteinase cleavage site is conserved in mammalian species but is not present in chicken. The pig aggrecan contains a putative KS-binding region with the sequence TIQTVT located within the IGD [30]. This hexameric sequence is conserved among mouse, rat, human and chick.

KS domain

The KS domain is located at the C-terminus of the G2 domain, and is coded for by exon 11. The KS domain shows low levels of similarity among species. The size of the domain also varies in different species. The protein sequence similarity of the KS domain is 93.8% for mouse and rat, 76.6% for mouse and human, and 45.2% for mouse and chicken. Putative hexameric KS attachment sites have been reported as E-(E,K)-P-F-P-S and E-E-P-(S,F)-P-S by analysis of the sequence of human [8] and bovine [31] molecules. Since these sequences are not found in mouse, rat or chicken, the key sequence for KS attachment may be different from those sequences in human and bovine.

CS domain

The CS domain is the largest domain, and is located in the middle of the aggrecan core protein. This domain can be divided into two subdomains, CS1 and CS2, based on a difference in specific repeated sequences. However, the whole domain is coded for by a single exon of 3482 bp. The CS domain shows 88.3% amino acid sequence identity between the mouse and rat, although the mouse sequence shows a 10-amino-acid deletion and two additions of 1 and 3 amino acids at different sites. The mouse CS domain contains 120 Ser-Gly repeat sequences. Studies on CS-attachment sites using xylosyl activity of peptides suggested that the putative CS-recognition sequence is S-G-X-G [32]. However, there are only four S-G-X-G sequences in the mouse CS domain. Partial proteolytic digestion followed by HF deglycosylation of chick aggrecan demonstrated the CS-attachment sequence, (D,E)-X-S-G [33]. The mouse and rat CS domain contains 45 repeats of the tetrapeptide sequence. Together with the data that rat chondrosarcoma aggrecan has 80–100 CS chains [6], this indicates that (D,E)-X-S-G may be the candidate for the CS-recognition site rather than S-G-X-G, although some other sequences may also be involved in CS attachment. Like the rat aggrecan CS domain, the mouse CS domain also contains 11 complete or partial repeats of a 40-residue unit and four complete and two partial repeats of a 100-residue sequence [(4X)S-G(2X)-S-G)$_7$(30X)], indicative of a common root sequence, (4X)SG(2X)SG.

G3 domain

The most C-terminal domain G3 is a composite of three structural motifs; EGF-like, lectin-like and CRP-like motifs. The coding region for EGF-like modules is known to be alternatively spliced. Two EGF-like modules, EGF1 [34] and EGF2 [11], have been reported by cDNA sequencing. The human aggrecan gene apparently encodes both EGF1 and EGF2. In mouse, Fulop et al. [11] confirmed that the EGF2 sequence occurs in aggrecan mRNA by cDNA sequencing. They speculated that a potential location of EGF1 is located approximately 1 kb downstream from exon 12 for the CS domain by PCR analysis with mouse genomic DNA using primers from human EGF1. Sequencing of

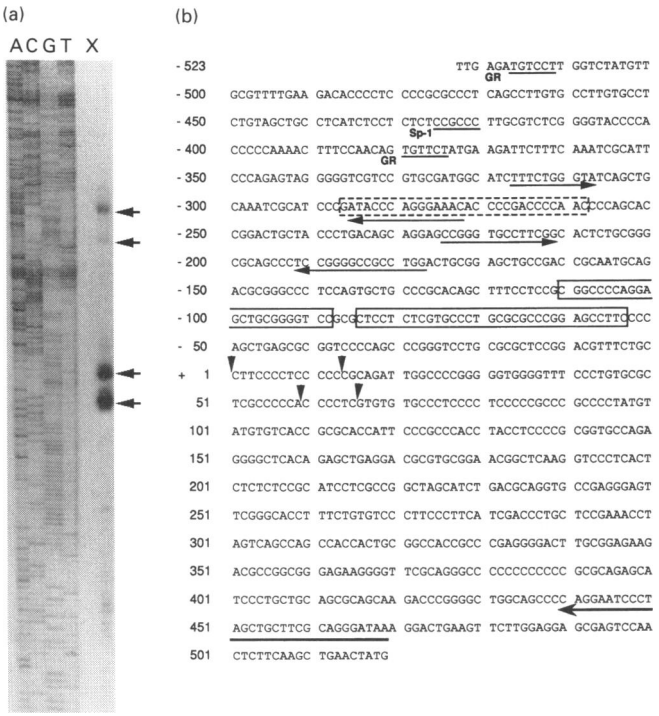


Figure 4 Determination of the transcriptional start site

(**a**) For primer extension of aggrecan mRNA, a 30-mer oligonucleotide complementary to nucleotide residues 177–207 from exon 1 (arrow in **b**) was used to extend cDNA by reverse transcriptase with newborn mouse sternal mRNA. The reaction products were resolved on a denaturing sequencing gel (lane X) with the corresponding sequencing reaction on the left (lanes A, C, G, T). The arrows show the extension products. (**b**) Nucleotide sequence of the promoter and 5′-untranslated region of exons 1 and 2. Four transcription start sites are marked by arrowheads. +1 represents the most upstream transcription start site. Motifs of Sp-1 and the glucocorticoid receptor (GR) are indicated. Arrows show the inverted repeats. Closed boxes and a dashed box show the sequences with sequence identity to the promoters of type-II collagen gene and of link protein gene respectively.

the corresponding genomic segment revealed the presence of a 94 bp sequence with 73.4% identity to human EGF1 at 797–890 bp downstream from exon 12 (data not shown). However, this sequence is not flanked by potential donor and acceptor sequences for splicing. We also could not detect any PCR product with mouse genomic DNA using the primers from the human EGF1 sequence. It is possible that there is no EGF1 sequence in mouse aggrecan. EGF2 is found in several species such as human, bovine, dog, rat and mouse with highly conserved nucleotide sequences. Chick also has an almost identical sequence in the intron preceding the G3 coding exons [18]. According to previous reports [11,18], the exon for EGF2 should be located at the region approximately 2.5–3 kb upstream of the exon for the lectin-like domain. Although we did not obtain a clone containing this domain, sequencing of the PCR product using mouse genomic DNA as a template confirmed that EGF2 is coded for by exon 13.

The other two motifs (lectin-like and CRP-like motifs) are also well conserved among species. The nucleotide sequence of mouse shows 94.6%, 88.6% and 78.1% identity to rat, human and chicken respectively. The amino acid sequence identities of these motifs are 98.1%, 88.3% and 78.9% to rat, human and chicken respectively. The lectin-like subdomain is encoded by exons 14, 15 and 16, and the CRP-like subdomain by exons 17 and 18. Introns of this region are relatively short and five exons are packed within a 3.6 kb region. A similar gene structure for this domain is observed in chicken. The intron–exon boundaries are also conserved between the two species. The lectin-like domain is most similar to C-type carbohydrate-recognition domains (CRDs). As suggested [35,36], the lectin-like subdomain of aggrecan can be categorized into groups encoded by three separate exons, such as the asialoglycoprotein receptor, CD23 and the Kupffer cell receptor. Versican and neurocan also contain similar structural motifs at the C-termini.

Transcription start site and promoter sequence

Primer extension revealed four transcription start sites which are clustered within 70 bp (Figure 4). The positions of the transcription start are different from those of the rat aggrecan gene, although the gene structure is quite similar (Y. Yamada, unpublished work). Numbering the most upstream site as +1, the G+C content within the 523 bp upstream sequence exceeds 65%, similar to the promoter region of both the rat link protein and type-II collagen genes. Within the first 523 bp of the upstream sequence (Figure 4b), there was no TATAA sequence. Two glucocorticoid receptor-binding sequences (TGTTCT/C) and one GGGCGG sequence (Sp-1 site) are located at −517 to −512, −380 to −375, and −426 to −421 respectively. There are several homologous direct repeat sequences including sequences between −317 to −308 and −274 to −285, and between −226 to −212 and −178 to −202. In addition, a region between −54 and −111 shows sequence identity to a sequence of the promoter (−103 to −132) of the rat type-II collagen gene. This sequence is highly conserved in both the rat and mouse type-II collagen genes [37] and is known to be important for type-II collagen gene promoter activity (Y. Yamada, unpublished work). Another stretch of sequence from −287 to −259 shows sequence similarity to a sequence of the promoter (−82 to −60) of rat link protein. These sequences may play a role in cartilage-specific gene expression.

Elucidation of the aggrecan gene structure will be useful in the study of the transcriptional regulation of the gene in development and pathological states. In addition, it will facilitate a better understanding of the function of aggrecan *in vivo*. Recently we have identified a 7 bp deletion in exon 5 of the aggrecan gene of *cmd* mice [19]. The mutation occurs at the B subdomain of the G1 domain and leads to a termination codon within exon 6, resulting in a potentially truncated polypeptide of 36000-M_r. Although heterozygous *cmd* mice show a normal phenotype, the homozygous mice die soon after birth due to respiratory failure. It would be interesting to examine whether *cmd* mice could be rescued by introducing the normal aggrecan gene. It is also conceivable to study the function of an individual domain of aggrecan using transgenic mice models in normal and/or *cmd* mice.

We thank Dr. Hynda Kleinman and Dr. Paul Krebsbach for the critical reading of the manuscript. This work was partially supported by Seikagaku Corporation Grant.

REFERENCES

1 Hascall, V. C., Heinegard, D. K. and Weight, T. N. (1991) in Cell Biology of Extracellular Matrix (Hay, E. D., ed.), pp. 149–175, Plenum Press, New York and London
2 Hardingham, T. E. and Fosang, A. J. (1992) FASEB J. **6**, 861–870
3 Kimata, K., Barrach, H.-J., Brown, K. S. and Pennypacker, J. P. (1981) J. Biol. Chem. **256**, 6961–6968
4 Li, H., Schwalz, N. B. and Vertel, B. M. (1993) J. Biol. Chem. **268**, 23504–23511
5 Goetinck, P. F., Stirpe, N. C., Tsonis, P. A. and Carlone, D. (1987) J. Cell Biol. **105**, 2403–2408

6 Hascall, V. and Kimura, J. (1982) Methods Enzymol. **82**, 769–800
7 Doege, K., Sasaki, M., Horigan, E., Hassell, J. R. and Yamada, Y. (1987) J. Biol. Chem. **262**, 17757–17767
8 Doege, K. J., Sasaki, M., Kimura, T. and Yamada, Y. (1991) J. Biol. Chem. **266**, 894–902
9 Krueger, R. C., Fields, T. A., Mensch, J. R., Jr. and Schwartz, N. B. (1990) J. Biol. Chem. **265**, 12088–12097
10 Chandrasekaran, L. and Tanzer, M. L. (1992) Biochem. J. **288**, 903–910
11 Fulop, C., Walcz, E., Valyon, M. and Glant, T. T. (1993) J. Biol. Chem. **268**, 17377–17387
12 Doege, K., Rhodes, C., Sasaki, M., Hassell, J. R. and Yamada, Y. (1990) in Extracellular Matrix Genes: Molecular Biology of Cartilage Proteoglycan (Aggrecan) and Link Protein (Sandel, L. J. and Boyd, C. D., eds.), pp. 137–155, Academic Press Inc., San Diego
13 Zimmermann, D. R. and Ruoslahti, E. (1989) EMBO J. **8**, 2975–2981
14 Iozzo, R. V., Naso, M. F., Cannizzaro, L. A., Wasmuth, J. J. and McPhrerson, J. D. (1992) Genomics **14**, 845–851
15 Rauch, U., Karthikeyan, L., Maurel, P., Margolis, R. U. and Margolis, R. K. (1992) J. Biol. Chem. **267**, 19536–19547
16 Iwata, M., Wight, T. N. and Carlson, S. S. (1993) J. Biol. Chem. **268**, 15061–15069
17 Doege, K. and Yamada, Y. (1988) Collagen Relat. Res. **8**, 486
18 Tanaka, T., Har-El, R. and Tanzer, M. L. (1988) J. Biol. Chem. **263**, 15831–15835
19 Watanabe, H., Kimata, K., Line, S., Strong, D., Gao, L., Kozak, C. A. and Yamada, Y. (1994) Nature Genetics **7**, 154–157
20 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn., vol. 1, pp. 2.73–2.79, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
21 Chomczynski, P. and Sacchi, N. (1987) Anal. Biochem. **162**, 156–159
22 Mount, S. M. (1982) Nucleic Acids Res. **10**, 459–472
23 Rhodes, C., Savagner, P., Line, S., Sasaki, M., Chirigos, M., Doege, K. and Yamada, Y. (1991) Nucleic Acids Res. **19**, 1933–1939
24 Stamenkovic, I., Amiot, M., Pesando, J. M. and Seed, B. (1989) Cell **56**, 1057–1062
25 Fosang, A. J., Neame, P. J., Last, K., Hardingham, T. E., Murphy, G. and Hamilton, J. A. (1992) J. Biol. Chem. **267**, 19470–19474
26 Fosang, A. J., Neame, P. J., Hardingham, T. E., Murphy, G. and Hamilton, J. A. (1991) J. Biol. Chem. **266**, 15579–15582
27 Fosang, A. J., Last, K., Knauoer, V., Neame, P. J., Murphy, G., Hardingham, T. E., Tschesche, H. and Hamilton, J. A. (1993) Biochem. J. **295**, 273–276
28 Sandy, J. D., Neame, P. J., Boynton, R. E. and Flannery, C. R. (1990) J. Biol. Chem. **266**, 8683–8685
29 Mok, M. T., Ilic, M. Z., Handley, C. J. and Robinson, H. C. (1992) Arch. Biochem. Biophys. **292**, 442–447
30 Barry, F. P., Gaw, J. U., Young, C. N. and Neame, P. J. (1992) Biochem. J. **286**, 761–769
31 Antonsson, P., Heinegard, D. and Oldberg, A. (1989) J. Biol. Chem. **264**, 16170–16173
32 Bourdon, M. A., Krusius, T., Campbell, S., Schwartz, N. B. and Ruoslahti, E. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 3194–3198
33 Krueger, R. C., Fields, T. A., Hildreth IV, J., and Schwartz, N. B. (1990) J. Biol. Chem. **265**, 12075–12087
34 Baldwin, C. T., Reginato, A. M. and Prockop, D. J. (1989) J. Biol. Chem. **264**, 15747–15750
35 Hoyle, G. W. and Hill, R. L. (1991) J. Biol. Chem. **266**, 1850–1857
36 Bezouska, K., Crichlow, G. V., Rose, J. M., Taylor, M. E. and Drickamer, K. (1991) J. Biol. Chem. **266**, 11604–11609
37 Metsaranta, M., Toman, D., de Crombrugghe, B. and Vuorio, E. (1991) J. Biol. Chem. **266**, 16862–16869

Received 3 August 1994/3 January 1995; accepted 11 January 1995