The Structure of the Rat Aggrecan Gene and Preliminary Characterization of Its Promoter*

(Received for publication, April 26, 1994, and in revised form, August 29, 1994)

Kurt J. Doege‡§§, Katherine Garrison‡§, Silvija N. Coulter†, and Yoshi Yamada**

From the ‡Research Unit, Shriners Hospitals for Crippled Children and the §Departments of Biochemistry and Molecular Biology, Cell Biology and Anatomy, Oregon Health Sciences University, Portland, Oregon 97201 and the **Laboratory of Developmental Biology, NIDR, National Institutes of Health, Bethesda, Maryland 20892

Aggrecan is a major structural component of cartilage extracellular matrix and a specific gene product of differentiated chondrocytes. cDNA clones have been used to isolate rat aggrecan genomic clones from phage and cosmid libraries, producing over 80 kilobases (kb) of overlapping DNA containing the complete rat aggrecan gene, including 12 kb of 5'- and 8 kb of 3'-flanking DNA. DNA sequencing shows 18 exons, most of which encode structural or functional modules; exceptions are domains G1-B and G2-B, which are split into two exons and the G3 lectin domain, which is encoded by three exons. There is one expressed epidermal growth factor-like exon and in addition a non-expressed "pseudo-exon" encoding a heavily mutated epidermal growth factor-like domain. Intron sizes have been determined by restriction mapping and inter-exon polymerase chain reaction; a 30-kb intron separates exons 1 and 2. Exon 1 has been mapped by primer extension and S1 nuclease protection; it encodes 381 base pairs (bp) of 5'-untranslated sequence. There is a minor promoter which initiates transcription an additional 68 bp 5' of the major promoter start site. DNA sequence is reported for a 529-bp fragment encompassing exon 1, including 120 bp of 5'-flanking DNA comprising the promoter. This promoter is lacking the TATAA or CCAAT elements but has several putative binding sites for transcription factors. A 922-bp DNA fragment with 640-bp 5'-flanking DNA and 282-bp exon 1 sequence showed higher promoter activity in transfected chondrocytes than in fibroblasts, is completely inactive in the reverse orientation, and is strongly enhancer in the forward direction by the SV40 enhancer.

Aggrecan, the large aggregating proteoglycan of cartilage, is a major component of the wet weight of cartilage and confers many of the physical properties of cartilage essential to its function (1). Aggrecan is a marker for cartilage and is expressed abundantly only in this tissue, but has been reported at low levels in other tissues (2 and references therein). More recently, monoclonal antibodies have been used to distinguish aggrecan from other aggregating proteoglycans in a tissue distribution study, and some non-cartilage tissues in chicken were found to be immunologically reactive with anti-aggrecan antibodies (3). Several cDNA sequences have been reported for aggrecans from different species: rat (4, 5), chicken (6, 7, 29), bovine (8, 9), human (10, 11), and mouse (12). Application of these probes to blotting and in situ hybridization studies detects expression of aggrecan in developing embryos only in cartilaginous tissues or notochord (13). A recent study using the polymerase chain reaction has demonstrated low levels of aggrecan mRNA in chick embryo calvarium (14), but the expression of aggrecan is more restricted to cartilage than that of type II collagen, for example (15). Steady state levels of mRNA for type II collagen in chondrocytes are at least 10-fold higher than for aggrecan or link protein (16), and the tissue distribution studies may reflect a sensitivity threshold. To address questions of gene regulation in chondrocyte differentiation, it will be important to determine quantitatively the transcriptional levels of the various chondrocyte marker genes in different tissues during development. It will also be important to examine the role that alternative splicing may play in distinguishing different stages of chondrocyte differentiation, as seen with type II collagen (17) and potentially aggrecan (18).

Aggrecan is the first described member of a gene family of large aggregating proteoglycans, which also includes versican (19), neurocan (20), and brevican (21). Versican (PG-M) has a wider tissue distribution than aggrecan and is highly expressed in chondroprogenitor cells, but decreases in abundance with chondrogenesis, at least in chickens (3, 22); neurocan and brevican appear restricted to neural tissues. These family members share many structural features including an amino-terminal domain, G1, which is related to link protein and binds to hyaluronan and link protein through subdomains A, B, and B'. Proteoglycans of this group also possess a COOH-terminal domain (G3 in aggrecan) composed of a C-type lectin motif (LEC) (23), a complement regulatory protein-related motif (CRP) (24), and one or two repeats of epidermal growth factor-like motifs (EGF) (25). This particular group of motifs is shared with the selectins (26), which are cell adhesion proteins. The G1 and G3-like domains in most of these proteoglycans are separated by long, structurally extended sequences, apparently unrelated, which provide attachment points for glycosaminoglycans and other carbohydrate chains; these domains are called keratan sulfate and chondroitin sulfate in aggrecan. Aggrecan is the only family member with a third globular domain, G2, which contains additional link-related motifs B and B'; G2 is adjacent to G1, separated by an extended region called interglobular domain.

*This work was supported by a grant from Shriners of North America (to K. J. D.). 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 (1) and EMBL Data Bank with accession number(s) J03485 and U13974.

†To whom correspondence should be addressed: Research Unit, Shriners Hospital for Crippled Children, 3101 S. W. Sam Jackson Park Rd., Portland, OR 97201. Fax: 503-221-3451.

‡Present address: Biology Dept., Bucknell University, Lewisburg, PA 17837.

1 The abbreviations used are: LEC, lectin; EGF, epidermal growth factor; CRP, complement regulatory protein; PCR, polymerase chain reaction; CAT, chloramphenicol acetyltransferase; bp, base pair(s); kb, kilobase(s); PIPES, 1,4-piperazinediethanesulfonic acid; nt, nucleotide(s).
The aggrecan gene has been mapped to human chromosome 15q26.1 (27) and to mouse chromosome 7 (12). While aggrecan has not been linked to classical familiar forms or other heritable disorders of the skeletal system (28), there are a number of animal mutants which are apparently defective in this gene, including the nanomelic chicken (29), and the cartilage-matrix-deficient mouse (cmd) (30). The altered expression of aggrecan seems to be a hallmark of osteoarthritis and other degenerative diseases of cartilage. An understanding of the structure and regulation of the aggrecan gene will enhance our ability to identify mutations and altered expression of this gene and develop therapies for debilitating conditions of joint tissues and abnormal skeletal growth.

We report here the first complete description of the gene structure for a member of this proteoglycan family, along with some of the preliminary rat and human aggrecan exon structure and regulation of expression. This work provides a starting point for pursuing an understanding of the processes by which aggrecan expression is regulated.

EXPERIMENTAL PROCEDURES

Isolation of Genomic Clones—Two A phage genomic libraries were screened using rat aggrecan cDNAs as probes. A partial EcoRI rat genomic library in Charon 4A (gift of Dr. Tom Sargent, NICHD, NIH) was screened using a mixture of labeled cDNAs encoding the entire aggrecan structural sequence: inserts from clones pRPG1-5 (5); 4 x 10

DNA Sequencing—Genomic fragments from phage clones were subcloned into pUC18 and sequenced directly from CsCl density gradient-purified single-stranded DNA sequences. Reaction products were visualized and quantitated with addition of 75 pM of ice-cold 4 mM ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 pM of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

PCR Analysis—PCR amplification of DNA or reverse-transcribed RNA templates was accomplished using standard methods as described previously (31). Annealing temperatures were calculated by the PRIMER program (kindly provided by Drs. S. Lincoln, M. Daly, and E. Lander, Massachusetts Institute of Technology Center for Genome Research). For amplification of the EGF2 domain, primers which flank the putative insertion site of the EGF exon in the cDNA were used: bp 5549-5569 and 6602-6622 (numbered including the EGF2 exon sequence). EcoRI recognition sites and spacer sequences were added at the 5'-ends for ease of cloning.

Promoter-Reporter Plasmid Constructs—The series of CAT vectors available from Promega was used. The initial promoter constructs, p604CAT+ and p604CAT− are the forward and reverse orientations, respectively, of a 922-bp SalI fragment from a genomic clone EcoRI BamHI fragment encompassing exon 1. The EcoRI-PvuI fragment was rendered blunt-ended by T4 DNA polymerase and ligated to HindIII linkers for cloning into the HindIII site of the Basic-CAT vector. This fragment was then cloned in both orientations into the CAT vector with the SV40 enhancer in the downstream BamHI site to give p604CAT-En+ and p604CAT-En−.

Transfections—Primary chick embryo sternal chondrocytes or tendon fibroblasts were obtained from day 16 embryos of white leghorn chickens; cells were liberated from tissue by treatment with 0.3% collagenase and 0.025 M ammonium chloride, pH 7.8, lysed by three successive cycles of freeze-thaw, and centrifuged at 14,000 x g, 4 °C for 30 min. The extract was quick-frozen and stored at −80 °C.

β-Galactosidase assay was performed as described (34). CAT activity was assayed by liquid scintillation chromatography as described (34) using volumes of extract normalized to equivalent β-galactosidase activities. Reaction products were visualized and quantitated described previously (4), using the acid guanidinium/phenol method (38). Poly(A)+ mRNA was prepared by oligo(dT) chromatography (34). Oligonucleotide primers were end-labeled using γ32P-ATP and polynucleotide kinase to a specific activity of approximately 2 x 106 counts/min/μg. 2 x 105 counts/min of primer was used with 10 μg of poly(A)+ RNA in a volume of 25 μl containing 5 μM sodium phosphate, pH 6.8, and 200 μM EDTA. The samples were annealed at 68 °C and then slowly cooled to 4 °C. The sample was then diluted to 50 μl containing 56 mM potassium chloride, 50 mM Tris-HCl, pH 8.3, 5 mM diethiothreitol, 7.5 mM MgCl2, 1 μM each of the four deoxynucleotide triphosphates, 1 μM of RNA (Promega), and 400 units of T7 RNA polymerase and reverse transcriptase (Promega). The reaction was incubated for 2 h at 42 °C, heat-inactivated at 90 °C, phenol/chloroform extracted, and the nucleic acids precipitated with 66% ethanol, 0.3 M sodium acetate, 10 μg of carrier RNA. The products were dissolved in 80% formamide, 0.1% each xylene cyanol and bromphenol blue tracking dyes, heated at 80 °C, and loaded on a 7.5% polyacrylamide sequencing gel for sizing relative to a sequence reaction from the same region of DNA, or in some cases end-labeled restriction fragments.

5'-Nucleotide Protection—The method described (34) for double-stranded DNA probes was followed. End-labeled fragment (10 μg/μl) was co precipitated with 10 μg of either poly(A)+ RNA from rat chondrosarcoma or yeast tRNA. The pelleted nucleic acids were dissolved in 80% formamide, 0.4 M PIPES, pH 6.4, 0.4 M NaCl, 1 mM EDTA, denatured for 10 min at 90 °C, and annealed for 3 h at 52-62 °C, according to GC content (34). The DNA-RNA hybrids were subjected to 5% polyacrylamide gel digestion in 300 μl of 0.05 x ammonium acetate, 0.25 x NaCl, 0.01 μM zinc chloride containing 6.25 μg of denatured salmon sperm DNA; between 10 and 90 units of S1 nuclease was added, and samples were incubated 30 min at 37 °C. Reactions were stopped with addition of 75 μl of ice-cold 4 x ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 μg of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

Primer Extension Analysis—RNA was prepared from Swarm rat chondrosarcoma cells released from matrix by collagenase digestion as described previously (5). Primer extension products were visualized and quantitated with addition of 75 μl of ice-cold 4 x ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 μg of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

Nucleic acid hybridization was performed (31). After hybridization, filters were washed three times with 2 x standard saline citrate (SSC) at 60 °C, and two times with 0.2 x SSC, 0.1% SDS at 60 °C.

Primer extension products were visualized and quantitated with addition of 75 μl of ice-cold 4 x ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 μg of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

Proteins were prepared (31). After hybridization, filters were washed three times with 2 x standard saline citrate (SSC) at 60 °C, and two times with 0.2 x SSC, 0.1% SDS at 60 °C.

Primer extension products were visualized and quantitated with addition of 75 μl of ice-cold 4 x ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 μg of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

Biochemistry and Genetics (31). After hybridization, filters were washed three times with 2 x standard saline citrate (SSC) at 60 °C, and two times with 0.2 x SSC, 0.1% SDS at 60 °C.

Primer extension products were visualized and quantitated with addition of 75 μl of ice-cold 4 x ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 μg of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

Biochemistry and Genetics (31). After hybridization, filters were washed three times with 2 x standard saline citrate (SSC) at 60 °C, and two times with 0.2 x SSC, 0.1% SDS at 60 °C.

Primer extension products were visualized and quantitated with addition of 75 μl of ice-cold 4 x ammonium acetate, pH 4.5, 0.1 M EDTA, and 25 μg of RNA, extracted with phenol/chloroform (1:1), and the nucleic acids precipitated with isopropyl alcohol for analysis on denaturing polyacrylamide gel electrophoresis as described for the primer extension products.

Biochemistry and Genetics (31). After hybridization, filters were washed three times with 2 x standard saline citrate (SSC) at 60 °C, and two times with 0.2 x SSC, 0.1% SDS at 60 °C.
in Fig. 1. These two groups of clones from different libraries were constructed by Southern hybridization with BamHI, HindIII, KpnI, SmaI, and XbaI. The exon boundaries yield three different clones encoding all but the signal sequence and 5′-untranslated region of the aggrecan structural sequence. Three additional clones were obtained, shown as A5, h4, and A7 in Fig. 1. These two groups of clones from different libraries encoded all the exons of the rat aggrecan gene but did not overlap in the first intron. To clone this missing portion, a cosmid library was screened using cDNA probes for exons 1-4; the resulting 42-kb clone is shown as Cos77 in Fig. 1.

Analysis of Genomic Clones—EcoRI fragments of the genomic clones were subcloned into pUC18, and rough maps of the coding regions were constructed by Southern hybridization analysis following cleavage with the restriction enzymes BamHI, HindIII, KpnI, SmaI, and XbaI. The exon boundaries were precisely determined by sequencing subcloned genomic fragments with exon-specific primers, and intron sizes were mapped by a combination of restriction blot analysis with exon-specific oligonucleotide probes, sizing of inter-exon PCR products, and sequencing. These data are summarized graphically in Fig. 1 and in more detail in Fig. 2. The exons are distributed within a 65 kb length, much of which is taken up by a 30-kb intron separating exons 1 and 2. There are 18 exons, most of which correspond to structural domains of the molecule; notable is the extremely large (3.5 kb) exon encoding the entire chondroitin-sulfate attachment region of the protein. The 5′-leader sequence, the signal peptide, G1-A, G1-B, G2-B′, the interglobular domain, keratan sulfate, G3-EGF, and G3-CRP domains are also encoded by single exons. Exceptions are the G1-B, G2-B, and G3-LEC domains, which are encoded as two, two, and three exons, respectively. The 3′-terminal exon 18 is composed of a short coding sequence, the termination codon, and, presumably, the entire untranslated trailer sequence. In addition, there appears to be a "pseudo-exon" encoding an additional EGF-like domain as discussed below; this is designated as X in Figs. 1 and 2. The internal introns are all of the symmetrical phase 1 type (39), except for those which split the structural domains G1-B (intron 4), G2-B (intron 8), and G3-LEC (intron 15); these are phase 2-2, 2-2, and 0-0, respectively.

During the course of sequencing these exons, an error was discovered in the reported cDNA sequence for exon 6 (G1-B′) (5); at 1210–1217 should be corrected from GGAGGCTGG to GGAGGCTCG. The corresponding amino acid sequence changes are from (CRTVG) to (CRRLG). The rat amino acid sequence now is identical to the corresponding human sequence (11). This error has been corrected in the most recent GenBank entry.

Cloning and Analysis of EGF-like Repeat Exons—A reverse transcriptase PCR product from rat chondrocyte and chondrosarcoma mRNA was previously identified (11) which was thought to be derived from a minor subset of agranlic transcript containing a single EGF-like repeat; this identification has been confirmed by cloning and sequencing of the PCR product (not shown). Surprisingly, this rat aggrecan EGF sequence (40) was more similar to the EGF2 motif in human versican (19) than to the first reported EGF motif in human aggrecan (10) (Fig. 4). The location of this rat EGF-like exon was shown to be 3 kb upstream of the first lectin-like exon of G3 (Figs. 1 and 2) by Southern blotting and by PCR amplification between this exon and the ends of the genomic phage clones A1 and A31. A second EGF-like sequence was detected in the rat aggrecan gene by low stringency Southern blot hybridizations, using an oligonucleotide probe from the human aggrecan EGF1-like sequence. The positive 1-kb KpnI-BamHI fragment of A1 was cloned and sequenced and found to harbor an EGF1-like motif (Fig. 3) about 300 bp downstream from the KpnI site. The deduced translation product of this new sequence is compared by the multiple sequence alignment program PILEUP (37) to other EGF sequences (Fig. 4) including a recently described second expressed EGF motif (EFG2) from human aggrecan (40). The rat aggrecan EGF1 sequence shown in Figs. 3 and 4 has undergone loss of 3 of the 6 cysteine residues conserved throughout this motif family, which is 5 amino acid portion deleted, and has also mutated invariant proline and tyrosine residues. These structural differences make it unlikely that the sequence motif can retain functionality in the protein, and indeed the 5′-intron-exon junction is lacking an appropriate splice acceptor sequence. Furthermore, this has not been possible to detect expression of this exon in rat RNA by specifically primed reverse transcriptase PCR (not shown). As this rat exon is located in an homologous position to the expressed EGF1 exon in human aggrecan (40), it appears to be a lost-function...
Fig. 2. Summary of exon and intron size and location in the rat aggrecan gene. Exon number is shown with corresponding domain (abbreviations discussed in text). The exon size in base pairs and amino acid residues is given along the sequence surrounding the intron boundary (exon sequence in upper case, intron in lower case) and the nucleotide-numbered position of the exon in the complete cDNA sequence. The intron size in base pairs is presented, as measured by inter-exon PCR (all but 1 and 13 which were estimated by restriction mapping). Exon 18 extends indefinitely as 3’-untranslated trailer sequence; the gap shown in the upper case sequence is at the end of the published cDNA sequence (5).

![Genomic DNA sequence encoding a second, unexpressed EGF-like motif in the rat aggrecan gene.](image)

*Fig. 3. Genomic DNA sequence encoding a second, unexpressed EGF-like motif in the rat aggrecan gene.* A rat genomic fragment which hybridized to an oligonucleotide sequence from the human aggrecan EGF1 sequence (10) was cloned and sequenced as described under “Experimental Procedures.” The translated open reading frame shown in single letter code matched well with the human EGF1 amino acid sequence. A stop codon limits the 5’-end of the coding sequence (asterisk), while a consensus splice-donor sequence (boxed) appears at the 3’-end of the homologous sequence. Two potential 5’-splice acceptor sequences are also boxed, but they lie outside the region of homology and upstream of the stop codon.

pseudo-exon in rat; it has accordingly been designated X in Figs. 1 and 2.

### Intron Analysis

The exon positions were mapped initially by Southern blot hybridization using exon-specific probes, and most of the intron sizes were then confirmed by inter-exon PCR; the exceptions were introns 1, 2, and 13 which at 30, 5, and 3.5 kb were too large (as the ragEGF1X pseudo-exon is not expressed, it is considered to be contained within intron 12). A portion of intron 13 was amplified from total genomic DNA, using an upstream primer from exon 13 (EGF2) and a downstream primer from the 5’-end of A31, to give a product of 1.1 kb (not shown). This product size was consistent with the non-overlapping phage clones A1 and A31 being contiguous, with a common terminal EcoRI site, as shown in Fig. 1.

The size of intron 1 was determined from restriction mapping of a single cosmid clone which overlapped the phage clones A5 and A4. The size of the non-overlapping “gap” region between the phage clones was 14 kb in the cosmid clone, and the authenticity of this non-overlapping portion of the cosmid clone was supported by Southern blot analysis of total rat genomic DNA, using two different probes within the central part of the

<table>
<thead>
<tr>
<th>EXON</th>
<th>DOMAIN</th>
<th>EXON SIZE (BP/AA)</th>
<th>ACCEPTOR SITE</th>
<th>EXON SEQUENCE</th>
<th>DONOR SITE</th>
<th>EXON SIZE IN CDNA</th>
<th>INTRON SIZE (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5’-UT</td>
<td>381/-</td>
<td>CCACT</td>
<td>TCCAG</td>
<td>gtaaa</td>
<td>1-381</td>
<td>30.0</td>
</tr>
<tr>
<td>2</td>
<td>SP</td>
<td>76/23</td>
<td>tcacag</td>
<td>TGAACCT AGC</td>
<td>gtaaa</td>
<td>1-381</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>G1-A</td>
<td>384/129</td>
<td>cacag</td>
<td>AC CAT D R</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.8</td>
</tr>
<tr>
<td>4</td>
<td>G1-B</td>
<td>174/58</td>
<td>tacag</td>
<td>GC ATT G I</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>5</td>
<td>G1-B’</td>
<td>127/42</td>
<td>aacag</td>
<td>A TAC Y</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>6</td>
<td>G1-B’</td>
<td>294/98</td>
<td>cacag</td>
<td>GT GAG G E</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>7</td>
<td>G2-B</td>
<td>175/58</td>
<td>cacag</td>
<td>GA GTT G Y</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>8</td>
<td>G2-B</td>
<td>128/42</td>
<td>cacag</td>
<td>A TAC Y</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>9</td>
<td>G2-B’</td>
<td>294/98</td>
<td>cacag</td>
<td>GG GAA G E</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>11</td>
<td>KS</td>
<td>204/68</td>
<td>ttcag</td>
<td>GT GTG G V</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>3471/1157</td>
<td>tacag</td>
<td>GTG AGC G I</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>13</td>
<td>EXON 3-</td>
<td>99/33</td>
<td>cgc cta</td>
<td>TGT GAT C D</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>14</td>
<td>G3-EGF1</td>
<td>114/38</td>
<td>tgcag</td>
<td>AC AGT D I</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>15</td>
<td>G3-EGF2</td>
<td>159/53</td>
<td>tgcag</td>
<td>AC CAG D Q</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>16</td>
<td>G3-LEC</td>
<td>81/27</td>
<td>tacag</td>
<td>AA AT K N</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>17</td>
<td>G3-LEC</td>
<td>145/49</td>
<td>cacag</td>
<td>CAA TTT Q F</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
<tr>
<td>18</td>
<td>G3-3’</td>
<td>139+/24</td>
<td>cgccag</td>
<td>CC A AC F N</td>
<td>gtaaa</td>
<td>1-381</td>
<td>1.4</td>
</tr>
</tbody>
</table>

![Multiple sequence alignment of G3 EGF motifs.](image)
14-kb gap region (data not shown). The close agreement of the restriction map of COS77 in the portions which overlap with phage clones and the corroborated structure of the 14-kb cosmid non-overlapping region in total genomic DNA indicates that no major rearrangement has occurred in this portion of the cosmid clone. One anomaly was seen in clone λ7, which is apparently missing an EcoRI restriction site found in both COS77 and at the end of A11. It is not known whether this difference may represent a real polymorphism or instead is some type of artifact.

Determination of Exon 1 Boundary—Several cDNA clones for rat aggrecan were isolated which extended 5′ of the translation start site, and these clones all contained varying lengths of the same untranslated leader sequence (5).2 Genomic clones were isolated with probes to this leader sequence, which was found to reside in a single exon. A 1.8-kb EcoRI-BamHI genomic fragment from the 5′ genomic clone which contained the 5′-leader exon was sequenced, and a 532-bp portion encompassing the exon is shown as Fig. 5. The 5′-end of this exon was mapped by restriction map of COS77 in the portions which overlap with phage clones and the corroboration of the structure of the 14-kb gap region (data not shown). The close agreement of the restriction start site mapped in Fig. 5, although these larger products were not sized as precisely. Likewise, the S1 protection analysis using a cDNA rather than genomic probe fragment showed complete protection of exons 1–3 by chondrosarcoma RNA, but not by tRNA, indicating there are no other major promoters which exclude exon 1 (Fig. 7).

Examination of the sequence surrounding the putative transcription start site (Fig. 5) showed neither an apparent TATAAA sequence, nor a CCAAT box. There were, however, several GC-boxes (5′ splice sites) typical of such TATAA-less promoters (41). Other potential transcription factor binding sites were identified by database search (TFD, kindly provided by Dr. David Ghosh, NCBI, NIH) including several AP-2 sites (42) and a site for NFκ-b (43) at -90 bp, which interestingly is conserved in the same position in the rat Col2al promoter sequence (44).

Analysis of Promoter Function—Several constructs were prepared placing DNA fragments from the transcriptional start region into the pCAT family of reporter vectors (Promega), as summarized in Fig. 8. The initial fragment which was tested, P640, had 640 bp of 5′-flanking DNA and 282 bp of the first exon, and this was cloned in forward (+) and reverse (−) orientation in the pCAT-Basic vector. As shown in Table I, the P640CAT construct reproducibly drove higher levels of CAT expression than the (−) version when transfected into freshly isolated chick embryo sternal chondrocytes. This level of expression is roughly 25% of that obtained using the SV40 promoter and enhancer to drive the CAT gene (pCAT-Control). When the SV40 enhancer was included in the construct with the forward-oriented P640 fragment (p640CAT-EN+), the CAT expression was almost 8-fold higher than the pCAT-Control vector. The reverse-orientation promoter in the same enhancer construct (p640CAT-EN−) gave background levels of CAT activity. Expression levels for both the enhanced and unenhanced promoters were substantially lower in chick embryo tendon fibroblast primary cells than in chondrocytes, relative to the pCAT-Control plasmid in both cell types. Larger fragments
Rat chondrosarcoma. The primer used is shown in Fig. 5 and begins at primer (order ACGT).

The genomic fragment was an end-labeled marker is exon 12 template DNA sequenced with a primer beginning primer extension mapping of the end of aggrecan RNA transcripts from matically, with numbers referring to the start of exon 1.

+84; extension products are in trans-fected chick sternal chondrocytes (data not shown).

have been tested for promoter activity in this system; the largest extends an additional 3.2 kb upstream of p640 but produces similar levels of CAT activity as the shorter construct in trans-fected chick sternal chondrocytes (data not shown).

DISCUSSION

This paper reports the first complete description of a gene from the aggrecan-like family of proteoglycans. The rat aggrecan gene has been cloned from both phage and cosmid libraries, yielding over 80 kb of overlapping genomic clones. The coding sequences are contained in 63 kb and are split into 18 exons, which are rather larger than usual (45) and tend to encode structural or functional domains, with some exceptions. The G1 and G2 B motifs are each split into two exons, unlike the gene for link protein, where these are single exons (46). The Ba and Bb division occurs between sequences coding for 2 cysteines known to be disulfide linked in the protein, and this is a clear example of a structural motif split between exons. The G1-A and the G1- and G2-B' motifs are not split but are in single exons, as seen for the link protein gene. Another similarity to the link protein gene is the size of the first intron, 30 kb or larger for both genes (47). The gene encoding CD44, which also contains a link B motif, follows the aggrecan pattern using two exons to encode this domain, with the division occurring in a similar position, splitting the pair of cysteines (48). The EGF and CRP motifs in G3 are encoded as structural units by single exons, as is seen in the selectins (49, 50). The LEC domain, however, is split into three exons, similarly to the rat asialo-glycoprotein receptor (51), but unlike the selectins, which encode the lectin domain as a single exon (49, 50).

An extreme example of the correlation of gene product structure-function with exon structure is seen in the nearly 3.5-kb exon 12, which encodes the entire chondroitin sulfate or glycosaminoglycan attachment region of aggrecan. This region was defined as containing all the serine-glycine sequences in the molecule, which occur in several different repeating patterns (5). This size of exon is quite unusual for an internal exon not containing untranslated sequence (45) and may be related to the extensive rearrangement and duplication thought to be involved in the generation of this array of repeats.

The introns show typical 5' -donor (g-t-a/w) and 3'-acceptor (c-a-g) splice sites (52), and for the most part are of the symmetrical phase 1 type; exceptions are those between the Ba and Bb exons, which are phase 2–2 in both G1 and G2. This phase
The set of three G3 lectin exons, where intron 15, between transfections within one experiment.

Activity, and CAT activity is expressed as percentage conversion of pBR322 sequences are shown as light and dark stippled boxes, respectively. The SV40 promoter is indicated by vertical stripes, and the SV40 enhancer is indicated by diagonal stripes.

**Table 1**

Promoter-driven CAT activity in transfected cultured cells

Chick sternal chondrocytes (CSC) or chick tendon fibroblasts (FIB) were cotransfected with the indicated constructs and pSVP as internal standard; extracts were prepared and assayed as described under "Experimental Procedures." Extracts were normalized for β-galactosidase activity, and CAT activity is expressed as percentage conversion of chloramphenicol to the acetylated form. Data are averages of triplicate transfections within one experiment.

<table>
<thead>
<tr>
<th></th>
<th>CSC</th>
<th>FIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>p640CAT+</td>
<td>26</td>
<td>6</td>
</tr>
<tr>
<td>p640CAT-</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>p640CAT-EN+</td>
<td>792</td>
<td>181</td>
</tr>
<tr>
<td>p640CAT-EN-</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>pCAT-Control</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

| p640CAT+/p640CAT- | 2.4 ± 0.4 |
| p640CAT+/pCAT-Control | 0.26 ± 0.05 |

- 3 Cat activities expressed relative to pCAT-Control activity in the same experiment.
- 4 p640CAT+ activity expressed relative to either p640CAT- or pCAT-Control in chick sternal chondrocytes; mean and standard error of these ratios for three different experiments are shown.

change requires the flanking exons to be spliced together or result in loss of the reading frame. A similar situation is found in different species (40).

The EGF-like motif expressed at low levels in rat aggrecan is of the EGF2 class. This same rat EGF2 sequence was recently reported as part of another study on alternative splicing of G3 in different species (40). An EGF1-encoding exon in the rat aggrecan gene has undergone extensive mutation and appears to be a non-functional relic (Figs. 3 and 4); we have referred to this as a pseudo-exon 13x. Two human aggrecan EGF exons were mapped by PCR to positions corresponding to those of the rat aggrecan exons 13x and 13 (40). Human aggrecan transcripts more frequently possess an EGF1-like sequence than the EGF2 motif and can express them together at low levels

The sequence in Fig. 5 shows this promoter to be of the TATA box class, and the sequencing of the upstream region of the rat aggrecan gene has undergone extensive mutation and appears to be a non-functional relic (Figs. 3 and 4); we have referred to this as a pseudo-exon 13x. Two human aggrecan EGF exons were mapped by PCR to positions corresponding to those of the rat aggrecan exons 13x and 13 (40). Human aggrecan transcripts more frequently possess an EGF1-like sequence than the EGF2 motif and can express them together at low levels

The sequence in Fig. 5 shows this promoter to be of the TATA box class, and the sequencing of the upstream region of the rat aggrecan gene has undergone extensive mutation and appears to be a non-functional relic (Figs. 3 and 4); we have referred to this as a pseudo-exon 13x. Two human aggrecan EGF exons were mapped by PCR to positions corresponding to those of the rat aggrecan exons 13x and 13 (40). Human aggrecan transcripts more frequently possess an EGF1-like sequence than the EGF2 motif and can express them together at low levels

The sequence in Fig. 5 shows this promoter to be of the TATA box class, and the sequencing of the upstream region of the rat aggrecan gene has undergone extensive mutation and appears to be a non-functional relic (Figs. 3 and 4); we have referred to this as a pseudo-exon 13x. Two human aggrecan EGF exons were mapped by PCR to positions corresponding to those of the rat aggrecan exons 13x and 13 (40). Human aggrecan transcripts more frequently possess an EGF1-like sequence than the EGF2 motif and can express them together at low levels

The sequence in Fig. 5 shows this promoter to be of the TATA box class, and the sequencing of the upstream region of the rat aggrecan gene has undergone extensive mutation and appears to be a non-functional relic (Figs. 3 and 4); we have referred to this as a pseudo-exon 13x. Two human aggrecan EGF exons were mapped by PCR to positions corresponding to those of the rat aggrecan exons 13x and 13 (40). Human aggrecan transcripts more frequently possess an EGF1-like sequence than the EGF2 motif and can express them together at low levels

3 H. Watanabe, L. Gao, S. Sugiyama, K. Doige, K. Kimata, and Y. Yamada, manuscript submitted for publication.
consistent with the 2–3 bp “range” detected for the start site of aggrecan transcripts. The 120 bp of 5′-flanking sequence is 72% G+C, and contains four potential AP-2 sites (42), two of which overlap with potential SP1 sites. The first exon sequence is also GC-rich at 66% and also contains a cluster of four overlapping potential AP-2 sites. The overlapped sequence at -105 to -64 is of particular interest, since it is the most conserved sequence (81%) in a comparison with the rat type II collagen (COL2A1) promoter (44), where it occurs in the same relative position (-123 to -103). This sequence is also a potential binding site in the aggrecan promoter (but not the COL2A1 promoter) for NF-kb, a factor known to interact with various cytokines (43), which may also play a role in chondrocyte gene regulation.

The link protein gene promoter (56) also has conserved sequences relative to the control in chondrocytes as does the unenhanced fragment by the heterologous SV40 enhancer; this suggests the potential AP-2 sites. The overlined sequence at -105 to -84 is of particular interest, since it is the most conserved sequence (81%) in a comparison with the rat type II collagen (COL2A1) promoter (44), where it occurs in the same relative position (-123 to -103). This sequence is also a potential binding site in the aggrecan promoter (but not the COL2A1 promoter) for NF-kb, a factor known to interact with various cytokines (43), which may also play a role in chondrocyte gene regulation.

Definitive evidence of such cell-specific elements in the aggrecan gene will depend on the demonstration of loss or gain of function in constructs with the elements present or absent. Furthermore, the regulation of endogenous aggrecan gene expression at the transcriptional level has not been characterized quantitatively, either in vivo or in different cell types in culture.

It is assumed that fibroblasts are transcribing the aggrecan gene at a much lower level than are primary chondrocytes, based on the steady state levels of aggrecan mRNA in these cells (4), but the relative contribution of new transcription and message stabilization to aggrecan steady state mRNA levels has not been evaluated for these cells. The chick primary chondrocyte system was used in the characterization of the promoters for the rat type II collagen and link protein genes (55, 56), making this the most appropriate system in which to initiate comparable studies on aggrecan. We are currently seeking to establish a cell culture system with a known degree of transcriptional activation of the endogenous aggrecan gene relative to other chondrocyte-expressed genes so that we may have a baseline to help identify aggrecan DNA sequences which confer the appropriate regulation in transfection studies.

REFERENCES

Rat Aggrecan Gene and Promoter