Structure of the human aggrecan gene: exon–intron organization and association with the protein domains

Wilmot B. VALHMU,*§ Glyn D. PALMER,* Pamela A. RIVERS,* Sohei EBARA,* Jan-Fang CHENG,† Stuart FISCHER‡ and Anthony RATCLIFFE*

*Orthopaedic Research Laboratory, Department of Orthopaedic Surgery and ‡Department of Biochemistry and Molecular Biophysics, Columbia University, New York NY 10032, U.S.A.; and †Lawrence Berkeley Laboratories, Berkeley, CA 94720, U.S.A.

The complete exon–intron organization of the human aggrecan gene has been defined, and the exon organization has been compared with the individual domains of the protein core. A yeast artificial chromosome containing the aggrecan gene was selected from the Centre d'Etude du Polymorphisme Humaine yeast artificial chromosome library. A cosmid sublibrary was created from this, and direct sequencing of individual cosmids was used to provide the exon–intron organization. The human aggrecan gene was found to be composed of 19 exons ranging in size from 77 to 4224 bp. Exon 1 is non-coding, whereas exons 2–19 code for a protein core of 2454 amino acids with a calculated mass of 254379 Da. Intron 1 of the gene is at least 13 kb. Overall, the sizes of the 18 introns range from 0.5 to greater than 13 kb. Each intron begins with a GT and ends with an AG, thus obeying the GT/AG rule of splice-junction sequences. The entire coding region is contained in 39.4 kb of the gene. The organization of exons is strongly related to the specific domains of the protein core. The A loop of G1 and the interglobular domain are encoded by exons 3 and 7 respectively. The B and B′ loops of G1 are encoded by exons 4–6, and those of G2 are encoded by exons 8–10. These sets of exons, coding for the B and B′ loops, are identical in size and organization. This is supported by the intron classes associated with these exons. Exon 11 codes for the 5′ half of the keratan sulphate-rich region, and exon 12 codes for the 3′ half of the keratan sulphate-rich region as well as the entire chondroitin sulphate-rich region. G3 is encoded by exons 13–18, including the alternatively spliced epidermal growth factor-like and complement regulatory protein-like domains. The correspondence between the exon organization and the protein domains argues strongly for modular assembly of the aggrecan gene.

INTRODUCTION

The proteoglycan aggrecan is the quantitatively major non-collagenous component of the extracellular matrix of cartilage, and it is also found in smaller but significant amounts in other connective tissues. Aggrecan consists of an extended multidomain protein core with a predicted size of approx. 220–250 kDa [1–3], to which many keratan sulphate (KS) and chondroitin sulphate (CS) glycosaminoglycan chains are attached [4,5]. At the N-terminus of the protein core is a globular domain (G1) [6], which non-covalently and specifically binds to hyaluronan. Each hyaluronan–aggrecan interaction is stabilized by a separate link protein, and the binding of many aggrecan molecules to a chain of hyaluronan leads to the formation of macromolecular aggregates which effectively become immobilized within the collagenous network of cartilage. G1 is separated from a second highly homologous globular domain (G2) by a linear interglobular domain (IGD) [7,8]. The protein core extends further into the KS- [9] and CS- [10] attachment regions. At the C-terminus are two alternatively spliced epidermal growth factor-like regions (EGF1 [1] and EGF2 [3]), a lectin (LEC)-like G3 domain [11], and an alternatively spliced domain which bears sequence similarity to complement-regulatory proteins (CRPs) [12]. In the extracellular matrix, some aggrecan molecules lack the G3 domain [13,14], perhaps because of proteolytic cleavage. Ultrastructural characterization of embryonic chick aggrecan core protein revealed that only 53% of the molecules contained the G3 domain [13].

The G1 domain has a characteristic three-looped structure. This includes an N-terminal A loop, which shows amino acid sequence similarity to the immunoglobulin superfamily, and two further loops (B and B′) which have homology with each other and are termed the proteoglycan tandem repeats (PTRs) [15]. G2 is composed of a second pair of PTRs and, despite the homology with the PTR motifs of the G1 domain, does not appear to bind to hyaluronan [7], and its function remains unknown. Link protein also has the A, B and B′ loops, replicating the structure of the G1 domain [15,16]. Other hyaluronan-binding proteins with similar structural domains include the human hyaluronan receptor CD44 [17,18], rat neurocan [19], human versican [20], bovine brevican [21], rat and cat BEHAB [22] and human TSG-6 [23], suggesting that this domain is derived from a common ancestral gene.

Complete coding sequences of human [2], rat [24] and chicken [25,26] aggrecan have been obtained, and partial structures of the chicken [27], rat and human [28] genes have previously been

Abbreviations used: G1, globular domain 1; G2, globular domain 2; G3, globular domain 3; IGD, interglobular domain; PTR, proteoglycan tandem repeat; KS, keratan sulphate; CS, chondroitin sulphate; LEC, lectin; EGF, epidermal growth factor; CRP, complement-regulatory protein; SP, signal peptide; YAC, yeast artificial chromosome; UTR, untranslated region; TTE, Tris/taurine/EDTA buffer; LAM-1, leucocyte adhesion molecule 1.

§ To whom correspondence should be addressed.

The nucleotide sequence data reported for the 5′ UTR, 3′ UTR and splice junction sequences of the 18 introns of the human aggrecan gene will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession numbers U13192, U13613, U22194, U22195, U22196, U22197, U22198, U22199, U22200, U22201, U22202, U22203, U22204, U22205, U22206, U22207, U22208, U22209, U22210, U22211, U22212, U22213, U22214, U22215, U22216, U22217, U22218, U22219, U22220, U22221, U22222, U22223, U22224, U22225, U22226, U22227 and U22228.

described. The complete structure of the rat aggrecan gene has now been reported [29]. In addition, the human gene has recently been mapped to chromosome 15q25→p26.2 [30] and 15q26 [31] by two independent groups. Aggrecan is known to be involved in disease processes, and its rate of synthesis [32] and post-translational modifications [33] are known to change in cartilage development and pathology. Fatal mutations of the aggrecan core protein gene that result in premature termination of translation have been described for the nanomelic chick [26] and the cartilage matrix deficiency (*cmd*) mouse [34]. Although alternatively spliced domains have been described for the aggrecan gene, little is known about the mechanisms controlling the expression of alternative forms of the gene. In order to define the relationship between the aggrecan gene structure and the protein domains, as well as to provide insight into the mechanisms underlying alternative expression of the gene, we have determined the exon–intron-junction sequences and the complete structure of the human aggrecan gene. The results indicate a strong correlation between exon organization and the protein domains and suggest a modular assembly of the gene. The experimental approach that was developed for this study, using a yeast artificial chromosome (YAC) library and direct cosmid sequencing, provides a general method for complete and rapid characterization of relatively large genes.

MATERIALS AND METHODS

Isolation of the human aggrecan YAC

A probe for the human aggrecan gene (nucleotides +137 to +492 of the cDNA sequence [2]), representative of part of the coding sequence of the G1 domain, was prepared by PCR amplification [35] of human genomic DNA using *Taq* polymerase (Stratagene, La Jolla, CA, U.S.A.) and 20 bp primers designed from the published cDNA sequence [2]. To prepare a probe for the G3 region of the gene, another pair of primers were designed to amplify the human aggrecan 3′ end (nucleotides +6899 to +7103 of the cDNA [2]). Amplification of human genomic DNA with this second pair of primers generated a 1.6 kb PCR product of the aggrecan gene. The probes were then used to screen the Centre d'Edute du Polymorphisme Humain (C.E.P.H.) YAC library [36] by Southern-blot hybridization, and a positive clone was isolated and screened further by PCR, pulsed-field gel electrophoresis and Southern-blot hybridization using the primers and probes described above.

Subcloning of the aggrecan YAC into cosmids

Yeast spheroplasts were prepared from 30 ml cultures of the human aggrecan YAC clone (C.E.P.H. clone no. 85D1; 450 kb) as described by Phillipsen et al. [37] using yeast lytic enzyme (ICN Biochemicals, Costa Mesa, CA, U.S.A.). High-molecular-mass DNA was then isolated from the spheroplasts and subcloned into the Supercos 1 cosmid vector (Stratagene) [38–40]. Briefly, spheroplasts were lysed in 10 mM NaCl/20 mM Tris/HCl (pH 8)/1 mM EDTA containing 100 μg of proteinase K/ml and 0.5 % SDS overnight at 50 °C. The lysate was extracted with phenol/chloroform and transferred to a fresh 50 ml centrifuge tube. Then 1 vol. of ethanol was layered over the lysate, and the DNA was spooled from the interface through the layer of ethanol. A partial *Mbo*I digestion [39] of the total yeast DNA was performed to prepare DNA with an average size of approx. 50 kb. The digestion DNA was dephosphorylated with calf intestinal alkaline phosphatase (Promega, Madison, WI, U.S.A.), purified by phenol/chloroform extraction, and ligated to *Xba*I–*Bam*HI-digested Supercos 1 vector. The ligation mix was then packaged using the Gigapack II XL (Stratagene) λ phage packaging extract [40]. NM554 host cells were infected with the packaged phage and cultured on 15 cm Luria broth/ampicillin plates at 37 °C overnight. Replica nylon filters were prepared from the plates [41] and screened with biotinylated probes for the 5′ region and the G3 region of the human aggrecan gene using the PhotoGene DNA-detection kit (Life Technologies/Gibco BRL, Grand Island, NY, U.S.A). Positive clones were isolated and screened further by PCR and Southern-blot hybridization.

*Eco*RI restriction mapping of cosmids

The ends of the cosmid clones were sequenced with T7 and T3 primers, and complementary primers were designed from the sequences generated. Biotinylated probes (200–300 bp) were then generated by PCR amplification in the presence of biotin-labelled dATP (Life Technologies/Gibco BRL). Southern blots of partial *Eco*RI digests of each cosmid were then hybridized with the biotinylated probes and detected using the PhotoGene DNA-detection kit.

Sequencing strategy

The general strategy was to design primers (primers 1–40; Figure 1) specific for the human aggrecan cDNA [1–3] and use them in direct cosmid sequencing to determine the structural organization and exon–intron-junction sequences of the human aggrecan gene. Sequence-walking primers were initially designed at three distinct sites of the human aggrecan cDNA, i.e. the regions coding for the G1-A module, the KS-attachment region, and the 3′ untranslated region (UTR). Each primer was selected such that the G+C content was as close to 50 % as possible. Beginning at each of the sites indicated above, direct cosmid sequencing progressed in both directions. In order to map the exon–intron junctions, five rounds of sequence walks were performed. In round 1, splice-junction sequences were generated using primers 1–5 (Figure 1). The directions and lengths of sequences obtained are indicated by arrows in Figure 1. Primer positions are indicated by the open boxes at the tails of the arrows. The next round of sequencing utilized primers 6–20 in order to map and sequence additional splice junctions. As exons 13, 14 and 18 are alternatively spliced, their splice sites were predicted by comparing the published cDNA sequences [1–3], obviating the need for initial sequence walks in adjacent exons before designing primers for exons 13, 14 and 18. Sequences generated at primer positions 3 and 18 (Figure 1) were obtained by sequencing the ends of cosHA-G3 and cosHA-1 respectively with the T7 promoter primer. The third, fourth and fifth rounds of sequencing used primers 21–31, primers 32–35 and primers 36–41 respectively. To verify the exon–intron-junction sequences obtained, intron-specific primers (primers 42–76) (Table 1, Figure 1) for reverse or second-strand sequencing were designed at an average distance of about 120 bp from each splice acceptor or splice donor site and used to sequence the exon–intron junctions.

This strategy provided complete sense and complementary sequence for all of the exons except exon 12, which has been shown to be at least 2.5 kb in size [25,28,29]. The size of exon 12 was confirmed by PCR amplification and restriction analysis, in combination with data obtained for its exon–intron-junction sequences.

Sequencing of human aggrecan exon–intron junctions

Three cosmid clones (cosHA-1, cosHA-2 and cosHA-G3), subcloned from the human aggrecan YAC and containing the

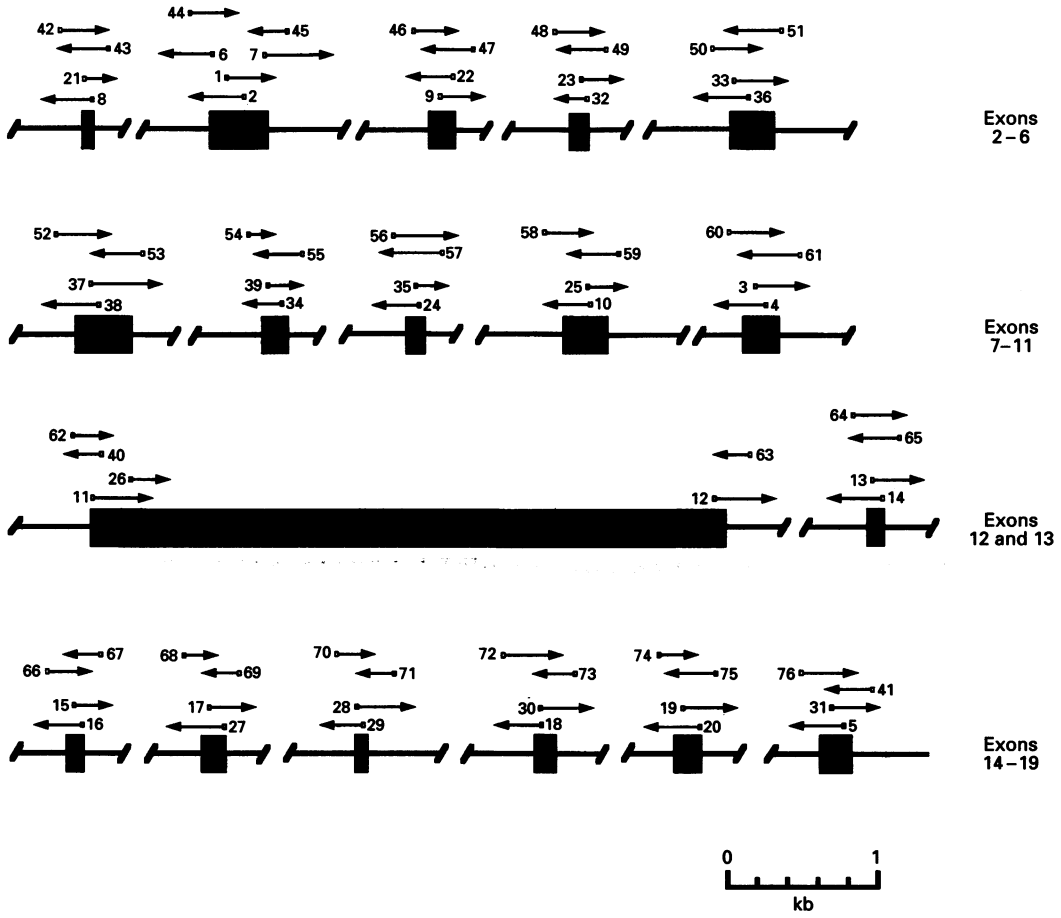

Figure 1 Schematic representation of the sequencing strategy

Sequence-walking primers (20-mers) were designed from the published human aggrecan cDNA sequences [1–3] and used to directly sequence YAC-derived human aggrecan cosmids in order to determine the exon–intron splice-junction sequences of the gene. Solid bars represent exons, and the interrupted lines indicate introns. The numbers and open boxes at the tails of the arrows indicate sequencing primers, and the arrows indicate the directions of sequencing and the lengths of sequences obtained. The sequence of exon 1 (not shown) was determined by cloning and sequencing the 5′ UTR of the human aggrecan mRNA.

entire coding region of the human aggrecan gene (Figure 2), were used in sequencing and mapping the exon–intron junctions of the human aggrecan gene. Cosmid DNA was isolated from each clone using Qiagen-Tip 100 columns and buffers (Qiagen, Chatsworth, CA, U.S.A.) according to the supplier's instructions and submitted to the Columbia University DNA Synthesis and Sequencing Facility for automated sequencing (Applied Biosystems). A segment of the gene (encompassing nucleotides +800 to +1800 of the cDNA sequence [2]) proved difficult to sequence by direct cosmid sequencing. Therefore a 20 kb *Eco*RI fragment of cosHA-1 (Figure 2), containing the desired gene segment, was digested with *Apa*I and subcloned into the *Apa*I site of the pBluescript SK(+) vector for sequencing. Sequence results were submitted to the National Center for Biotechnology Information for database searching and sequence alignment using the BLASTN program [42].

The *Apa*I–*Eco*RI fragments generated at the ends of the 20 kb *Eco*RI fragment were also subcloned at the *Apa*I–*Eco*RI sites of pBluescript SK(+). A clone containing a 2.7 kb *Apa*I–*Eco*RI fragment which mapped in the CS-attachment region of the cDNA [2] was analysed further by PCR using primers designated at +4515 to +4534 (sense) and +4522 to +4503 (antisense). These primers were used along with the T7 and T3 promoter primers in the PCR reactions.

Determination of intron sizes

After mapping the locations of the human aggrecan exon–intron junctions, inter-exon PCR reactions were performed with the cosmids and selected sequence-walking primer pairs in order to determine the sizes of the introns. Two primers located in the exons immediately adjacent to a given intron (for example, primers 1 and 22, or primers 2 and 21 in Figure 1) were used in PCR to amplify the intron. PCR reactions were set up in 50 μl reaction volumes containing 2.5 units of *Taq* polymerase, 100 ng of each primer (sense or antisense) and 10–20 ng of cosmid template. Thirty-five cycles of 94 °C for 1 min, 60 °C for 1 min and 72 °C for 10 min were run overnight to amplify the introns. The sizes of the products were determined by agarose-gel electrophoresis and comparison with molecular size markers. The actual size of each intron was calculated by subtracting the number of coding bases, which were amplified by the primers, from the size of the respective PCR product.

Table 1 Intronic sequencing primers of the human aggrecan gene

Primer numbers are as indicated in Figure 1.

Primer number	Primer sequence
42	5′-CTCTGGAGATGATTCCAGGT-3′
43	5′-CGAGGGGGAATTATAACCCA-3′
44	5′-CATCCCCATCATAGAGACAG-3′
45	5′-ACAGAGGAAGCAGCTGAAGT-3′
46	5′-TATTCCCATAATTCTGCGAG-3′
47	5′-GGTGTGACCACGTCCTCTAA-3′
48	5′-AGAAGTTGCTGTTTCCTCCC-3′
49	5′-AGGAGGACAGAGTGTGTCAG-3′
50	5′-ATGGGACCAGGACTTTGGGA-3′
51	5′-TTCCCTCTAGGACCCCATTT-3′
52	5′-GGTGCCTGGCACTCTGTGGT-3′
53	5′-CAGCCTTCAACTGCGGGAAT-3′
54	5′-TAAGACAGCAGGAGTAGGCC-3′
55	5′-CACAGGGTATTGGGTGCTCA-3′
56	5′-AGACAACTGAAGACATCTCC-3′
57	5′-CCTGGAAAGCTGGGTGGAAT-3′
58	5′-AAAGGGGTGATGGAGCTGGT-3′
59	5′-ACACTGGGTGGTGACAAGAT-3′
60	5′-GCCTGTCCACCTCAGAGTCC-3′
61	5′-GTTGCCCATAGCAGAACACC-3′
62	5′-GTTGCTGAGCTCTTGAGAGA-3′
63	5′-TTACGTTACAGATGAGGCTC-3′
64	5′-GTGGACTTCAGGAACCTCAT-3′
65	5′-CCTGAGCCTCAGTGGACCTT-3′
66	5′-CAGTTCTCAGGAAAACCAGC-3′
67	5′-TGGCTCTCCAGATGAACCTC-3′
68	5′-TCCCTTCCCTTGAGGGCACA-3′
69	5′-TCAGTCTCACGAGGACCAAA-3′
70	5′-AAGCACGTGCCTTGCTCCTA-3′
71	5′-AGCAACAACCTGTTCACCAC-3′
72	5′-TGTGACTGATTCCCTAGAGG-3′
73	5′-GTCTGCCTTTTGGACAGGTG-3′
74	5′-CACTGTCAGATGTTGAGGCT-3′
75	5′-CTCAACCCTGTGGACGGTGT-3′
76	5′-GAAAGCCGATAAAGCCTCAG-3′

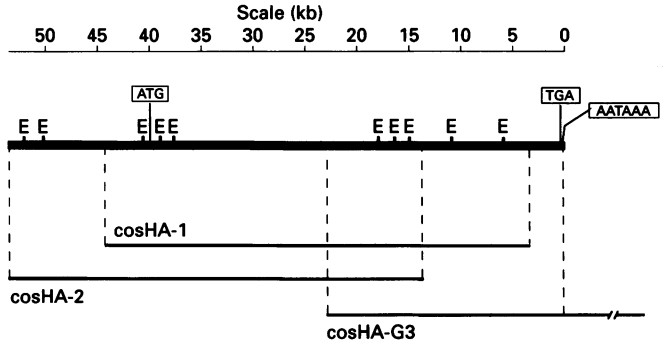

Figure 2 *Eco*RI map and positions of the selected cosmids on the human aggrecan gene

The YAC-derived cosmid clones were isolated by screening with biotinylated probes against the 5′ and 3′ ends of the coding sequence of the human aggrecan gene. Overlapping clones containing the entire coding region of the gene were identified by PCR amplification with primers specific for the 5′ and 3′ ends of the gene. The positions of the clones relative to the gene were determined by *Eco*RI restriction analysis and alignment of restriction fragments. The size of each insert in each cosmid was approx. 40 kb. The positions of the ATG translation-initiation codon, the TGA translation-stop codon and the AATAAA polyadenylation signal are indicated. The gene fragment is scaled from the AATAAA polyadenylation signal, as the cosmids do not contain the promoter and exon 1 of the gene. E represents *Eco*RI cleavage sites.

Cloning and sequencing of exon 1

Sequencing and Southern-blot hybridization revealed that the three cosmid clones used in this study did not contain exon 1 or the promoter region of the human aggrecan gene. Therefore the 5′ UTR of the human aggrecan mRNA was cloned using the AmpliFINDER RACE technique (Clontech, Palo Alto, CA, U.S.A.). Briefly, polyadenylated mRNA was purified, using the Micro-FastTract mRNA-isolation kit (Invitrogen, San Diego, CA, U.S.A.), from human chondrocytes isolated [43] from articular cartilage specimens obtained as a by-product of knee-replacement surgery. The mRNA sample was reverse-transcribed with avian-myeloblastosis-virus reverse transcriptase and an antisense primer (5′-GCCTCGCTGTCCTCGATGCC-3′) designed from the published human aggrecan cDNA sequence [2]. After hydrolysis of the mRNAs and purification of the first-strand cDNA, the AmpliFINDER anchor was ligated to the cDNA, and the 5′ UTR of the human aggrecan cDNA was amplified by PCR using *Taq* DNA polymerase, nested gene-specific primers (5′-GAACTTCAGTCCCACACACC-3′ or 5′-AGAGTTGGACTCCCTTCTCC-3′) and the anchor-specific primer. The PCR product was directly cloned into the PCR II TA cloning vector (Invitrogen) and sequenced in the forward and reverse directions using T7 and SP6 primers.

RESULTS AND DISCUSSION

Exon–intron organization of the human aggrecan gene

Sequencing of genomic cosmids of the human aggrecan gene revealed that the gene contains 19 exons and 18 introns (Table 2; see Figure 4). The exons ranged in size from 77 to 4224 bp. The putative mRNA, including all of the alternatively spliced exons, codes for a protein core of calculated molecular mass 254379 Da.

The sequences of the splice junctions (Table 2) strictly conformed to the GT/AG rule of splice-junction sequences [44]. The introns were shown to be primarily class I, where each intron interrupts the coding sequence between the first and second bases of the codon. Introns 4 and 8 were class II, and intron 16 was a class 0. The sizes of the introns, as estimated by PCR amplification (Figure 3) and *Eco*RI restriction mapping, ranged from 0.5 to > 13 kb (Table 2). Intron 1 was greater than 13 kb, and intron 13 was shown to be 8.4 kb in size. The remaining introns were approximately 3 kb or less. The data therefore showed that the entire coding region (exons 2–19 and introns 2–18) is contained in a 39.4 kb segment of the gene (Table 2, Figure 2).

The exon–intron organizations of the human (Figure 4) and rat [29] aggrecan genes are quite similar. One notable difference between the gene structures of the human and rat aggrecan genes is that the first rat aggrecan EGF-like domain was mutated, thus precluding its expression.

The complex nature of the human aggrecan gene (Figure 4) and the structural similarity of the different domains to domains of other genes indicate a modular structure. With this in mind, we have searched the sequence databases for proteins with sequence similarity to the human aggrecan gene and compared the exon organizations of these proteins with that of aggrecan. The aggrecan gene structure is therefore presented in relation to its protein domains.

The 5′ UTR and signal peptide (SP)

Exon 1 provides the 5′ UTR of the human aggrecan mRNA (Figure 4). It is at least 375 bp long and has a base composition of

Table 2 Characteristics of the exon–intron junctions of the human aggrecan gene

Note that each intron begins with a GT and ends with an AG, a direct compliance with the GT/AG rule of splice-junction sequences. The protein domain boundaries are as defined by Doege et al. [2]. A protein domain(s) is coded by each exon or a combination of exons. Intron sizes are accurate within ± 0.1 kb. Amino acid numbering commences with the methionine of the translation-initiation codon as amino acid number 1. In addition to the 3′ UTR, exon 19 also codes for the 25-amino-acid C-terminus of aggrecan.

Protein domain	Exon No.	Exon Size (bp)	5′ Splice donor	Intron No.	Intron Size (kb)	Intron Class	3′ Splice acceptor	Amino acid at junction
5′ UTR	1	375	CTTCAAG.....	1	> 13	–	cctcccctcttcc**ag**GTGAACT	–
SP	2	77	ACTTCAG**gt**gaggacatccctat	2	2.4	I	acctctcccacac**ag**ACCATGA	Asp-24
G1-A	3	384	GTGAAAG**gt**gagagcctcccaca	3	1.0	I	gtatgtgtcctgc**ag**GCATCGT	Gly-152
G1-B	4	175	CTGTCAG**gt**gagccctagcccat	4	1.6	II	tcgggttcctggc**ag**ATACCCC	Arg-210
	5	128	ATGGAGG**gt**gagctgccctgccc	5	1.5	I	ctgtgtccttcac**ag**GTGAGGT	Gly-253
G1-B′	6	294	TACACAG**gt**ggggcacggctggt	6	1.9	I	ttgcccctcccct**ag**GTGAAGA	Gly-351
IGD	7	378	CCAGGGG**gt**aagtagctgcccgt	7	1.5	I	ctcctccccaccc**ag**GGGTCGT	Gly-477
G2-B	8	175	CCGTCAG**gt**gaagccatgctcct	8	0.5	II	cactctcctttgc**ag**ATACCCC	Arg-535
	9	128	CTTGAGG**gt**acaagccacattct	9	1.4	I	tgcctctgcccccc**ag**GGGAGGT	Gly-578
G2-B′	10	294	TTCCGAG**gt**atgcagcctcactt	10	2.1	I	ccacatctcctttt**ag**GCATTTC	Gly-676
KS1	11	240	CTGCCAG**gt**tggtatggcttggg	11	2.9	I	ccttctttcctac**ag**GGATCCT	Gly-756
KS2/CS	12	4224	GCTGGAG**gt**attgtgatttttttc	12	0.9	I	ccctgggggttgc**ag**CCCCCGC	Ala-2164
EGF1	13	114	AACATAG**gt**aaggccctcattgg	13	8.4	I	cttggtttcttgc**ag**ACATTGA	Asp-2202
EGF2	14	114	GAGATTG**gt**acggccgtcttggc	14	3.2	I	tgacctgtgttgc**ag**ACCAGGA	Asp-2240
LEC	15	159	GTCAACA**gt**gagtgcggcggggc	15	0.5	I	tcacccttcccc**ag**ACAATGC	Asn-2293
	16	83	CCCCATG**gt**gagttctgctgtag	16	0.8	0	cactcccacccac**ag**CAATTTG	Met-2320/Gln
	17	145	GCCACAG**gt**aagctggcgcctgg	17	0.8	I	gtgtttgcccctc**ag**TGGCCTG	Val-2369
CRP	18	183	ACAGACC**gt**gagcatcaccccgg	18	0.5	I	tccctttcgtcct**ag**CCACCAC	Pro-2430
3′ UTR	19	212+	–	–	–	–	–	–
		Consensus:	A g AG**gt** ag C a				cc t ct tc G c ct c nc**ag** CA tt c tc ct A	

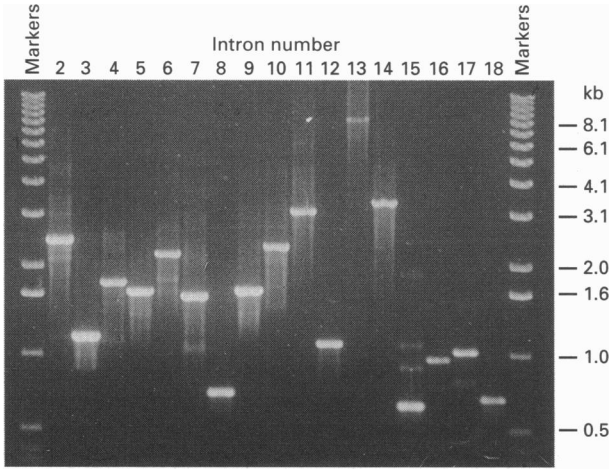

Figure 3 PCR amplification products of introns 2–18 of the human aggrecan gene

Selected pairs of sequence-walking primers which flanked each intron were used to amplify the introns. The products were fractionated on a 0.7% agarose gel (25 cm × 20 cm) containing ethidium bromide at 80 V. Electrophoresis buffer consisted of 0.5 × TTE buffer [45 mM Tris/taurine (pH 8.0)/1 mM EDTA]. The bands were visualized over UV transillumination and photographed.

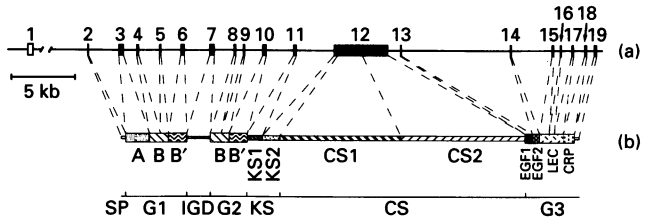

Figure 4 Correlation of the genomic map and protein domains of the human aggrecan gene

(**a**) Exon and intron organization of the aggrecan gene. The filled bars represent coding exons, and the open bar represents the non-coding exon 1. Lines between the bars indicate introns. Note that the entire coding region is contained in a 40 kb segment of the gene. (**b**) Schematic diagram of the protein domains of the aggrecan core protein. Domain boundaries are as defined by Doege et al. [2], except for the division of the KS-rich region into KS1 and KS2 domains. Note that several protein domains are encoded by specific single exons.

```
  1 GGCCCGACCA CCTACCTCCC CGCCGCTCCA GAGGGGGCTC GCAGAGCTGA GGACGCGCGC
 61 AGCGCTGCTC AAGGTCTCTC TCTCTCAGCA CCCTCGCCGG CCGGCGTCTG ACGCGGGTGC
121 CAGGGTCTCC GGGCACCTTT CAGTGTCCAT TCCCTCAGCC AGCCAGGACT CCGCAACCCA
181 GCAGTTGCCG CTGCGGCCAC AGCCCGAGGG GACCTGCGGA CAGGACGCCG GCAGGAGGAG
241 GGGTGCGCAG CGCCCGCGCA GAGCGTCTCC CTCGCTACGC AGCGAGACCC GGGCCTCCCG
301 GCCCCAGGAG CCCCCAGCTG CCTCGCCAGG TGTGTGGGAC TGAAGTTCTT GGAGAAGGGA
361 GTCCAACTCT TCAAGGTGAA CT
                ▲
            Intron 1
```

Figure 5 Sequence of the 5′ UTR of the human aggrecan gene

cDNA was synthesized from the 5′ UTR of the human aggrecan mRNA and cloned using the AmpliFINDER RACE and TA cloning techniques (see the Materials and methods section). Bases 1–375 represent the sequence of exon 1. The arrow indicates the insertion site of intron 1.

70% G-C and 30% A-T (Figure 5). Exon 2 is only 77 bp (Table 2) and codes for 7 nucleotides of the 5′ UTR, the ATG translation-initiation site and the putative SP of the protein [2]. The predicted cleavage site of SP, Ala^{16}-Ala^{17}, conforms to the

(−3, −1) rule [45]. Although the putative SP shows reasonable amino acid sequence similarity to those of a number of other genes, for example, mouse immunoglobulin μ heavy chain [46], no signal sequence similarity to other hyaluronan-binding proteins was apparent [19–23].

G1 domain

The G1 domain of aggrecan has the specific function of binding to hyaluronan, and it shows remarkable sequence similarity to hyaluronan-binding domains of other members of the family of hyaluronan-binding proteins (e.g. rat neurocan [19], human versican [20], bovine brevican [21] and rat and cat BEHAB [22]). The A loop of G1 is encoded by exon 3, the B loop by exons 4 and 5, and the B′ loop by exon 6. The intron class for introns 3, 5 and 6 is I, whereas that for intron 4 is II (Table 2). The presence of a class-II intron between exons 4 and 5 indicates that, within the G1 domain, these two exons cannot be alternatively spliced or duplicated independently of each other.

The exon organization of human aggrecan G1 is quite similar to that described for the chick [47] and rat [48] link proteins. The link protein exons 3 (375 bp in chick and 372 bp in rat), 4 (303 bp) and 5 (287 bp of coding sequence), which code for the link protein A, B and B′ loops respectively, correspond to the human aggrecan exon 3 (384 bp), the combined exons 4 (175 bp) and 5 (128 bp), and exon 6 (294 bp). It is noteworthy that the chick [47] and rat [48] link protein B loops are encoded by a single 303 bp exon whereas the G1-B loops of human (Table 2, Figure 4) and rat [29] aggrecan are each encoded by two distinct exons. Preliminary sequencing results obtained from two overlapping genomic clones of human versican identified two exons [49] that may be counterparts of the human aggrecan exons 3 (384 bp) and 6 (294 bp)

IGD and G2

The IGD is encoded solely by exon 7 (Figure 4), which is 378 bp long. There is little sequence similarity of the IGD to known sequences of other proteins, either at the amino acid or nucleotide level, suggesting that exon 7 is aggrecan-specific. However, a proline-rich segment (residues 403–425) showed significant sequence similarity to several proteins including human γ-glutamyltransferase [50], mycoplasma pneumoniae 30 K adhesin-related protein [51], rye omega seculin precursor [52] and the tegument protein of herpes virus saimiri [53].

The G2 domain protein structure consists of the PTR loops B and B′, which are homologous with the B and B′ loops of the G1 domain of aggrecan and link protein [54]. The B loop is encoded by exons 8 and 9, and the B′ loop by exon 10 (Figure 4). The sizes and organization of these exons are exactly the same as for the exons for the PTRs of the G1 domain (Table 2, Figure 4). Similarly, the classes of the introns are exactly the same. This clearly suggests that the PTRs of the G1 and G2 domains are derived from a common origin, either from a common ancestral gene or by duplication of the exons coding for the motifs within this gene.

Glycosaminoglycan (KS and CS)-attachment regions

The KS- and CS-attachment regions of aggrecan have been shown to exhibit sequence and size differences between species [2,9,24–26,28]. This variability in size is due primarily to the presence or absence of tandem repeat sequences in both the KS- and CS-rich domains [55]. Our sequencing data show that the KS-rich region of the human aggrecan gene is encoded by exon 11 (240 bp) and the 5′ end of exon 12 (Figure 4). The amino acid sequence encoded by exon 11 shows identity with equivalent sequences in bovine [9], mouse (GenBank accession number L07049), rat [24] and chick [25,26] aggrecan. This indicates that this region is conserved between species and is unlikely in itself to account for the variability in KS content in aggrecans of the different species. The exon 12-encoded portion of the KS-rich region (Figure 4) contains 12 consecutive hexapeptide repeats in human aggrecan [2,55] and 23 in bovine aggrecan [9,55]. Chick and rat aggrecans contain three to four copies of the hexapeptide repeats; however, they are poorly conserved [55]. In contrast with the exon 11-encoded segment, it seems likely that this part of the KS-rich region accounts for much of the variability in KS content between aggrecans from different species. Because of the sequence differences between the two segments of the KS-rich region encoded by exons 11 and 12 (Figure 4), we suggest designating the exon 11-encoded segment KS1 and the exon 12-encoded segment KS2. This is consistent with the criteria used in assigning CS1 and CS2 for the two distinct parts of the CS-rich region [2].

Exon 12 of the human aggrecan gene is 4224 bp long and also codes for the complete CS-attachment domain, consisting of both the CS1 and CS2 regions [2]. Confirmation of this size was obtained from PCR amplification, restriction analysis and plasmid subcloning, which showed no evidence of an intron in the region spanned by this exon (results not shown). Of the various complete cDNA sequences described so far for the aggrecan gene, the human aggrecan gene contains the largest CS-coding exon. The chick counterpart of the human aggrecan exon 12 is only 2856 bp long [25] and that of the rat aggrecan is 3741 bp [28,29]. The exceptionally large size of the human aggrecan exon 12, as well as those of chick and rat, is a feature shared by tandem-repeat-coding exons of several other genes. Among these are the 1.41 kb exon of the human tenascin gene [56] and the 14 kb exon of the human profilaggrin gene [57].

G3 region

The G3 region of aggrecan consists of two alternatively spliced EGF-like domains [1,3], a LEC-like domain and an alternatively spliced CRP-like domain [1,2]. The human aggrecan exons 13 and 14 code for the EGF-like domains (Figure 4). The LEC-like domain is encoded by exons 15, 16 and 17, and the CRP-like domain by exon 18. The G3 introns are all class I, except intron 16 which is class 0 (Table 2).

A major function of G3 appears to be contributing to intracellular trafficking of aggrecan [26,58], although which of its various domains is responsible for this function remains to be determined. The two EGF-like exons are both 114 bp in size (Table 2). The biological functions of these domains in aggrecan are not known, but it is likely that they are involved in Ca^{2+} binding. The third and fourth EGF-like domains of human protein S [59] and the first EGF-like domain of human coagulation factors IX and X [60] have been shown to possess high affinity Ca^{2+}-binding sites. Thus it is likely that the EGF-like domains of aggrecan bind Ca^{2+} and interact with the Ca^{2+}-dependent LEC-like carbohydrate-binding domain.

The LEC-like domain of aggrecan bears significant amino acid sequence similarity to several other proteins, including the human leucocyte adhesion molecule 1 (LAM-1) [61], the sea raven antifreeze protein [62] and the human macrophage mannose receptor [63,64]. This domain is encoded by three exons (Figure 4), similar to chicken hepatic lectin [65] and murine low affinity IgE Fc receptor [66,67] carbohydrate-recognition domains. Interestingly, the nucleotide sum of the three LEC-like exons of

```
  1 GAAGAGCTTC CAGGACGCAC CCAGGACGCT GAGCCCAGGA GCCTGCCAGG CTGACGTGCA
 61 TCCCACCCAG ACGGTGTCCT CTTCTTGTCG CTTTTTGTCA TATAAGGAAT CCCATTAAAG
121 AAGGAAAAAA ATAAATCCCA CATTTGTGTA TGCACCCACT CACCCCTCCA AATCAGCAAA
181 ACCGCATCTA ATTTGTCCGC CGAATGCCAA AGCAAAGCAA ACTTATTATA ACCCTTGGAC
241 TGAGTTTAGA GACATTTCTT
```

Figure 6 Genomic sequence of the 3′ UTR of the human aggrecan gene

The sequence of the 3′ UTR was obtained by directly sequencing cosHA-G3 using the primer 5′-CCACTGAGAAGAGCTTCCAG-3′ (nucleotides +7005 to +7024 of the human aggrecan cDNA [2]) as sense primer and the primer 5′-AGCGACAAGAAGAGGACACC-3′ (nucleotides +7084 to +7065) as antisense primer. The sequences generated were then confirmed by further sequencing of cosHA-G3 with an intron 18-specific primer (5′-GAAAGCCG-ATAAAGCCTCAG-3′) and a downstream antisense primer (5′-TCCCTGGAAAGGCACGATGG-3′) designed from the sequence obtained with the initial sequencing primer at +7005 to +7024. Two polyadenylation signals (bold-type) were identified.

aggrecan (Table 2) and the size of the single exon coding for the LEC-like domain of LAM-1 [61] are both 387 bp.

The human aggrecan CRP-like domain (Figure 4) is a single copy of the short consensus repeats commonly found in complement receptor proteins [68–70] and the selectin family of adhesion molecules [61,71,72]. Like the aggrecan gene, the selectins LAM-1 [61] and endothelial leucocyte adhesion molecule 1 [71,72] both contain EGF-, LEC- and CRP-like domains. However, there is a notable difference between aggrecan (Figure 4) and the selectins in the organization of the exons coding for these domains. The two exons coding for the aggrecan EGF-like domains are located 5′ to the LEC-like domain, whereas the exon coding for the single EGF-like domain of the selectins is located between the LEC-like domain and the first short consensus repeat or CRP-like domain.

The 3′ UTR

Exon 19 codes for the 25-amino-acid C-terminus of the core protein and the 3′ UTR of the aggrecan gene. Sequencing 260 bp of the 3′ UTR (Figure 6) revealed two polyadenylation signals in close proximity. A potential polyadenylation signal (ATTAAA) [73,74] and a true polyadenylation signal (AATAAA) are respectively located 114 bp and 130 bp downstream of the translation stop codon. It is not known whether additional polyadenylation signals exist 3′ to those described above. However, an upstream polyadenylation signal is located in intron 11 (results not shown), 34 bp from the splice acceptor site. The significance or function of this upstream polyadenylation signal is not known.

Coding sequence for the human aggrecan gene

Two modest differences were found between the coding sequence obtained in the present study and the previously published cDNA sequence [2]. In exon 7, the genomic sequence corresponding to positions 1562–1564 of the published human aggrecan cDNA sequence [2] was found to be TGC, whereas the cDNA sequence is CTG at the same positions. This difference results in the conversion of Pro-401 and Gly-402 of the predicted protein sequence [2] into Leu-401 and Arg-402. This change is consistent with published sequences of the aggrecan gene in other species, at both the nucleotide and amino acid levels [9,24,26]. In exon 10, the genomic sequence in the present study showed that a CTG (coding for leucine) should be inserted between nucleotides 1872 and 1873 of the published cDNA sequence [2]. Again this is consistent with the published cDNA sequences for other species.

The general experimental approach described in this study, i.e. the use of a YAC library and sequencing from cosmid clones derived from the YAC clone, is one that has potential for application to many studies of gene structure and regulation. This is particularly true for genes that are relatively large, as is aggrecan, because the complete gene is most likely to be contained in a single YAC clone. For relatively small genes, the direct screening of a cosmid library may prove more efficient, because the whole gene and its regulatory regions are likely to be enclosed within a single clone of this smaller size.

Although no specific functions have been defined for some of the domains or modules in aggrecan, their presence and conservation indicates that they play essential biological roles. The results presented here have allowed further insight into the organization of the aggrecan protein core. Examples include the observed differences between the exon-11- and exon-12-encoded segments of the KS region and their designation as KS1 and KS2 subdomains. The description of the exon–intron organization and the sequences of their splice junctions also allows an understanding of potential alternative splicing events that are likely to occur during expression of the human aggrecan gene.

This work was supported by grants from Searle and the Orthopaedic Research and Education Foundation. We thank Dr. Eftihia Cayanis, Mr. Baresh Chauhan and Mr. Phillips Perera for technical assistance.

REFERENCES

1 Baldwin, C. T., Reginato, A. M. and Prockop, D. J. (1989) J. Biol. Chem. **264**, 15747–15750
2 Doege, K. J., Sasaki, M., Kimura, T. and Yamada, Y. (1991) J. Biol. Chem. **266**, 894–902
3 Fulop, C., Walcz, E., Valyon, M. and Glant, T. T. (1993) J. Biol. Chem. **268**, 17377–17383
4 Hardingham, T. E. and Fosang, A. (1992) FASEB J. **6**, 861–870
5 Heinegard, D. and Oldberg, A. (1989) FASEB J. **3**, 2042–2051
6 Heingard, D. and Hascall, V. C. (1974) J. Biol. Chem. **249**, 4250–4256
7 Fosang, A. L. and Hardingham, T. E. (1989) Biochem. J. **261**, 801–809
8 Paulsson M., Yurchenco, P. D., Ruben, G. C., Engel, J. and Timpl, R. (1987) J. Mol. Biol. **197**, 297–313
9 Antonsson, P., Heinegard, D. and Oldberg, A. (1989) J. Biol. Chem. **264**, 16170–16173
10 Heinegard, D. and Hascall, V. C. (1974) Arch. Biochem. Biophys. **165**, 427–441
11 Doege, K., Hassell, J. R., Caterson, B. and Yamada, Y. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 3761–3765
12 Hourcade, D., Holers, V. M. and Atkinson, J. P. (1989) Adv. Immunol. **45**, 381–416
13 Dennis, J. E., Carrino, D. A., Schwartz, N. B. and Caplan, A. I. (1990) J. Biol. Chem. **265**, 12098–12103
14 Weidemann, H., Paulsson, M., Timpl, R., Engel, J. and Heinegard, D. (1984) Biochem. J. **224**, 331–333
15 Perkins, S. J., Nealis, A. S., Dudhia, J. and Hardingham, T. E. (1989) J. Mol. Biol. **206**, 737–748
16 Perkins, S. J., Nealis, A. S., Dunham, D. G., Hardingham, T. E. and Muir, I. H. (1991) Biochemistry **30**, 10708–10716
17 Goldstein, L. A., Zhou, D. F., Picker, L. J. et al. (1989) Cell **56**, 1063–1072
18 Stamenkovic, I., Amiot, M., Pesando, J. M. and Seed, B. (1989) Cell **56**, 1057–1062
19 Rauch, U., Karthikeyan, L., Maurel, P., Margolis, R. U. and Margolis, R. K. (1992) J. Biol. Chem. **267**, 19536–19547
20 Zimmermann, D. R. and Ruoslahti, E. (1989) EMBO J. **8**, 2975–2981
21 Yamada, H., Watanabe, K., Shimonaka, M. and Yamaguchi, Y. (1994) J. Biol. Chem. **269**, 10119–10126
22 Jaworski, D. M., Kelly, G. M. and Hockfield, S. (1994) J. Cell Biol. **125**, 495–509
23 Lee, T. H., Wisniewski, H. G. and Vilcek, J. (1992) J. Cell Biol. **116**, 545–557
24 Doege, K., Sasaki, M., Horigan, E., Hassell, J. R. and Yamada, Y. (1987) J. Biol. Chem. **262**, 17757–17767
25 Chandrasekaran, L. and Tanzer, M. L. (1992) Biochem. J. **288**, 903–910
26 Li, H., Swartz, N. B. and Vertel, B. M. (1993) J. Biol. Chem. **268**, 23504–23511
27 Tanaka, T., Har-El, R. and Tanzer, M. L. (1988) J. Biol. Chem. **263**, 15831–15835
28 Doege, K., Sasaki, M. and Yamada, Y. (1990) Biochem. Soc. Trans. **18**, 200–202
29 Doege, K. J., Garrison, K., Coulter, S. N. and Yamada, Y. (1994) J. Biol. Chem. **269**, 29232–29240
30 Just W., Klett C, Vetter, U. and Vogel, W. (1993) Hum. Genet. **92**, 516–518

31 Korenberg, J. R., Chen, X. N., Doege, K., Grover, J. and Roughley, P. J. (1993) Genomics **16**, 546–548
32 Sandy, J. D., Adams, M. E., Billingham, M. E. J., Plaas, A. and Muir, H. (1984) Arthritis Rheum. **27**, 388–397
33 Caterson, B., Mahmoodian, F., Sorrell, J. M. et al. (1990) J. Cell Sci. **97**, 411–417
34 Watanabe H., Kimata, K., Lime, S. et al. (1994) Nature Genet. **7**, 154–157
35 Saiki, R. K. (1990) in PCR Protocols: A Guide to Methods and Applications (Innis, M. A., Gelfand, D. H., Sninsky, J. J. and White, T. J., eds.), pp. 13–20, Academic Press, San Diego
36 Dausset, J., Ougen, P., Abderrahim, H. et al. (1992) Behring Inst. Mitt. **91**, 13–20
37 Phillipsen, P., Stotz, A. and Scherf, C. (1991) Methods Enzymol. **194**, 169–182
38 Evans, G. A., Lewis, K. and Rothenberg, B. E. (1989) Gene **79**, 9–20
39 Sambrook, J., Fritsch, E. F. and Maniantis, T. (1989) in Molecular Cloning: A Laboratory Manual, 2nd edn., pp. 9.24–9.28, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
40 Short, J. M. and Pollard, A. (1988) Strategies **1**, 5–6
41 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) in Molecular Cloning: A Laboratory Manual, 2nd edn., pp. 1.93–1.100, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
42 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) J. Mol. Biol. **205**, 403–410
43 Ishikawa, Y., Chin, J. E., Hubbard, H. L. and Wuthier, R. E. (1985) J. Cell. Physiol. **123**, 79–88
44 Mount, S. M. (1982) Nucleic Acids Res. **10**, 459–472
45 von Heijne, G. (1986) Nucleic Acids Res. **14**, 4683–4690
46 Gerstein, R. M., Frankel, W. N., Hsieh, C.-L. et al. (1990) Cell **63**, 537–548
47 Kiss, I., Deak, F., Mestric, S. et al. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 6399–6403
48 Rhodes, C., Doege, K., Sasaki, M. and Yamada, Y. (1988) J. Biol. Chem. **263**, 6063–6067
49 Iozzo, R. V., Naso, M. F., Cannizzaro, L. A., Wasmuth, J. J. and McPherson, J. D. (1992) Genomics **14**, 845–851
50 Wetmore, L. A., Gerard, C. and Drazen, J. M. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 7461–7465
51 Dallo, S. F., Chavoya, A. and Baseman, J. B. (1990) Infect. Immun. **58**, 4163–4165
52 Hull, G. A., Halford, N. G., Kreis, M. and Shewry, P. R. (1991) Plant Mol. Biol. **17**, 1111–1115
53 Albrecht, J.-C., Nicholas, J., Biller, D. et al. (1992) J. Virol. **66**, 5047–5058
54 Barta, E., Deak, F. and Kiss, I. (1993) Biochem. J. **292**, 947–949
55 Upholt, W. B., Chandrasekaran, L. and Tanzer, M. L. (1993) Experientia **49**, 384–392
56 Gulcher, J. R., Nies, D. E., Alexakos, M. J. et al. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 9438–9442
57 Gan, S.-Q., McBride, O. W., Idler, W. W., Markova, N. and Steinert, P. M. (1990) Biochemistry **29**, 9432–9440
58 Vertel, B. M., Walters, L. M., Grier, B., Maine, M. and Goetinck, P. F. (1993) J. Cell Sci. **104**, 939–948
59 Dahlback, B., Hildebrand, B. and Linse, S. (1990) J. Biol. Chem. **265**, 18481–18489
60 Handford, P. A., Mayhew, M., Baron, M., Winship, P. R., Campbell, I. D. and Brownlee, G. G. (1991) Nature (London) **351**, 164–167
61 Ord, D. C., Ernst, T. J., Zhou, L.-J. et al. (1990) J. Biol. Chem. **265**, 7760–7767
62 Hayes, P. H., Scott, G. K., Ng, N. F. L., Hew, C. L. and Davies, P. L. (1989) J. Biol. Chem. **264**, 18761–18767
63 Harris, N., Super, M., Rits, M., Chang, G. and Ezekowitz, R. A. B. (1992) Blood **80**, 2363–2373
64 Ezekowitz, R. A. B., Sastry, K., Bailly, P. and Warner, A. (1990). J. Exp. Med. **172**, 1785–1794
65 Bezouska, K., Crichlow, G. V., Rose, J. M., Taylor, M. E. and Drickamer, K. (1991) J. Biol. Chem. **266**, 11604–11609
66 Nunez, R. and Lynch, R. G. (1993) Pathobiology **61**, 128–137
67 Richards, M. L., Katz, D. H. and Liu, F. T. (1991) J. Immunol. **147**, 1067–1074
68 Fingeroth, J. D. (1990) J. Immunol. **144**, 3458–3467
69 Paul, M. S., Aegerter, M., Cepek, K., Miller, M. D. and Weis, J. H. (1990) J. Immunol. **144**, 1988–1996
70 Vik, D. P. and Wong, W. W. (1993) J. Immunol. **151**, 6214–6224
71 Becker-Andre, M., Van Huijsduijnen, R. H., Losberger, C., Whelan, J. and Delamarter, J. F. (1992) Eur. J. Biochem. **206**, 401–411
72 Collins, T., Williams, A., Johnson, G. I. et al. (1991) J. Biol. Chem. **266**, 2466–2473
73 Ayte, J., Gil-Gomez, G. and Hegardt, F. G. (1993) Gene **123**, 267–270
74 Kiss, I., Deak, F., Holloway, R. G. et al. (1989) J. Biol. Chem. **264**, 8126–8134

Received 16 August 1994/13 March 1995; accepted 22 March 1995