0014-4754/93/050384-09$1.50 + 0.20

Molecular cloning and analysis of the protein modules of aggrecans

W. B. Upholt, L. Chandrasekaran and M. L. Tanzer

Department of BioStructure and Function, School of Dental Medicine, University of Connecticut Health Center, 263 Farmington Avenue, Farmington (Connecticut 06030-3705, USA)

Abstract. The large aggregating chondroitin sulfate proteoglycan of cartilage, aggrecan, has served as a prototype of proteoglycan structure. Molecular cloning has elucidated its primary structure and revealed both known and unknown domains. To date the complete structures of chicken, rat and human aggrecans have been deduced, while partial sequences have been reported for bovine aggrecan. A related proteoglycan, human versican, has also been cloned and sequenced. Both aggrecan and versican have two lectin domains, one at the amino-terminus which binds hyaluronic acid and one at the carboxyl-terminus whose physiological ligand is unknown. Both lectins have homologous counterparts in other types of proteins. Within the aggrecans the keratan sulfate domain may be variably present and also has a prominent repeat in some species. The chondroitin sulfate domain has three distinct regions which vary in their prominence in different species. The complex molecular structure of aggrecans is consistent with the concept of exon shuffling and aggrecans serve as suitable prototypes for comprehending the evolution of multi-domain proteins.
Key words. Aggrecans; lectins; keratan sulfate; chondroitin sulfate; gene structure; gene evolution.

Introduction

Since this subject has already been well reviewed we have tried to develop new insights into aggrecan structure, evolution and biosynthesis, based upon comparison of species and comparison with other relevant proteins.

Molecular cloning of aggrecan and versican

Complete coding sequences for the large aggregating chondroitin sulfate proteoglycan core protein have been obtained for the rat[16], human[20] and the chicken[11]. The relative sizes and locations of the different cDNAs encoding the proteoglycan core protein from different species are illustrated in figure 1A. The sizes of the message for aggrecan from chicken and rat are estimated to be 8–9 kb[16,54] and 10 kb for the bovine[3], by Northern blot analyses. Partial sequences of the cDNA clones corresponding to the aggrecan mRNA were initially obtained which were then used as templates for primer extention or to rescreen a cDNA or a genomic DNA library. In the case of chicken, the initial clone[54] was obtained by screening an expression library with a polyclonal antiserum specific for the large proteoglycan of chicken cartilage[55]. This clone was 1.2 kb in length, encoding the carboxyl terminus of the protein. Partial sequence of a different region of the gene was isolated by Krueger et al.[40] from another expression library. The full length coding sequence corresponding to the mRNA for the chicken aggrecan was later obtained by PCR-based primer-extended cDNA cloning steps in combination with isolation and sequencing of genomic clones[11].

A partial cDNA clone of 872 bp encoding the carboxyl terminal globular domain of rat aggrecan was initially obtained by screening an expression library using a polyclonal antiserum specific for rat chondrosarcoma proteoglycan[19]. The complete coding sequence of rat aggrecan was obtained from the sequences of overlapping clones produced by a series of primer-extended cDNA cloning steps[16]. The complete coding sequence of human aggrecan was obtained by the same group[20]. Three large non-overlapping cDNA clones were obtained by screening a human chondrocyte cDNA library using rat aggrecan cDNA probes. The gaps in the sequence were filled in by sequencing across exons within genomic clones. Sequences of several overlapping human aggrecan cDNA clones, confirming alternative splicing of the EGF-like domain were obtained by Baldwin et al.[4].

Partial cDNA sequences encoding the bovine[49] and porcine[22] aggrecan carboxyl termini and the keratan sulfate domain of bovine[3] core proteins have been reported. Porcine sequences have also been directly determined[6].

cDNA clones for versican (fig. 1B) were initially obtained by screening an expression library using the antiserum against human fetal proteoglycan[41]. One of these clones was later used as a probe to screen a human placental cDNA library[62]. Three transcripts of versican ranging in size from 8–10 kb were seen in Northern blots probed with radiolabelled cDNA, and may be the result of alternative splicing of the mRNA.

Gene structures for the human and rat aggrecans as to exon size and localization have been characterized by sequencing and restriction-fragment hybridization

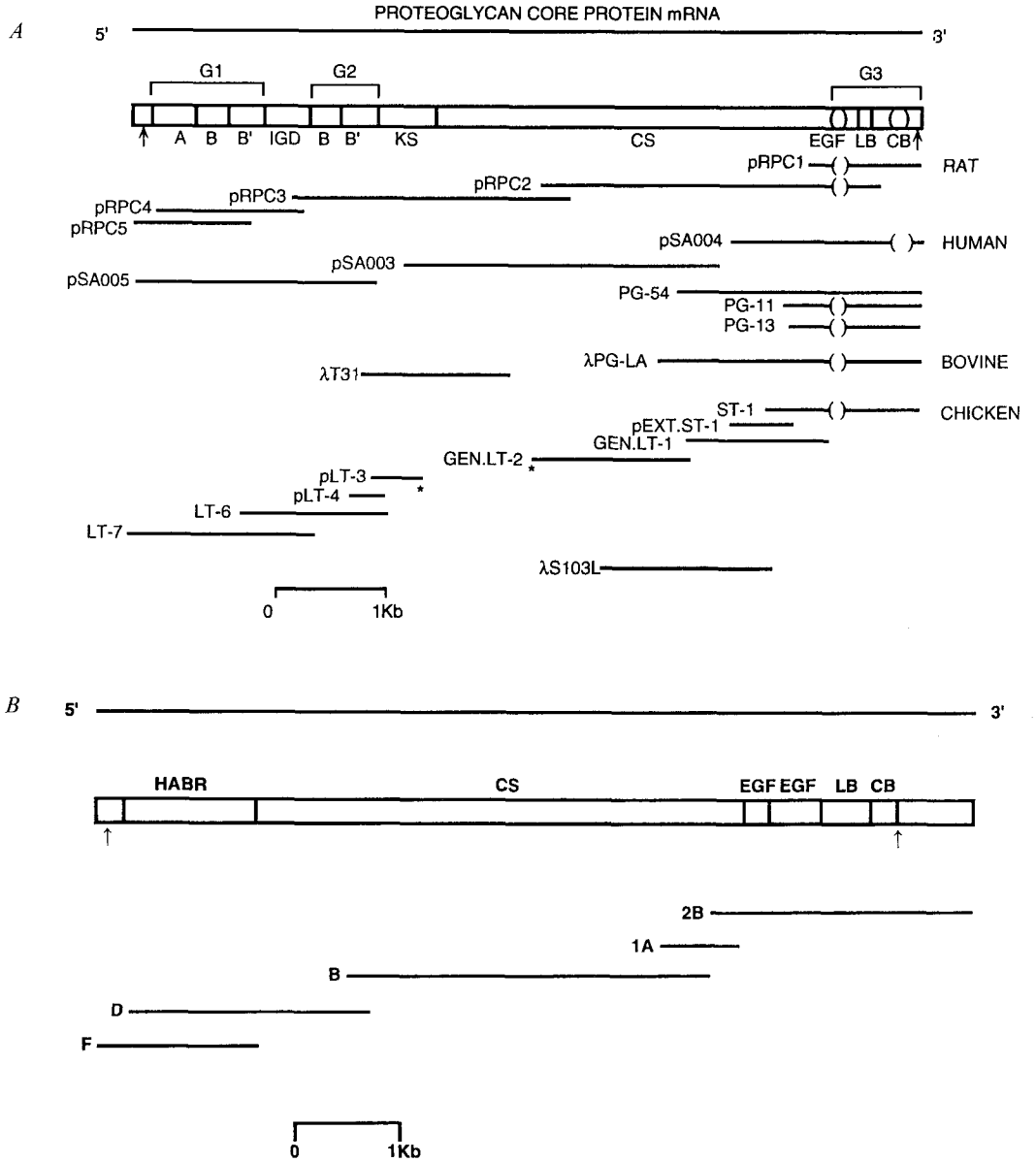

Figure 1. *A* Relative size and location of rat, human, bovine and chicken cDNAs encoding aggrecan. The mRNA and the domain structure of the translation product are shown at the top. A, B, B' are regions with homology to the link protein domains; KS, keratan sulfate rich region; CS, chondroitin sulfate attachment region; EGF, epidermal growth factor homology; LB, lectin binding protein homology; CB, complement B homology domain; arrows indicate the positions of the AUG and terminator codons. Parenthesis indicate the absence of alternatively spliced domains encoding the EGF homology domain or the complement B homology domain, * clones p LT-3 and Gen. LT-2 from chicken are overlapping clones with 72 basepairs of overlapping sequence. They are shown separately in order to align with the domain structures shown in the mRNA. References: ST-1, Sai et al.[54]; pRPC1-5, Doege et al.[16,19]; λPG-LA, Oldberg et al.[49]; λT31, Antonsson et al.[3]; PG-11, 13 and 54, Baldwin et al.[4]; λS103 L, Krueger et al.[40]; pSA003, 004 and 005, Doege et al.[20]; pEXT. ST-1, GEN. LT-1, GEN. LT-2, pLT-3, pLT-4, LT-6 and LT-7, Chandrasekaran and Tanzer[11]. *B* Relative size and location of the overlapping cDNA clones coding for versican. The mRNA and domain structure of the translation product are shown at the top. HABR, hyaluronic acid binding domain; CS, chondroitin sulfate attachment domain; EGF, epidermal growth factor homology; LB, lectin binding protein homology; CB, complement B homology domain. Arrows indicate AUG and terminator codons. References: 2B and 1A, Krusius et al.[41]; B, D and F, Zimmermann and Ruoslahti[62].

analysis[17,18]. The intron/exon organization of the G-3 encoding portion of the chicken aggrecan gene[57] has also been determined. The rat and human genes are comprised of at least 15 exons, excluding the alternatively spliced EGF-like exon in the case of the human gene.

Structure and evolution of the glycosaminoglycan attachment region of aggrecan

The chondroitin sulfate attachment region of the core protein of the large cartilage proteoglycan is encoded by a single large exon (exon 10) which is highly variable in

its size and sequence between species. This exon varies from 2856 bp in chicken[11] to 4223 bp in human[20]. This region is identified as the chondroitin sulfate attachment region by several criteria. First it is located at the appropriate region of the molecule and it contains large numbers of Ser-Gly pairs which are known to be the basic unit of the attachment region for xylosylation. Second, Krueger et al.[39] have isolated two peptides from this region of the chicken proteoglycan which are substituted and which they have subjected to amino acid sequencing before and following deglycosylation with hydrogen fluoride. They identified a number of unrecognized residues (substituted) prior to deglycosylation which were all confirmed to be serines either by amino acid sequencing following deglycosylation or by examination of the DNA sequence encoding the residues[11].

When the amino acid sequences encoded by exon 10 from various species are compared by dot plot analysis (fig. 2) patterns of similarities and differences emerge. Based on patterns of repeated sequences, the chondroitin sulfate attachment region of the rat proteoglycan (the first to be described) was divided into two regions named CS-1 and CS-2[16] (fig. 2A). Subsequent analysis of aggrecans of other species has shown the existence of at least two other tandem repeat domains within exon 10. The first, near the beginning of exon 10, in the region encoding the keratan sulfate-enriched domain, was identified from sequencing bovine cDNA clones[3] (fig. 2C), and the second, near the 3′-end of exon 10 was identified in chicken cDNA clones[11] (fig. 2B). Each of these repeated domains appears to be present in each of the four species although in some cases only a very few poorly conserved copies of the repeat are present, e.g., the KS and CS-1 domains in the chicken (see fig. 2B).

The first repeated sequence of exon 10, in the KS-rich domain, is present as about 23 consecutively repeated hexapeptides in bovine aggrecan and about 12 repeats in human aggrecan (fig. 2C). In rat and chick there are roughly 3–4 such repeats which are poorly conserved and can only be found by very careful alignment of the chicken or rat sequence with the bovine and human

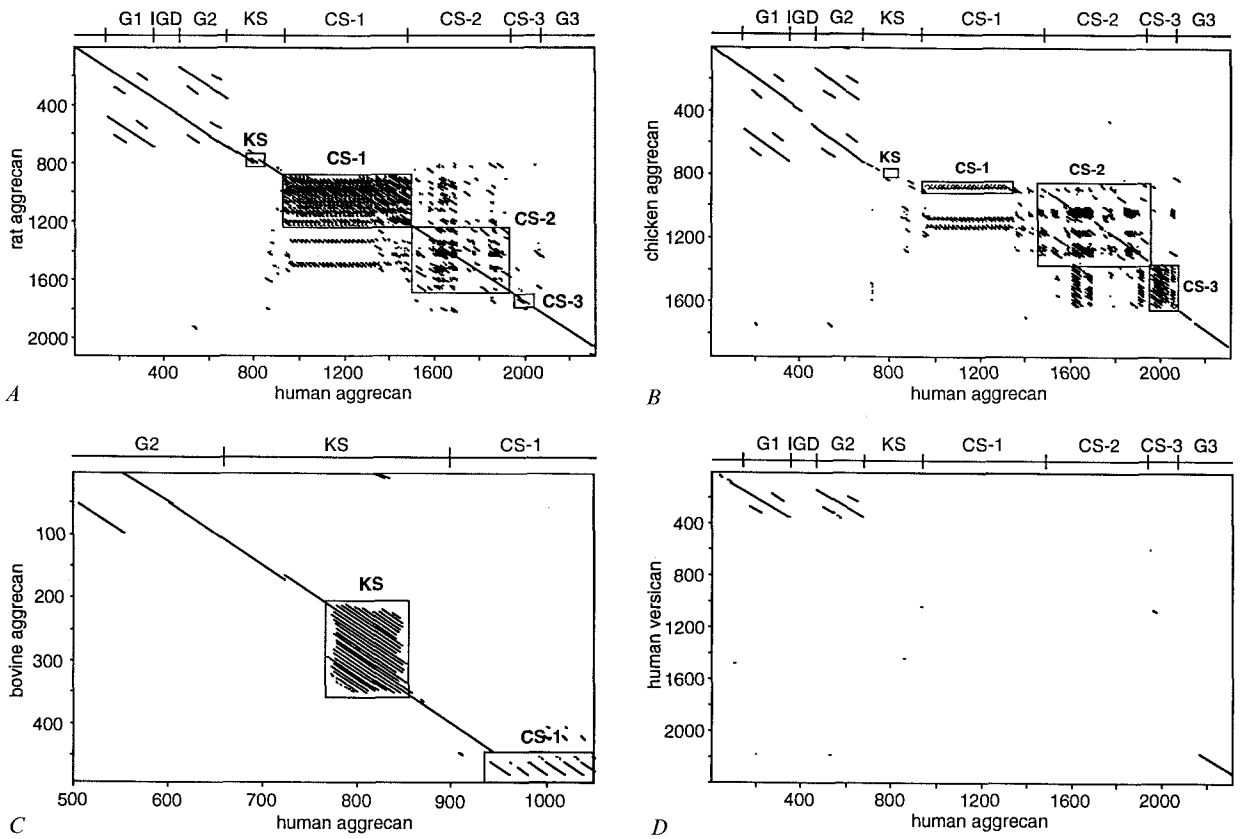

Figure 2. Dot plot analysis of the amino acid sequences of large chondroitin sulfate proteoglycan core proteins. Comparison of: *A* rat and human aggrecan sequences; *B* chicken and human aggrecan sequences; *C* bovine and human aggrecan sequences; *D* human versican and human aggrecan sequences. Dot plot analysis was done using MacVector Software with a window size of 35 and a minimum % score of 25. Numbering begins with the methionine start codon for all proteins except for the bovine in the bovine/human aggrecan comparison where only a portion of the sequences were compared since complete bovine cDNA sequence covering the entire coding sequence is not available. Regions corresponding to the KS, CS-1, CS-2 and CS-3 repeat domains are boxed. The various domains of human aggrecan are shown along the top of each panel. Sequences for comparison were obtained from the GenBank database: bovine[3,49]; chicken[11]; human[20]; rat[16]; human versican[62].

sequences. Based on its amino acid composition and position in the protein, it has been suggested that this repeat corresponds to the KS attachment sequence[3,20] although there is no direct evidence that KS is attached to this sequence. In fact the one putative KS-attachment site identified by Krueger et al.[39] is encoded by exon 9 (amino acid 728 of chick aggrecan) although the tryptic peptide which they identified as containing KS chains is encoded by parts of both exons 9 and 10 and does include the domain of chick aggrecan which corresponds to the repeated domain identified in the human and bovine sequences.

The CS-1 region consists of approximately 15 highly similar repeats of 20 amino acids in the rat and 29 repeats in the human. Although present in the bovine the number cannot be determined as the bovine cDNA is incomplete in this region (fig. 2C). In the human most of the repeats consist of 19 rather than 20 amino acids and 11 of the repeats are identical. Doege et al.[16] have suggested that these repeats have evolved from a primordial 10 amino acid sequence through a series of duplications, mutations, and amplifications. The chicken contains a residual CS-1 domain consisting of 4 nonuniformly spaced repeats of the sequence SGLPS (a part of the core sequence of the human and rat CS-1 single repeat unit) within a 40 amino acid span. Each of the SGLPS sequences is a part of a 10 amino acid repeat with 70–80% identity with one half of the CS-1 repeat.

The CS-2 domain appears to be essentially conserved in its entirety (but at a lower level of conservation as there is only approximately 50% amino acid sequence identity between the chicken domain and that of either the rat or human) in all 4 species and consists of 5–6 longer (approximately 100 amino acids) and more variable repeats. These repeats are also proposed to have arisen in part from the same basic 10 amino acid sequence as the CS-1 domain by a repeated series of amplification and fusion steps[16]. In the chicken, Chandrasekaran and Tanzer[11] have named this domain CS-1 based on its position (approximately amino acids 910–1400) which roughly corresponds to that of CS-1 in rat and human. Chicken aggrecan has an additional repeat sequence (CS-3 in fig. 2) for which no corresponding region was previously identified in rat, human, or bovine aggrecan. This region consists of a 20 amino acid sequence repeated 14 times (amino acids 1400–1630, named CS-2 in Chandrasekaran[11] based on its position) which is present only as a poorly conserved residual repeat of approximately 5 copies in the mammalian species (fig. 2B).

When the versican amino acid sequence[62] is compared with the aggrecan core protein sequence (fig. 2D) no similarity is seen in any portion of the CS-attachment region.

As mentioned above, Krueger et al.[39] have isolated two peptides from the CS-attachment region of the chicken proteoglycan which are substituted and which they have subjected to amino acid sequencing following deglycosylation with hydrogen fluoride. They identified a number of unrecognized residues (presumably substituted) which are all confirmed to be serines by examination of the cDNA sequence. In each case these serines are followed by glycines and all of the modified Ser-Gly pairs are separated from a second modified Ser-Gly pair by two amino acids. Five such doublets of Ser-Gly pairs are present in the sequence analyzed by Krueger et al.[39]. These five Ser-Gly doublets were aligned (fig. 3A) beginning two amino acids before the first Ser-Gly pair and extending two amino acids beyond the second Ser-Gly and analyzed using the PROFILEMAKE program of the GCG Sequence Analysis software, resulting in the consensus sequence of 10 amino acids shown in figure 3B. The primary characteristics of this sequence are a doublet of 2 Ser-Gly pairs separated by 2 amino acids and preceded by an acidic amino acid and a nonpolar amino acid. A length of 10 amino acids was chosen for analysis as the 5 doublets identified by Krueger et al.[39] occur as parts of two separately isolated tandem repeats with repeat lengths of 10 amino acids. If this profile sequence is used to search the entire chicken core protein sequence, 55 of the 94 Ser-Gly pairs in the entire gene are found in 34 such repeat units, all of which are encoded by exon 10. Examination of the CS-attachment regions of aggrecans of other species for which sequence is available (rat, human, and cow) shows that each of these species has related 10 amino acid repeats which include from 56–75% of the Ser-Gly repeats in the CS-region. All of the 10 amino acid Ser-Gly doublets from each species were aligned with the other 10 amino acid repeats from the same species and the frequencies of amino acids at each position of the 10 amino acid repeat were then compared between species. Since the repeat sequences in part appear to have evolved separately in each lineage through a repeated series of tandem duplications together with mutations (see below), the frequency of amino acids at any one position may be a species characteristic pattern rather than one necessary for the structure/function of the proteoglycan. This especially appears to be true for human aggrecan where 17 of the 55 ten amino acid repeats are identical (all a part of the highly conserved CS-1 repeat in human) and differ slightly from the consensus sequence seen in the other species. Examination of each of the species' patterns of amino acid frequencies in the 10 amino acid repeat gives a general consensus pattern (fig. 3C) which is common to all species. In general, sulfur-containing amino acids and polar amino acids other than serine and the acidic amino acids are underrepresented in the repeat sequence, compared to the rest of the aggrecan protein. Since the 2 Ser-Gly pairs separated by 2 amino acids

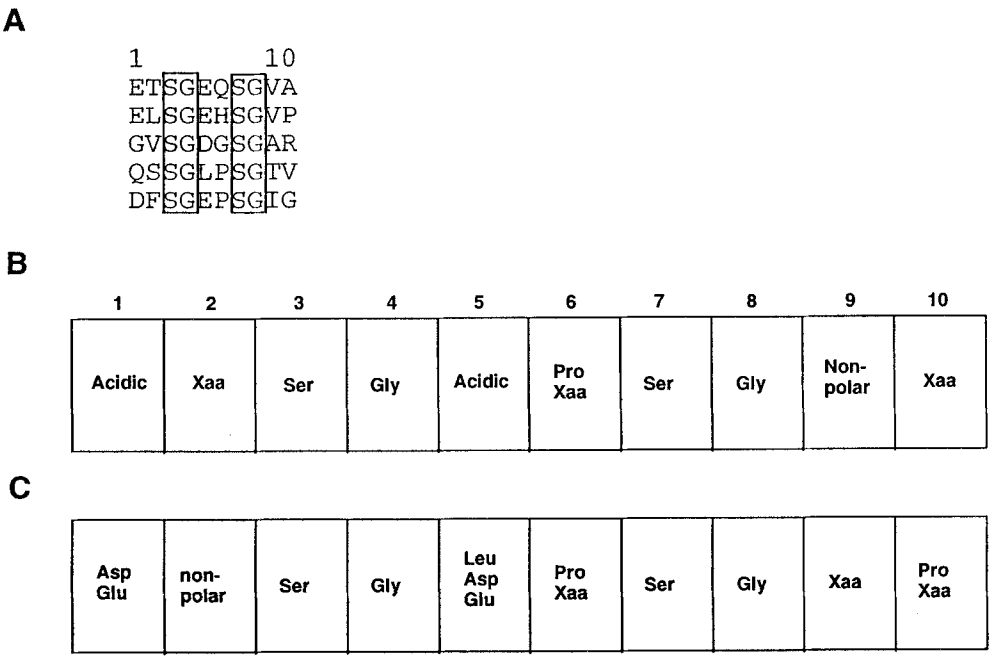

Figure 3. Consensus sequence for the 10 amino acid repeat sequence containing 2 Ser-Gly pairs. *A* Sequences identified by Krueger et al.[40] as being modified in the chick aggrecan core protein. *B* Consensus sequence derived from *A* and used to search sequences for similar repeats. Using appropriate criteria, all decamers with 2 Ser-Gly pairs separated by 2 amino acids were selected. *C* Summary consensus sequence derived from bovine, chicken, human and rat decamer repeat sequences.

were the criteria used to select the sequences for comparison these positions are invariant. The 1st 2 amino acids preferentially are an acidic amino acid at position 1 followed by a nonpolar amino acid at position 2. However, based on the amino acid sequences of the known attachment sites of the CS-enriched region identified by Krueger et al.[39] and the frequencies found in the aligned repeats, it is likely that essentially any amino acid can occur in either position. The position following the 1st Ser-Gly pair is predominantly a leucine or an acidic residue followed by a proline in position 6. With the exception of human aggrecan where 47% of the residues at position 9 are acidic (all a part of the highly conserved repeats in the CS-1 domain), there seems to be little preference at this position. Proline is the predominant amino acid in position 10 with the exception of the human gene again where this residue is predominantly valine, essentially all of which occur in the CS-1 repeat domain. Although the repeat sequence presumably contains the consensus sequence for xylosylation of the cartilage chondroitin sulfate proteoglycan, it has been identified as the consensus sequence representing the 10 amino acid repeat present in these core proteins. It is clear by examination of the sequence of chondroitin sulfate proteoglycans other than aggrecan, such as versican[62] or the rat membrane-spanning proteoglycan NG2[46], and from peptides used as xylosyl transferase substrates in vitro[7] that the SGXXSG doublet is not absolutely required for attachment of xylose to serine. However since the 10 amino acid repeat is conserved in aggrecan in all species examined, its conformation may be important for the proper attachment or function of chondroitin sulfates chains in this particular proteoglycan.

The existence of a doublet of two closely spaced Ser-Gly pairs separated by less than 6 amino acids was first predicted by Mathews[43] based on his analysis of the composition and size of tryptic-chymotryptic hydrolysates of proteoglycans from ox, sturgeon, and ray cartilage. Additional biochemical and electron-microscopic data were obtained by Heinegård's laboratory[34,58], supporting the clustering of polysaccharide chains in cartilage proteoglycan.

The 10 amino acid repeat sequence with 2 Ser-Gly pairs separated by two amino acids is one of the basic units of which the CS-1 and CS-2 repeat domains are composed, first described for rat[16] and bovine aggrecans[49] and subsequently for the human[20] and chicken aggrecans[11,40]. The 20 amino acid CS-3 repeat domain in the chicken core protein is related but has an alanine in place of the glycine of the 2nd Ser-Gly pair. For this reason that sequence was not included in the derivation of the consensus sequences above although it may well serve as a site for CS substitution.

Prior to, interspersed within, and following the tandem repeat domains are additional copies of the basic 10 amino acid repeat as well as incomplete remnants of that repeat. These data suggest that the entire exon may have evolved from tandem duplication of the basic 30 nucleotide sequence encoding the 10 amino acid repeat

followed by repeated subsequent single site mutagenesis and by repeated duplications, expansions, contractions, and corrections of the repeat sequences through out-of-register homologous recombination and gene conversion involving sister chromatids.

There are several other genes which have been identified which also have large or very large exons containing extensive tandem repeats. These genes include those for *Bombyx mori* silk fibroin[26], profilaggrin[27], the rat ventral prostate proline-rich polypeptides[12] and the salivary gland proline-rich proteins[2]. These exons vary in length from about 1 kb for the mouse PRP gene M14 to greater than 14 kb for the profilaggrin and silk fibroin genes. In these genes variations in the number of tandem repeats are seen between inbred stocks of *Bombyx mori*[26], between different closely related genes of the mouse proline-rich gene family[2], or in different alleles in the population for the human profilaggrin[27] and the rat ventral prostate proline-rich polypeptide genes[12]. Ann et al.[2] and De Clercq et al.[12] have extensively discussed hypotheses for the evolutionary history and significance of the repetitive domains of these proteins. These hypotheses involve mechanisms of duplication, amplification, and contraction of tandem repeats through unequal crossing over and gene conversion. These latter processes can result in concerted evolution of gene families and the maintenance of high sequence identity among individual members of a tandem repeat sequence. It is likely that the mechanisms of evolution of these genes are similar to that of exon 10 of aggrecan. These genes are very different from other genes encoding long repetitive sequences such as those for the interstitial collagens. The repeated Gly-Xaa-Yaa domains of the interstitial collagens are encoded by a large number of small exons most of which vary from 45 to 108 bp. Furthermore, there is essentially no variation in the number of Gly-Xaa-Yaa repeats in these collagens. These differences are consistent with the hypothesis of Alexander et al.[1] that one of the functions of intron interruptions in genes is to 'limit sequence amplification in genes . . . whose protein products cannot tolerate size variation'. By extension of this hypothesis, genes for products for which the precise length or number of repeats are relatively unimportant may more likely be encoded by larger exons. These would be more efficient in conserving the extent of the gene in the genome, and in the efficiency of RNA processing. The presence of multiple repeats within a single large exon would also provide for a greater rate of evolution of the gene. The structure of exon 10 of the core protein gene is also consistent with the proposal of Ohno[48] that 'oligomers (are) the primordial coding sequences of all polypeptide chains'. Ann et al.[2] have suggested that 'it seems plausible that genes created recently would be the most likely to retain evidence of primordial repeats', which is particularly relevant for the mouse salivary proline-rich protein genes which are members of a family of closely related similarly controlled genes[2].

The extensive differences between exon 10 of aggrecan in the various species are consistent with the above hypotheses regarding the role of repeat sequences in evolution and the role of introns in limiting amplification. It would appear that ongoing extensive rounds of reiterative amplification, mutation and contraction might have occurred in each of the individual lineages following their separation, thereby giving rise to tandem repeats in each of the species. These repeats may be very homogeneous within a single species such as the 11 identical 19 amino acid repeats in the human CS-1 domain yet may differ both with respect to sequence and with respect to copy number in other species. These differences could account for the fairly wide range of sizes of exon 10 in the different species. Nevertheless, the variation in the number of Ser-Gly pairs in SGXXSG doublets does not vary to the same extent as the variation in size of the exon. In the case of the chicken, which has the smallest exon 10, there is a higher density of SGXXSG doublets. In some regions of the chicken gene amino acids of the tandem basic repeat of 10 amino acids have been lost resulting in 2–3 overlapping 10 amino acid repeats.

Conjectures concerning aggrecan structure and biosynthesis

Aggrecan is a large, complex multi-domain macromolecule whose major biological function is to impart resiliency to cartilage. The ability of aggrecan to impart resiliency reflects its three-dimensional structure; that structure emerges from both its underlying protein backbone and its carbohydrate substituents. Those carbohydrate substituents, both N-linked oligosaccharides and O-linked glycosaminoglycans (GAGs), are added to the protein as it traverses the interior of the cell en route to the cell surface, where the finished molecule is then deposited into the extracellular milieu.

Multi-domain proteins

Current opinion is that there are a limited number of unique protein domains (motifs, regions) which are utilized in the formation of large, multi-domain proteins[5,23,24,50]. At the gene level, evolutionary processes have provided for exon shuffling of the motifs, enabling proteins to be perfected and for new proteins to emerge, in order to fulfill specific tissue requirements[21]. The extracellular matrix evolved with the advent of multicellular organisms and, in the case of animals, sponges represent the earliest multicellular ancestors. Sponges have a well-developed matrix, composed of collagen and proteoglycans[35,44]. In higher animals, proteoglycans are ubiquitous in tissues and organs with cartilaginous tissues being especially rich in

proteoglycans. The most abundant proteoglycan of cartilage, aggrecan, notably has six different domains (G1, IGD, G2, KS, CS, G3), and each one is a specialized, unique entity and also includes subdomains (fig. 1). Thus, aggrecan is a prototypical example of a complex multi-domain protein, one which has evolved to fulfill the role of imparting resiliency to cartilage. The IGD and G2 domains have unknown function and the physiological ligand recognized by the lectin part of the G3 domain is also unknown.

Aggrecan structure

The globular domains of aggrecan, G1, G2, and G3 most likely have very organized structures, as indicated by their electron microscopic appearance[14,51], their lectin-binding properties[25,30], their neutron and X-ray scattering patterns[52], and their predicted secondary structures[15]. It is less certain that the remaining domains, IGD, KS and CS have well-organized secondary structures but experimental data are needed to determine this point. The overall structure of aggrecan is supported by gene cloning studies in several species. Although there are substantial differences in the size of specific domains, comparing species, the linear, tandem arrangement of domains in aggrecans is identical. There are also differences in the amino acid patterns of individual domains, comparing species, but those differences seem to be variations on a basic theme. For example, the avian KS domain does not have a hexapeptide repeat seen in mammalian aggrecan. Nevertheless, all aggrecans appear designed to serve the same physiological purpose.

Aggrecan biosynthesis

There are at least two major events which occur during the post-translational processes which lead to a mature aggrecan molecule[32]. One event is that, following removal of its signal peptide, aggrecan must undergo correct folding, i.e., each individual domain must adopt its characteristic conformation. Concurrently, posttranslational modification of encoded consensus sequences occurs, with initiation of N-linked oligosaccharides and O-linked GAG chains. Both types of substituents must achieve their mature forms prior to aggrecan's emergence from the cell. Whether recognition processes for carbohydrate addition involve more than primary structure determinants, e.g., secondary structure as well, is not clear. One hint is that not all consensus sites, either N- or O-linked become substituted[6,39], implying that conformation is also a factor in determining such modifications. Conversely, it is not known whether glycosylation affects domain conformation.

Aggrecan folding may also be mediated by intracellular chaperon proteins, as found for other nascent proteins[28]. If so, one possibility is that each domain may have its own chaperon or repertoire of chaperons. Such accessory proteins could also help mediate the temporal and spatial relationships involved in carbohydrate addition, consensus site recognition and progressive folding into the mature conformation. If, as postulated, each domain develops into its mature form independent of its neighboring domains, overall aggrecan structure does not require correlation and integration of folding processes between neighboring domains.

Aggrecan biosynthesis in avian embryonic chondrocytes occurs in specific intracellular pre-Golgi loci, compared to type II collagen[60]. This observation may contrast with results seen for other types of proteoglycans in other cell types[9]. However, direct comparison cannot be made because these data were from kinetic experiments rather than from immunocytological analyses. At any rate, segregation of nascent aggrecan within chondrocytes may reflect the need for folding of multiple domains as well as for initiation and maturation of numerous GAG chains. Segregation may also reflect the postulated role of the G3 domain in signalling intracellular routing. This postulate is supported by data from nanomelic chickens (see below). Interestingly, in avian embryonic chondrocytes the intracellular processes occur very rapidly, yielding a finished aggrecan molecule within 15–20 min[10]. In contrast, rat chondrosarcoma cells yield a completed aggrecan molecule after 60–80 min, perhaps due to lower levels of galactosyltransferase I in these cells[10].

Potentially, the nanomelic mutation of chickens will provide insight into the routing process. The mutant cartilage contains $<5\%$ of normal proteoglycans but has a normal level of type II collagen[29]. The genetic lesion seems to be restricted to aggrecan and genetic linkage of the nanomelic phenotype to the aggrecan gene has been made[59]. Nanomelic aggrecan lacks GAG chains and also lacks the G3 domain but it is substituted by immature N-linked oligosaccharides[47]. Moreover, nanomelic chondrocytes retain the capacity to synthesize chondroitin sulfate chains, consistent with the idea that nanomelic aggrecan is not properly routed within the cell. Although the general consensus sequence for O-glycosylation of serine by xylose has been postulated to be acidic-acidic-Xxx-Ser-Gly-Xxx-Gly, based on in vitro enzymatic assays[37], transfection data indicate that threonine can partially substitute for serine and that conformation of the glycosylation site may be of critical importance[42]. In the case of KS glycosylation of aggrecan, less information is available about putative consensus sequences. Chicken aggrecan differs from mammalian aggrecans in lacking a hexapeptide repeat sequence in its KS domain, a domain which is known to contain KS chains[13,33,39]. Clearly, more data are needed concerning the nature of both CS and KS consensus sequences, especially data obtained from cellular experiments.

Recent information, obtained using the Golgi inhibitor Brefeldin A, clearly demonstrates that CS chain initiation occurs prior to the trans-Golgi network, while chain elongation and sulfation are associated with that network[56]. These results, obtained in a human melanoma cell line and a human B-lymphoblastoid cell line, are comparable to CS formation in cartilage cells[31,32,36,38,61]. It is now generally accepted that Golgi-mediated processes are universal, even comparing yeast cells and animal cells[45,53]. Indeed, there appear to be endogenous proteins of the Golgi complex which are common to a wide spectrum of eukaryotic cells[8].

In summary, as a protein with multiple domains, potentially with unique conformations, plus extensive post-translational modifications, aggrecan provides an unusual opportunity for correlating its domain structures with its pathway of biosynthesis. Such correlation is possible by using contemporary methods of structure analysis in conjunction with cellular transfection methods. The composite results should provide substantial insight into the molecular anatomy and biogenesis of aggrecan. By analogy, such information may aid in understanding other proteoglycans, especially those which share comparable domains with aggrecan.

1 Alexander, F., Young, P. R., and Tilghman, S. M., Evolution of the albumin: α-fetoprotein ancestral gene from the amplification of a 27 nucleotide sequence. J. molec. Biol. *173* (1984) 159–176.

2 Ann, D. K., Smith, M. K., and Carlson, D. M., Molecular evolution of the mouse proline-rich protein multigene family: Insertion of a long interspersed repeated DNA element. J. biol. Chem. *263* (1988) 10887–10893.

3 Antonsson, P., Heinegård, D., and Oldberg, Å., The keratan-sulfate-enriched region of bovine cartilage proteoglycan consists of a consecutively-repeated hexapeptide motif. J. biol. Chem. *264* (1989) 16170–16173.

4 Baldwin, C. T., Reginato, A. M., and Prockop, D. J., A new epidermal growth factor-like domain in the human core protein for the large cartilage-specific proteoglycan. J. biol. Chem. *264* (1989) 15747–15750.

5 Baron, M., Norman, D. G., and Campbell, I. D., Protein modules. Trends biochem. Sci. *16* (1991) 13–17.

6 Barry, F. P., Gaw, J. U., Young, C. N., and Neame, P. J., Hyaluronan binding region of aggrecan from pig laryngeal cartilage. Amino acid sequence, analysis of N-linked oligosaccharides and location of the keratan sulfate. Biochem. J. *286* (1992) 761–769.

7 Bourdon, M. A., Krusius, T., Campbell, S., Schwartz, N. B., and Ruoslahti, E., Identification and synthesis of a recognition signal for the attachment of glycosaminoglycans to proteins. Proc. natl Acad. Sci. USA *84* (1987) 3194–3198.

8 Brändli, A. W., Mammalian glycosylation mutants as tools for the analysis and reconstitution of protein transport. Biochem. J. *276* (1991) 1–12.

9 Brion, C., Miller, S. G., and Moore, H.-P. H., Regulated and constituitive secretion. Differential effects of protein synthesis arrest on transport of glycosaminoglycan chains to the two secretory pathways. J. biol. Chem. *267* (1992) 1477–1483.

10 Campbell, S. C., and Schwartz, N. B., Kinetics of intracellular processing of chondroitin sulfate proteoglycan core protein and other matrix components. J. Cell Biol. *106* (1988) 2191–2202.

11 Chandrasekaran, L., and Tanzer, M. L., Molecular cloning of chicken aggrecan: structural analyses. Biochem. J. *288* (1992) 903–910.

12 De Clercq, N., Hemschoote, K., Devos, A., Peeters, B., Heyns, W., and Rombauts, W., The 4.4 kilodalton proline-rich polypeptides of the rat ventral prostate are the proteolytic products of a 637-kilodalton protein displaying highly repetitive sequences and encoded in a single exon. J. biol. Chem. *267* (1992) 9884–9894.

13 De Luca, S., Lohmander, L. S., Nilsson, B., Hascall, V. C., and Caplan, A. I., Proteoglycans from chick limb bud chondrocyte cultures. Keratan sulfate and oligosaccharides which contain mannose and sialic acid. J. biol. Chem. *255* (1980) 6077–6083.

14 Dennis, J. E., Carrino, D. A., Schwartz, N. B., and Caplan, A. I., Ultrastructural characterization of embryonic chick cartilage proteoglycan core protein and the mapping of a monoclonal antibody epitope. J. biol. Chem. *265* (1990) 12098–12103.

15 Doege, K., Rhodes, C., Sasaki, M., Hassell, J. R., and Yamada, Y., Molecular biology of cartilage proteoglycan (aggrecan) and link protein, in: Extracellular Matrix Genes, pp. 137–155. Eds L. J. Sandell and C. D. Boyd. Academic Press, New York, 1990.

16 Doege, K., Sasaki, M., Horigan, E., Hassell, J. R., and Yamada, Y., Complete primary structure of the rat cartilage proteoglycan core protein deduced from cDNA clones. J. biol. Chem. *262* (1987) 17757–17767.

17 Doege, K., Sasaki, M., and Yamada, Y., Rat and human cartilage proteoglycan (aggrecan) gene structure. Biochem. Soc. Trans. *18* (1990) 200–202.

18 Doege, K., and Yamada, Y., Gene structure of the rat cartilage proteoglycan core protein. Collagen rel. Res. *8* (1988) 486.

19 Doege, K., Fernandez, P., Hassell, J. R., Sasaki, M., and Yamada, Y., Partial cDNA sequence encoding a globular domain at the C terminus of the rat cartilage proteoglycan. J. biol. Chem. *261* (1986) 8108–8111.

20 Doege, K., Sasaki, M., Kimura, T., and Yamada, Y., Complete coding sequence and deduced primary structure of the human cartilage large aggregating proteoglycan, aggrecan. Human specific repeats, and additional alternatively spliced forms. J. biol. Chem. *266* (1991) 894–902.

21 Dorit, R. L., and Gilbert, W., The limited universe of exons. Current Opinion Struct. Biol. *1* (1991) 973–977.

22 Dudhia, J., and Hardingham, T. E., cDNA sequences to human and porcine cartilage proteoglycan and link protein. Trans. Orthopaed. Res. Soc. *14* (1989) 8.

23 Engel, J., Common structural motifs in proteins of the extracellular matrix. Current Opinion Cell Biol. *3* (1991) 779–785.

24 Engel, J., Domains in proteins and proteoglycans of the extracellular matrix with functions in assembly and cellular activities. Int. J. Biol. Macromolec. *13* (1991) 147–151.

25 Fosang, A. J., and Hardingham, T. E., Isolation of the N-terminal globular protein domains from cartilage proteoglycans. Identification of the G2 domain and its lack of interaction with hyaluronate and link protein. Biochem. J. *261* (1989) 801–809.

26 Gage, L. P., and Manning, R. F., Internal structure of the silk fibroin gene of *Bombyx mori* I. The fibroin gene consists of a homogeneous alternating array of repetitious crystalline and amorphous coding sequences. J. biol. Chem. *255* (1980) 9444–9450.

27 Gan, S.-Q., McBride, O. W., Idler, W. W., Markova, N., and Steinert, P. M., Organization, structure, and polymorphisms of the human profilaggrin gene. Biochemistry *29* (1990) 9432–9440.

28 Gething, M.-J., and Sambrook, J., Protein folding in the cell. Nature *355* (1992) 33–45.

29 Goetinck, P. F., Studies of the avian chondrodysplasia mutant, nanomelia. Path. Immunopath. Res. *7* (1988) 73–75.

30 Halberg, D. F., Proulx, G., Doege, K., Yamada, Y., and Drickamer, K., A segment of the cartilage proteoglycan core protein has lectin-like activity. J. biol. Chem. *263* (1988) 9486–9490.

31 Hardingham, T. E., and Fosang, A. J., Proteoglycans: many forms and many functions. FASEB J. *6* (1992) 861–870.

32 Hascall, V. C., Heinegård, D. K., and Wight, T.N., Proteoglycans: Metabolism and pathology, in: Cell Biology of Extracellular Matrix, pp. 149–175. Ed E. D. Hay. Plenum Press, New York 1991.
33 Haynesworth, S. E., Carrino, D. A., and Caplan, A. I., Characterization of the core protein of the large chondroitin sulfate proteoglycan synthesized by chondrocytes in chick limb bud cell cultures. J. biol. Chem. *262* (1987) 10574–10581.
34 Heinegård, D., and Hascall, V. C., Characterization of chondroitin sulfate isolated from trypsin-chymotrypsin digest of cartilage proteoglycans. Archs Biochem. Biophys. *165* (1974) 427–441.
35 Hunt, S., Evolution of Mucopolysaccharides and Related Molecules. Polysaccharide-Protein Complexes in Invertebrates. Academic Press, New York 1970.
36 Jackson, R. L., Busch, S. J., and Cardin, A. D., Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes. Physiol. Revs *71* (1991) 481–539.
37 Kearns, A. E., Campbell, S. C., Westley, J., and Schwartz, N. B., Initiation of chondroitin sulfate biosynthesis: a kinetic analysis of UDP-D-xylose:core protein β-D-xylotransferase. Biochemistry *30* (1991) 7477–7483.
38 Kjellén, L., and Lindahl, U., Proteoglycans: structure and interactions. A. Rev. Biochem. *60* (1991) 443–475.
39 Krueger, R. C. Jr, Fields, T. A., Hildreth, J. IV, and Schwartz, N. B., Chick cartilage chondroitin sulfate proteoglycan core protein. I. Generation and characterization of peptides and specificity for glycosaminoglycan attachment. J. biol. Chem. *265* (1990) 12075–12087.
40 Krueger, R. C. Jr, Fields, T. A., Mensch, J. R. Jr, and Schwartz, N. B., Chick cartilage chondroitin sulfate proteoglycan core protein. II. Nucleotide sequence of cDNA clone and localization of the S103L epitope. J. biol. Chem. *265* (1990) 12088–12097.
41 Krusius, T., Gehlsen, K. R., and Ruoslahti, E., A fibroblast chondroitin sulfate proteoglycan core protein contains lectin-like and growth factor-like sequences. J. biol. Chem. *262* (1987) 13120–13125.
42 Mann, D. M., Yamaguchi, Y., Bourdon, M. A., and Ruoslahti, E., Analysis of glycosaminoglycan substitution in decorin by site-directed mutagenesis. J. biol. Chem. *265* (1990) 5317–5323.
43 Mathews, M. B., Comparative biochemistry of chondroitin sulfate-proteins of cartilage and notochord. Biochem. J. *125* (1971) 37–46.
44 Mathews, M. B., Polyanionic Glycans of Other Tissues. Connective Tissue. Macromolecular Structure and Evolution. Springer-Verlag, New York 1975.
45 Mellman, I., and Simons, K., The Golgi complex: in vitro veritas? Cell *68* (1992) 829–840.
46 Nishiyama, A., Dahlin, K. J., and Stallcup, W. B., The expression of NG2 proteoglycan in the developing rat limb. Development *111* (1991) 933–944.
47 O'Donnell, C. M., Kaczman-Daniel, K., Goetinck, P. F., and Vertel, B. M., Nanomelic chondrocytes synthesize a glycoprotein related to chondroitin sulfate proteoglycan core protein. J. biol. Chem. *263* (1988) 17749–17754.
48 Ohno, S., Repeats of base oliogmers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. J. molec. Evol. *20* (1984) 313–321.
49 Oldberg, Å., Antonsson, P., and Heinegård, D., The partial amino acid sequence of bovine cartilage proteoglycan, deduced from a cDNA clone, contains numerous Ser-Gly sequences arranged in homologous repeats. Biochem. J. *243* (1987) 255–259.
50 Patthy, L., Modular exchange principles in proteins. Current Opinion Struct. Biol. *1* (1991) 351–361.
51 Paulsson, M., Mörgelin, M., Wiedemann, H., Beardmore-Gray, M., Dunham, D., Hardingham, T., Heinegård, D., Timpl, R., and Engel, J., Extended and globular protein domains in cartilage proteoglycans. Biochem. J. *245* (1987) 763–772.
52 Perkins, S. J., Nealis, A. S., Dunham, D. G., Hardingham, T. E., and Muir, I. H., Molecular modeling of the multidomain structures of the proteoglycan binding region and the link protein of cartilage by neutron and synchrotron X-ray scattering. Biochemistry *30* (1991) 10708–10716.
53 Rothman, J. E., and Orci, L., Molecular dissection of the secretory pathway. Nature *355* (1992) 409–415.
54 Sai, S., Tanaka, T., Kosher, R. A., and Tanzer, M. L., Cloning and sequence analysis of partial cDNA for chicken cartilage proteoglycan core protein. Proc. natl Acad. Sci. USA *83* (1986) 5081–5085.
55 Sparks, K. J., Lever, P. L., and Goetinck, P. F., Antibody binding of cartilage-specific proteoglycans. Archs Biochem. Biophys. *199* (1980) 579–590.
56 Spiro, R. C., Freeze, H. H., Sampath, D., and Garcia, J. A., Uncoupling of chondroitin sulfate glycosaminoglycan synthesis by Brefeldin A. J. Cell Biol. *115* (1991) 1463–1473.
57 Tanaka, T., Har-El, R., and Tanzer, M. L., Partial structure of the gene for chicken cartilage proteoglycan core protein. J. biol. Chem. *263* (1988) 15831–15835.
58 Thyberg, J., Lohmander, S., and Heinegård, D., Proteoglycans of cartilage: Electron-microscopic studies on isolated molecules. Biochem. J. *151* (1975) 157–166.
59 Velleman, S. G., and Clark, S. H., The cartilage proteoglycan deficient mutation, nanomelia, contains a DNA polymorphism in the proteoglycan core protein gene that is genetically linked to the nanomelia phenotype. Matrix *12* (1992) 66–72.
60 Vertel, B. M., Velasco, A., LaFrance, S., Walters, L., and Kaczman-Daniel, K., Precursors of chondroitin sulfate proteoglycan are segregated within a subcompartment of the chondrocyte endoplasmic reticulum. J. Cell Biol. *109* (1989) 1827–1836.
61 Wight, T. N., Heinegard, D. N., and Hascall, V. C., Proteoglycans: Structure and function, in: Cell Biology of Extracellular Matrix, pp. 45–78. Ed. E. D. Hay. Plenum Press, New York 1991.
62 Zimmermann, D. R., and Ruoslahti, E., Multiple domains of the large fibroblast proteoglycan, versican. EMBO J. *8* (1989) 2975–2981.