The partial amino acid sequence of bovine cartilage proteoglycan, deduced from a cDNA clone, contains numerous Ser-Gly sequences arranged in homologous repeats

Åke OLDBERG, Per ANTONSSON and Dick HEINEGÅRD

Department of Medical and Physiological Chemistry, University of Lund, P.O. Box 94, S-221 00 Lund, Sweden

We have determined the sequence of a partial cDNA clone encoding the C-terminal region of bovine cartilage aggregating proteoglycan core protein. The deduced amino acid sequence contains a cysteine-rich region which is homologous with chicken hepatic lectin. This lectin-homologous region has previously been identified in rat and chicken cartilage proteoglycan. The bovine sequence presented here is highly homologous with the rat and chicken amino acid sequences in this apparently globular region. A region containing clusters of Ser-Gly sequences is located N-terminal to the lectin homology domain. These Ser-Gly-rich segments are arranged in tandemly repeated, approx. 100-residue-long, homology domains. Each homology domain consists of an approx. 75-residue-long Ser-Gly-rich region separated by an approx. 25-residue-long segment lacking Ser-Gly dipeptides. These dipeptides are arranged in 10-residue-long segments in the 100-residue-long homology domains. The shorter homologous segments are tandemly repeated some six times in each 100-residue-long homology domain. Serine residues in these repeats are potential attachment sites for chondroitin sulphate chains.

INTRODUCTION

Proteoglycans composed of a core protein and covalently attached glycosaminoglycans serve as structural components in the extracellular matrix of tissues (for review see Heinegård & Paulsson, 1984). The major cartilage proteoglycan is built from a protein core of M_r 340 000 (Upholt *et al.*, 1979) substituted with some 100 chondroitin sulphate chains, some 30 keratan sulphate chains, some 40 O-glycosidically linked oligosaccharides and a few N-glycosidically linked oligosaccharides (Heinegård *et al.*, 1985a). The chondroitin sulphate chains are attached to serine residues of the core protein and available data indicates that a Ser-Gly dipeptide provides the recognition sequence for xylosyl transferase, providing the first step in polysaccharide chain elongation (Coudron *et al.*, 1980). The keratan sulphate chains and the O-glycosidically linked oligosaccharides are attached to serine or threonine, but no recognition sequence for the N-acetylgalactosaminyl transferase has been identified.

The protein core contains several substructures. The N-terminal portion is not substituted with glycosaminoglycan chains but forms a structure allowing specific interaction with hyaluronate (Heinegård & Hascall, 1974a). This part of the molecule can be visualized upon electron microscopy of rotary shadowed proteoglycans. It contains two globular domains separated by a short filament (Wiedemann *et al.*, 1984). Only the first globule at the N-terminal end interacts with hyaluronate, while the function of the second globule is not known. The other portions of the proteoglycan also form distinct substructures. The keratan sulphate chains are enriched in a domain containing an abundance of glutamic acid/glutamine and proline residues near the second globule. The chondroitin sulphate chains are concentrated in a serine- and glycine-rich domain occupying the

major portion of the protein core (Heinegård & Axelsson, 1977).

Proteoglycans isolated from cartilage are heterogeneous in size, presumably as a result of extracellular proteolytic degradation of the core proteins (Heinegård *et al.*, 1986). Interestingly, a proteoglycan subpopulation having the largest core protein contains a C-terminal globular structure identified by electron microscopy (Paulsson *et al.*, 1987). Only a minor proportion of the proteoglycans in nasal cartilage from a 2-year-old cow contain this C-terminal globule. It appears possible that the C-terminal part of the proteoglycan may be cleaved in a slow progressive process, probably in the matrix. In support, labelling experiments indicates that there is a slow transition from a larger proteoglycan to a somewhat smaller molecule (Heinegård *et al.*, 1986). It should be stressed, however, that all proteolytically degraded proteoglycans are distinct from the group of small proteoglycans also present in cartilage (Heinegård *et al.*, 1985b).

The C-terminal globular structure has recently been identified by cDNA cloning and sequencing, using mRNA from rat chondrosarcoma (Doege *et al.*, 1986) and chicken chondrocytes (Sai *et al.*, 1986). The deduced amino acid sequences are homologous and both contain a stretch of amino acids with homology with chicken liver lectin. This lectin binds oligosaccharides containing a non-reducing terminal N-acetylglucosamine (Sikder *et al.*, 1983). The two cDNA clones, which were 0.8 kb and 1.2 kb, hardly extended into the region of the core protein substituted with chondroitin sulphate chains, since they contained few Ser-Gly sequences. The present work was initiated to determine the structure of the C-terminal globular domain and to reveal the structural features of the chondroitin sulphate-rich region.

MATERIALS AND METHODS

Affinity purification of antibodies

Proteoglycan monomers, i.e. A1-D1, were prepared from tracheal cartilage of a 2 month calf fetus by sequential equilibrium centrifugation in CsCl without and with 4 M-guanidinium chloride as described elsewhere (Heinegård *et al.*, 1985a). In separate experiments it has been shown that proteoglycans in this preparation contain the C-terminal globule. Rabbits were immunized (Heinegård *et al.*, 1985b) and antibodies were purified by a series of immunoadsorption steps. The rationale for the approach is that immunological cross-reactivity between the hyaluronate-binding region from bovine and rat sources is very poor with rabbit polyclonal antibodies (Wieslander & Heinegård, 1980). Furthermore, the second globule in the N-terminal region is present in a fragment containing the keratan-sulphate-rich region isolated after trypsin digestion of proteoglycan aggregates as described elsewhere (A1-T-A1 in Heinegård & Axelsson, 1977). This structure appears to show extensive cross-reactivity between species. It was thus considered that antibodies to the second globule could be adsorbed by passing the antiserum through an affinity column with bound keratan sulphate-rich fragment. Antibodies to structures in the chondroitin sulphate-rich region, as well as in the C-terminal third globule, could then be adsorbed to an affinity column with bound chondroitinase-treated rat chondrosarcoma proteoglycan monomer. Because of their poor cross-reactivity, antibodies to the hyaluronate-binding region do not bind to this affinity matrix.

Rat chondrosarcoma proteoglycan monomer was prepared as described above for the bovine proteoglycan. After chondroitinase digestion and purification by chromatography on Sepharose 6B as described elsewhere (Hascall & Heinegård, 1974), the recovered core preparation was coupled to CNBr-activated Sepharose (Pharmacia Biotechnology) according to the description of the manufacturer. The keratan sulphate-rich fragment was coupled similarly. Columns (1.8 ml) containing 5 mg of the respective antigen coupled to Sepharose were packed in BioRad disposable columns. The antiserum (5 ml) was sequentially passed through the keratan sulphate-peptide matrix and the rat chondrosarcoma core protein matrix. Unbound material was eluted from the column containing immobilized core protein with 50 ml of 0.15 M-NaCl in 5 mM-sodium phosphate, pH 7.4. Bound material was eluted with 3 M-KSCN in 5 mM-sodium phosphate, pH 7. Fractions containing material with absorbance at 280 nm were immediately pooled and dialysed against 0.15 M-NaCl/5 mM-sodium phosphate, pH 7.4. This affinity-purified antibody was subsequently used to screen the cDNA library for clones containing epitopes enriched in the C-terminal portion of the proteoglycan.

Preparation and sequencing of cDNA clones

RNA was purified from chondrocytes which were isolated from bovine articular cartilage as previously described (Sommarin & Heinegård, 1983). The methods used for construction and screening of the λ gt11 library, RNA-transfer blot analysis and nucleic acid sequence analysis have been described (Oldberg *et al.*, 1986).

RESULTS AND DISCUSSION

Identification of cDNA clones

A bovine chondrocyte λ gt11 cDNA library was prepared and screened with an affinity-purified antibody against the large aggregating chondroitin sulphate proteoglycan (PG-LA). Screening of approx. 100 000 plaques resulted in the identification of the clone λ PG-LA 5. This 800 bp-long cDNA hybridized to a 12 kb mRNA in transfer blot analysis of poly(A)-containing RNA from bovine chondrocytes (results not shown). This mRNA size is consistent with a PG-LA core protein of estimated M_r 340 000 (Upholt *et al.*, 1979). To obtain longer cDNA clones, allowing more amino acid sequence information from the C-terminal region of the core protein to be obtained, we decided to screen the cDNA library with λ PG-LA 5 as the probe. This led to the isolation of a 2.3 kb cDNA clone. This clone, λ PG-LA 51, hybridized with an identical 12 kb mRNA species as λ PG-LA 5 in transfer blot analysis of chondrocyte RNA, indicating that the cDNA clones were derived from the same mRNA.

Nucleotide and deduced amino acid sequence of λ PG-LA 51

The λ PG-LA 51 cDNA insert was composed of two *Eco*RI fragments. The nucleotide sequence of the longer 1.9 kb fragment was determined after random fragmentation (Bankier & Barrell, 1983). The shorter 370 bp fragment was ligated into the *Eco*RI site of phage M13 and sequenced. The orientation of the 370 bp fragment relative to the 1.9 kb *Eco*RI fragment was determined by sequence analysis using specific synthetic oligonucleotides as primers (Suzuki *et al.*, 1985). The results also confirmed the presence of a single *Eco*RI site on λ PG-LA 51.

The λ PG-LA 51 nucleotide sequence is 2263 bp long and codes for a 719 amino-acid-residue polypeptide, which represents approx. 20% of the protein core (Fig. 1). The deduced amino acid sequence contains a 20-residue-long sequence which has previously been identified in a CNBr peptide derived from bovine nasal cartilage PG-LA (Périn *et al.*, 1984). The C-terminal amino acid sequence is also homologous with the sequence of rat (Doege *et al.*, 1986) and chicken (Sai *et al.*, 1986) PG-LA and contains 10 cysteine residues which could be involved in intrachain disulphide bonds. This region of the core protein has a globular appearance when studied by electron microscopy after rotary shadowing (Paulsson *et al.*, 1987). A region of this globular domain (amino acid residues 513–631 in Fig. 1) has been shown to be homologous with chicken liver lectin (Doege *et al.*, 1986; Sai *et al.*, 1986). It is not known, however, if this region actually binds *N*-acetylglucosamine or other carbohydrates. The rat, chicken and bovine PG-LA sequences are highly conserved from the region with lectin homology, i.e. cysteine residue 509 in Fig. 1 to the C-terminus. The amino acid sequence N-terminal to this cysteine residue is less conserved between the species. The putative lectin region shows a homology with more than 80% identity, while the more N-terminal region contains about 30% conserved amino acid residues. It is possible that the cysteine-rich, lectin-homologous globular region has important functions and that therefore the primary structure of this segment has been conserved during

[illegible]

Ser-Gly dipeptides representing putative attachment sites for chondroitin sulphate chains are underlined. The homology domains (I-III) are indicated by vertical lines. Amino acid residues different from the bovine sequence in the rat chondrosarcoma proteoglycan (Doerge *et al.*, 1986) are shown under the bovine sequence. The sequence underlined with a broken line has previously been determined by amino acid sequence analysis (Périn *et al.*, 1984).

The amino acid sequence deduced from λ PG-LA 51 upstream of the putative lectin domain contains numerous Ser-Gly sequences. It has previously been shown that serine residues having a C-terminal glycine residue are preferred sites for the xylose transferase and

therefore potential sites for attachment of chondroitin sulphate chains. In support, chondroitin sulphate peptides isolated from cartilage proteoglycans have been shown to contain a glycine residue next to the substituted serine residue (Johnson & Baker, 1973). Furthermore, the single dermatan sulphate chain of a small proteoglycan from bovine skin is bound to a Ser-Gly structure (Pearson *et al.*, 1983) as are the heparin chains of a mast cell proteoglycan (Robinson *et al.*, 1978) and the chondroitin sulphate chains of a proteoglycan from a yolk sac carcinoma (Bourdon *et al.*, 1985). Thus it is likely that at least a portion of the Ser-Gly sequences are substituted with chondroitin sulphate chains.

Fig. 2. Homology domains in the proteoglycan amino acid sequence deduced from λPG-LA 51

The domains indicated in Fig. 1 are shown aligned and identical amino acid residues are boxed. A 10-amino-acid-residue repeat and the predicted secondary structure (Garnier *et al.*, 1978) of the homology domains are shown.

Internal homology domains

The chondroitin sulphate-binding region of the core protein is arranged in domains containing 12–15 Ser-Gly dipeptide repeats. The amino acid sequence deduced from λPG-LA 51 contains one *N*-terminal incomplete and two complete domains. Alignment of the domains shows that they are homologous (Fig. 2). The homology domains consist of 72–76-residue-long regions rich in Ser-Gly sequences separated by 27–28-residue-long regions devoid of Ser-Gly dipeptides. The Ser-Gly repeats located in the *C*-terminal region of the chondroitin sulphate-rich region are not arranged in homology domains. The approx. 100-amino-acid-residue domains are highly homologous in the Ser-Gly-rich regions and homologous to a lesser extent in the regions separating each cluster of Ser-Gly sequences. The presence of tandemly repeated homology domains suggests that the chondroitin sulphate-substituted region in the proteoglycan core protein has evolved through a series of gene duplications. Presumably the entire chondroitin sulphate-rich region is composed of several repeats of this 100-residue homology domain.

Analysis of the secondary structure of the homology domains (Garnier *et al.*, 1978) predicts that the regions separating the Ser-Gly-rich regions primarily assume an α-helical structure. The Ser-Gly-containing regions in each domain appear to prefer random as well as β-pleated sheet structures.

Indicated in Fig. 2 is a 10-residue-long amino acid sequence which is repeated some six times in the longer homology domains. Investigation of the Ser-Gly sequences deduced from λPG-LA 51 showed that 70% of the dipeptides could be accounted for in the sequence shown in Fig. 3. These 10-amino-acid-residue segments are also present in parts of the core protein without homology with the 100-residue homology domain. The typical features in this sequence are an acidic amino acid residue

in position 2, a Ser-Gly sequence in positions 4 and 5 and preferentially another Ser-Gly dipeptide in positions 8 and 9. This short homology repeat is presumably important for how the serine residues become substituted with glycosaminoglycans or oligosaccharides. Interestingly, this 10-residue sequence is similar to a sequence in the chondroitin sulphate proteoglycan isolated from a yolk sac carcinoma (Bourdon *et al.*, 1985). This proteoglycan contains a 48-amino-acid residue region which exclusively consists of Ser-Gly repeats with no intervening sequences and contains some 12–15 chondroitin sulphate chains. Also in the yolk sac carcinoma proteoglycan the first Ser-Gly dipeptide is preceded by an acidic amino acid residue and followed by a Ser-Gly sequence in the same positions as the conserved residues in the 10-amino-acid-residue sequence in Fig. 3. Similarly the serine residue substituted with dermatan sulphate in the proteoglycan isolated from calf skin is part of a Ser-Gly dipeptide preceded at the position two residues *N*-terminally by an aspartic acid residue (Pearson *et al.*, 1983). At present it is unclear to what extent the serine residues in the 10-amino-acid repeats are substituted with chondroitin sulphate. The structural similarities to other proteoglycans, however, suggest that serine residues at position 4 in Fig. 3 carry a polysaccharide chain. This would lead to six or seven chondroitin sulphate chains per 100-amino-acid-residue homology domain. If all serine residues in the 10-amino-acid-residue repeat were substituted with chondroitin sulphate, the total would be nine to eleven chains per 100-residue homology domain.

Previous studies using fragmentation or proteoglycans with proteolytic enzymes have suggested that the chondroitin sulphate chains are distributed in clusters along the core protein. Each cluster contains an average of four chains, but clusters containing as many as ten chains were observed (Heinegård & Hascall, 1974b). Further indications for clustering of chondroitin sulphate

1	2	3	4	5	6	7	8	9	10
Pro 38%	Asp 48%	Leu 22%			Glu 30%	Pro 30%		Gly 61%	Ile 17%
Leu 19%	Glu 52%	Ser 22%	Ser	Gly	Leu 26%	Ala 22%	Ser 70%	Ala 13%	Leu 17%
Thr 19%		Val 17%			Thr 9%	Thr 22%		Val 13%	Thr 17%

Fig. 3. The 10-amino-acid-residue sequence representing 70% of the Ser-Gly sequences deduced from λPG-LA 51

The percentage of amino acid residues at each position represent a summary of 36 Ser-Gly dipeptides.

chains along the core protein have been obtained by electron microscopy of spread molecules (Thyberg *et al.*, 1975). The somewhat larger number of presumptive glycosaminoglycan-binding Ser-Gly sequences observed in each homology domain in the present study may indicate that serine residues are only partially substituted with chondroitin sulphate. It is possible that some of the serine residues are substituted with keratan sulphate or *O*-glycosidically linked oligosaccharides rather than chondroitin sulphate.

A search in the National Biomedical Research Foundation protein sequence data bank revealed homologies with a number of mouse, rabbit and human immunoglobulin κ and λ light chains. The strongest homology, 38% identical amino acid residues spanning over a 71-residue-long sequence, was observed between a region in the proteoglycan located C-terminal to the 100-residue homology domains (residues 246–315, Fig. 1) and the variable region of mouse M167 immunoglobulin κ light chain (Rudikoff & Potter, 1978). The homology involves the more conserved framework sequences in the variable region of the light chains. Consequently a number of other κ and λ light chains were homologous with the proteoglycan. The most frequently observed homology with the light chains was between residues 302 and 315 in the proteoglycan sequence. The significance of this homology is not known, but may be a result of convergence, suggesting that the homologous regions have a similar function in the two proteins.

We are grateful to Ms. Sara Axelsson for skilful technical assistance. This work was supported by grants from the Swedish Medical Research Council (07478 and 05668), Folksam Yrkeskadors Stiftelse, Kocks Stiftelse, Crafoords Stiftelse, Österlunds Stiftelse and the Medical Faculty, University of Lund.

REFERENCES

- Bankier, A. T. & Barrell, B. C. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., ed.), vol. B5-08, pp. 1–34, Elsevier, Ireland.
- Bourdon, M. A., Oldberg, Å., Pierschbacher, M. D. & Ruoslahti, E. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1321–1325.
- Coudron, C., Loerner, T., Jacobson, I., Rodén, L. & Swartz, N. B. (1980) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **39**, 1671 (abstr.).
- Doege, K., Fernandez, P., Hassel, J. R., Sasaki, M. & Yamada, Y. (1986) *J. Biol. Chem.* **261**, 8108–8111.
- Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120**, 97–120.
- Hascall, V. C. & Heinegård, D. (1974) *J. Biol. Chem.* **249**, 4232–4241.
- Heinegård, D. & Axelsson, I. (1977) *J. Biol. Chem.* **252**, 1971–1979.
- Heinegård, D. & Hascall, V. C. (1974a) *Arch. Biochem. Biophys.* **165**, 427–441.
- Heinegård, D. & Hascall, V. C. (1974b) *J. Biol. Chem.* **249**, 4250–4256.
- Heinegård, D. & Paulsson, M. (1984) in *Extracellular Matrix Biochemistry* (Piez, K. A. & Reddi, A. H., eds.), pp. 277–328, Elsevier, New York.
- Heinegård, D., Sheehan, J., Sommarin, Y., Wieslander, J. & Paulsson, M. (1985a) *Biochem. J.* **225**, 95–105.
- Heinegård, D., Björne-Persson, A., Cöster, L., Franzén, A., Gardell, S., Malmström, A., Paulsson, M., Sandfalk, R. & Vogel, K. (1985b) *Biochem. J.* **230**, 181–194.
- Heinegård, D., Franzén, A., Hedbom, E. & Sommarin, Y. (1986) *CIBA Found. Symp.* **124**, 69–88.
- Johnson, A. H. & Baker, J. R. (1973) *Biochem. Soc. Trans.* **1**, 277–279.
- Oldberg, Å., Franzén, A. & Heinegård, D. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8819–8823.
- Paulsson, M., Mörgelin, M., Wiedemann, H., Beardmore-Gray, M., Dunham, D., Hardingham, T., Heinegård, D., Timpl, R. & Engel, J. (1987) *Biochem. J.*, in the press.
- Pearson, C. H., Winterbottom, N., Fature, D. S., Scott, P. G. & Carpenter, M. R. (1983) *J. Biol. Chem.* **258**, 15101–15104.
- Périn, J.-P., Bonnet, F., Jollès, J. & Jollès, P. (1984) *FEBS Lett.* **176**, 37–42.
- Robinson, H. M., Horner, K. A., Höök, M., Ögren, S. & Lindahl, U. (1978) *J. Biol. Chem.* **253**, 6687–6693.
- Rudikoff, S. & Potter, M. (1978) *Biochemistry* **17**, 2703–2707.
- Sai, S., Tanak, T., Kosher, R. A. & Tanzer, M. L. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 5081–5085.
- Sikder, S. K., Kabat, E. A., Steer, C. J. & Ashwell, G. (1983) *J. Biol. Chem.* **258**, 12520–12525.
- Sommarin, Y. & Heinegård, D. (1983) *Biochem. J.* **214**, 777–784.
- Suzuki, S., Oldberg, Å., Hayman, E. G., Pierschbacher, M. D. & Ruoslahti, E. (1985) *EMBO J.* **4**, 2519–2524.
- Thyberg, J., Lohmander, S. & Heinegård, D. (1975) *Biochem. J.* **151**, 157–166.
- Upholt, W. B., Vertel, B. M. & Dorfman, A. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 4847–4851.
- Wiedemann, H., Paulsson, M., Timpl, R., Engel, J. & Heinegård, D. (1984) *Biochem. J.* **224**, 331–333.
- Wieslander, J. & Heinegård, D. (1980) *Biochem. J.* **187**, 687–694.

Received 5 December 1986/2 February 1987; accepted 3 February 1987