Molecular cloning and sequence analysis of the cDNA for small proteoglycan II of bovine bone

Agnes A. DAY,* Camille I. McQUILLAN, John D. TERMINE and Marian R. YOUNG

Bone Research Branch, National Institute of Dental Research, National Institutes of Health, Bethesda, MD 20892, U.S.A.

The cDNA for the full-length core protein of the small chondroitin sulphate proteoglycan II of bovine bone was cloned and sequenced. A 1.3 kb clone (λ Pg28) was identified by plaque hybridization with a previously isolated 1.0 kb proteoglycan cDNA clone (λ Pg20), positively identified previously by polyclonal and monoclonal antibody reactivity and by hybrid-selected translation *in vitro* [Day, Ramis, Fisher, Gehron Robey, Termine & Young (1986) *Nucleic Acids Res.* **14**, 9861–9876]. The cDNA sequences of both clones were identical in areas of overlap. The 360-amino-acid-residue protein contains a 30-residue propeptide of which the first 15 residues are highly hydrophobic. The mature protein consists of 330 amino acid residues corresponding to an M_r of 36 383. The core protein contains three potential glycosaminoglycan-attachment sites (Ser-Gly), only one of which is within a ten-amino-acid-residue homologous sequence seen at the known attachment sites of related small proteoglycans. Comparisons of the published 24-residue *N*-terminal protein sequence of bovine skin proteoglycan II core protein with the corresponding region in the deduced sequence of the bovine bone core protein reveals complete homology. Comparison of the cDNA-derived sequences of bovine bone and human embryonic fibroblast proteoglycans shows a hypervariable region near the *N*-terminus. Nucleotide homology between bone and fibroblast core proteins was 87 %, and amino acid homology was 90 %.

INTRODUCTION

Biological functions for proteoglycans (PGs) may be as diverse as the tissues in which they are found. These functions include conferment of compressive properties to cartilage [1,2], interaction with collagen fibrils during fibrillogenesis [2–4], stabilization of basement membranes and other connective tissues [5–7] and the regulation of biological clocks [8]. The small proteoglycans of bone [5,9,10] are typically representative of one form of these macromolecules, as they appear in a variety of connective tissues [10]. We previously isolated cDNA clones for the most abundant bone proteoglycan (PG-II) and determined the distribution of its mRNA in relation to that of small proteoglycans in skin, cornea, smooth muscle, articular cartilage and tendon [11]. Our studies also found that PG-II expression was sensitive to viral transformation, as are those of other connective-tissue proteins [11]. In the present paper we describe the complete primary sequence of the bovine bone PG-II core protein, deduced from these cDNA clones, and compare it with those of other proteoglycan core proteins from bovine skin [12,13] and human placenta [14,15].

MATERIALS AND METHODS

Screening of the bacteriophage library

The isolation, purification and partial characterization of the λ Pg20 cDNA was reported previously [11]. Insert from the 1.0 kb cDNA clone was nick-translated (Amersham), and 3.2×10^7 c.p.m. was used to rescreen a

previously constructed bovine bone λ gt11 cDNA library [16] by the plaque hybridization method of Young & Davis [17]. Of the 40 000 recombinants screened, 75 positive clones were identified and eight were rescreened to purity. Isolation of selected insert clones by *Eco*RI digestion and agarose-gel electrophoresis [18] were performed. Clone λ Pg28 was selected for its size (1.3 kb), which closely correlated with the estimated size of the mRNA for the core protein of bovine bone PG-II [11].

Subcloning and sequence determination

Inserts for λ Pg20 and λ Pg28 were subcloned into M13mp18 and M13mp19 bacteriophage sequencing vectors. The sequencing strategy consisted of (A) ordered fragment sequencing with the use of subcloned restriction fragments of the inserts, (B) M13 rapid deletion subcloning (Cyclone System, IBI) and (C) utilization of synthetic 17-mer and 20-mer oligonucleotides based upon sequence derived from methods A and B. The oligonucleotides were synthesized on a DNA synthesizer 380B (Applied Biosystems). The sequences of the oligonucleotides are as follows from 5' to 3':

MY18, TGGTGATTTTGTGGCA
MY19, GATTGCAGAGATATTGTTG
MY20, ACTTTAAGAACCTGAAGAAC
MY21, CTCTTCAGGAGCTGCGTGTC
MY22, CCATTTTCAATGCCTGAGCT
MY23, TCCGATGTGTCTATGTGCGT
MY24, TTATAAAAATGAGGGTTTTC
MY25, GAGCACTTTCCTGAAGTTCC

Abbreviation used: PG, proteoglycan.

* To whom correspondence should be addressed.

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00712.

The Sanger dideoxynucleotide chain-termination method [19], with 5'-[α - ^{32} P]ATP as the radioactive label (Amersham), was used for all subcloned fragments. The M-13 primers (17-mer and 40-mer included in sequencing kit) that hybridized to vector sequences were used for sequencing methods A and B.

Computer analysis

Analyses of cDNA sequences were facilitated by utilizing programs of the University of Wisconsin Genetics Computer Group (UWGCG). The NBRF protein database and GenBank nucleotide database were searched for homologies by using the Integrated Database and Extended Analysis System for Nucleic Acids and Proteins (IDEAS).

RESULTS

Isolation of full-length cDNA clones for bovine bone PG-II

After isolation of selected clones, λ Pg28 was chosen for sequencing on the basis of its size correlation with the message for PG-II protein. A 1.0 kb λ Pg20 clone was previously isolated [11] by immunoscreening and was used to confirm and complement the sequence derived from λ Pg28.

Determination of the nucleotide sequence

The restriction map analysis of λ Pg20 and λ Pg28 as well as the sequencing strategy are shown in Fig. 1. The complete coding sequence of the protein was determined by sequencing both strands by utilizing restriction-enzyme-generated fragments of clones and synthetic oligonucleotides (20-mers) derived from these sequences.

The results generated a continuous cDNA sequence of 1250 bases including 5' and 3' flanking sequences (Fig. 2). The ATG initiation codon is preceded by 87 nucleotide residues containing two termination codons in the same reading frame. The coding region of the cDNA consists of 1080 bases corresponding to 360 amino acid residues. The termination codon (TAG) occurs at bases 1168–1170. A potential polyadenylation signal, ATAA, is 23 bases downstream of the stop codon and 30 bases upstream of the poly(A) tail.

Analysis of the deduced core protein of PG-II

Analysis of hydropathy for the deduced amino acid sequence showed an *N*-terminal 15-amino-acid-residue hydrophobic region typical of a signal peptide [20]. A comparison of this amino acid sequence with published *N*-terminal core protein sequences for the small proteoglycans of bovine skin [12,13] and human bone [9] reveals an additional 15 *N*-terminal amino acid residues that are not part of the secreted tissue protein and that may represent a propeptide. Three possible glycosaminoglycan-attachment sites (Ser-Gly) are present in the protein, at Ser-4, Ser-160 and Ser-296. There were three potential acceptor sites for asparagine (*N*)-linked glycosylation, located at Asn-182, Asn-233 and Asn-274, all in the latter half of the protein.

The predicted 360-amino-acid-residue precursor protein has an M_r of 39 809, and the mature core protein is composed of 330 amino acid residues and has an M_r of 36 383. There is an internal homologous repeat of 19 residues present in the mature protein (residues 46–64 and 186–204). The homologies for these two sequences are 57% at the nucleotide level and 53% at the protein level.

Comparison of our cDNA sequence for the small

Fig. 1. Restriction-endonuclease map and sequencing strategy of cDNA clones λ Pg20 and λ Pg28

The coding region of the PG-II sequence is indicated by the hatched bar. Arrows indicate the direction and length of sequences read utilizing both M-13 bacteriophage deletion mutants (Cyclone sequencing) and ordered fragments (restriction digest fragments subcloned into M-13, thin arrows). MY indicates oligonucleotide primed sequencing (heavy arrows).

1	TTG GGT AAT AAG GGA AAC CAT TAT CTT ATT TAG CCT TTA CAA CAC TCT CAC AAG GTT CCT	60
61	CGG TTG TGA AAA TAT ACA AGA GAA ATC <u>ATG</u> AAG GCA ACT ATC ATC TTT CTC TTG GTT GCA	120
	<u>Met</u> Lys Ala Thr Ile Ile Phe Leu Leu Val Ala	
121	CAA GTT TCC TGG GCT GGA CCA TTT CAA CAG AAA GGC TTA TTT GAC TTT ATG CTG GAA GAT	180
	Gln Val Ser Trp Ala Gly Pro Phe Gln Gln Lys Gly Leu Phe Asp Phe Met Leu Glu Asp	
181	GAG GCT TCT GGG ATA GGC CCG GAA GAG CAC TTT CCT GAA GTT CCT GAA ATA GAG CCT ATG	240
	Glu Ala Ser Gly Ile Gly Pro Glu Glu His Phe Pro Glu Val Pro Glu Ile Glu Pro Met	
241	GGC CCA GTC TGC CCC TTC CCG GGT CAG TGC CAT CTG CGA GTT GTC CAG TGT TCT GAT CTG	300
	Gly Pro Val Cys Pro Phe Arg Cys Gln Cys His Leu Arg Val Val Gln Cys Ser Asp Leu	
301	GGT CTG GAA AAA GTG CCA AAA GAC CTT CCT CCT GAT ACT GCG CTG CTG GAC CTG CAA AAC	360
	Gly Leu Glu Lys Val Pro Lys Asp Leu Pro Pro Asp Thr Ala Leu Leu Asp Leu Gln Asn	
361	AAC AAA ATA ACT GAG ATC AAA GAT GGA GAC TTT AAG AAC CTG AAG AAC CTT CAT ACA CTG	420
	Asn Lys Ile Thr Glu Ile Lys Asp Gly Asp Phe Lys Asn Leu Lys Asn Leu His Thr Leu	
421	ATT CTC ATC AAC AAC AAA ATT AGC AAA ATC AGC CCT GGG GCA TTT GCT CCT TTG GTG AAA	480
	Ile Leu Ile Asn Asn Lys Ile Ser Lys Ile Ser Pro Gly Ala Phe Ala Pro Leu Val Lys	
481	TTG GAA CGA CTT TAT CTT TCC AAG AAT CAA CTG AAG GAA TTG CCA GAG AAA ATG CCC AAA	540
	Leu Glu Arg Leu Tyr Leu Ser Lys Asn Gln Leu Lys Glu Leu Pro Glu Lys Met Val Lys	
541	ACT CTT CAG GAG CTG CGT GTC CAT GAG AAC GAG ATC ACC AAA GTG CGG AAA TCT GTG TTC	600
	Thr Leu Gln Glu Leu Arg Val His Glu Asn Glu Ile Thr Lys Val Arg Lys Ser Val Phe	
601	AAT GGA TTG AAC CAG ATG ATC GTC GTA GAA CTT GGC ACC AAC CCG CTG AAG AGC TCA GGC	660
	Asn Gly Leu Asn Gln Met Ile Val Val Glu Leu Gly Thr Asn Pro Leu Lys Ser Ser Gly	
661	ATT GAA AAT GGA GCC TTT CAG GGA ATG AAG AAG CTC TCC TAC ATC CGC ATT GCT GAC ACA	720
	Ile Glu Asn Gly Ala Phe Gln Gly Met Lys Lys Leu Ser Tyr Ile Arg Ile Ala Asp Thr	
721	AAT ATA ACT ACC ATC CCT CAA GGT CTT CCT CCT TCC CTT ACT GAA TTA CAT CTC GAT GGC	780
	Asn Ile Thr Thr Ile Pro Gln Gly Leu Pro Pro Ser Leu Thr Glu Leu His Leu Val Gly	
781	AAC AAA ATC ACC AAA GTT GAT GCA GCT AGC CTG AAA GGA CTG AAT AAT TTG GCT AAG TTG	840
	Asn Lys Ile Thr Lys Val Asp Ala Ala Ser Leu Lys Gly Leu Asn Asn Leu Ala Lys Leu	
841	GGA CTG AGT TTC AAC AGC ATC TCT GCG GTT GAC AAT GGC TCG TTG GCC AAC ACT CCT CAT	900
	Gly Leu Ser Phe Asn Ser Ile Ser Ala Val Asp Asn Gly Ser Leu Ala Asn Thr Pro His	
901	TTG AGG GAA CTT CAT TTG AAC AAC AAC AAG CTT GCT AAA GTG CCC GGT GGC GTG GCC GAT	960
	Leu Arg Glu Leu His Leu Asn Asn Asn Lys Leu Ala Lys Val Pro Gly Gly Val Ala Asp	
961	CAT AAG TAC ATT CAG GTT GTC TAC CTT CAC AAC AAC AAT ATC TCT GCA ATC GGC TCT AAC	1020
	His Lys Tyr Ile Gln Val Val Tyr Leu His Asn Asn Asn Ile Ser GCA Ile Gly Ser Asn	
1021	GAC TTC TGC CCA CCC GGA TAC AAC ACC AAA AAG GCT TCT TAT TCA GGA GTG AGC CTT TTC	1080
	Asp Phe Cys Pro Pro Gly Tyr Asn Thr Lys Lys Ala Ser Tyr Ser Gly Val Ser Leu Phe	
1081	AGC AAC CCA GTC CAG TAC TGG GAG ATC CAG CCA TCC ACC TTC CGA TGT GTC TAT GTG CGT	1140
	Ser Asn Pro Val Gln Tyr Trp Glu Ile Gln Pro Ser Thr Phe Arg Cys Val Tyr Val Arg	
1141	GCT GCC GTT CAG CTT GGA AAC TAC AAG TAC CTC CCA AGA AAA CCC TCA TTT TTA TAA CCT	1200
	Ala Ala Val Gln Leu Gly Asn Tyr Lys Lys Lys Lys Lys Lys Lys Lys Lys Lys Lys Lys	
1201	GGC AAA ATC CCA TTA ATG TCA TTG CTC AGG AAA AAC AAA AAA AAA AAA AA	1250

Fig. 2. Nucleotide sequence of cDNA clones and corresponding amino acid sequence

Nucleotides are numbered to left and right of the Figure, and amino acid numbering occurs within. Amino acid residues occurring within the pre-propeptide are negatively numbered. Arrow indicates the cleavage site of the putative pre-propeptide. The initiation and termination codons are boxed. Open triangles represent possible glycosaminoglycan-attachment sites, and closed triangles highlight possible *N*-linked oligosaccharide-attachment sites. The 19-residue homologous repeats are underlined. Cysteine residues are indicated by closed circles. The potential polyadenylation signal is heavily underlined.

1	MKATIIIFLLV	LAQVSWAGPFQ	QKGLDFML	LEDEASGIGPE	EHFPEVPEIEP	MGPVPCFRCQ	60
		L	L	R	V	P	
61	CHLRVVQCS	D	LGLEKVPKD	LPPDTA	LLDLQNNKITE	IKDGFKNLKNLH	120
			D	T		A	V
121	ISPGAFAPL	V	V	KLERLYLSK	NQLKELPEK	MPKTLQELRVH	180
		T				A	VT
181	ELGTNPLKS	S	SGIENGAFQ	G	MKKLSYIRI	ADTNITIPQGL	240
					S		SR
241	SLKGLNNLA	K	LGLSFNSIS	A	V	DNGLANTPHL	300
						D	TR
301	HNNNISAIG	S	NDFCPPG	Y	N	TKKASYSGV	360
	VV	S		H		S	I

Fig. 3. Comparison of bovine bone PG-II and human fibroblast core protein sequences

The bar above the *N*-terminal part of the sequence indicates complete homology between the deduced bovine bone PG-II sequence and the bovine skin proteodermatan sulphate *N*-terminal sequence as derived via Edman degradation. The reverse font highlights the hypervariable region between bone PG-II (full sequence) and human fibroblast PG40 (indicated only where amino acid residues differ).

chondroitin sulphate PG-II of bovine bone with the nucleotide sequence of the dermatan sulphate proteoglycan expressed in human fibroblasts [14] reveals a 52% homology in the 5' untranslated region and an 82% homology of 82 bases of the 3' flanking sequences. The coding regions of these two proteins display an 87% homology at the nucleotide level and a 90% homology at the amino acid level (Fig. 3). There is an 11-amino-acid-residue hypervariable region (9% homology between the two core proteins) (residues 10–20 of the secreted protein). The signal-peptide regions are 90% homologous. Comparison of these sequences with the *N*-terminal amino acid residues of the core proteins from bovine skin [12] and tendon [21] proteodermatan sulphates shows complete homology between the bovine skin and bone species (Fig. 3).

DISCUSSION

We have isolated cDNA clones for the small proteoglycan II of bovine bone from a λ gt11 bovine bone cell library by using a combination of immunoscreening and plaque hybridization. We report here the complete cDNA coding sequence and the deduced amino acid sequence of the core protein.

The core protein contains a 30-amino-acid-residue propeptide, of which the first 15 residues are highly hydrophobic and probably represent a signal peptide. The mature core protein is 330 amino acid residues long and has an M_r of 36383. The tetrapeptide Ser-Gly-Xaa-Gly has been reported to be a recognition consensus sequence for glycosaminoglycan chain attachment to proteoglycan core proteins [22]. This consensus sequence occurs at only one of the three Ser-Gly repeats in bone PG-II and agrees with the observations by Chopra *et al.* [13] of Ser-4 as the site of glycosaminoglycan chain attachment for skin proteodermatan sulphate. The bone PG-II molecule is known to have one chain per molecule [23], and the adjacent amino acid sequence (Asp-Glu-Ala-Ser-Gly-Ile-Gly-Pro-Glu) is the same as that reported for the bovine skin, bovine tendon and human fetal membrane proteoglycan core proteins [13,14,21]. It is likely that the bovine bone protein also has an attached glycosaminoglycan chain at this site. Our deduced amino acid sequence reveals four cysteine residues closely

aligned in the *N*-terminal region of the PG-II core protein and two at the C-terminus. Fisher *et al.* [5] reported the presence of at least one *N*-linked oligosaccharide in bovine bone PG-II. There were three potential acceptor sites for *N*-linked glycosylation in the latter half of the protein.

Our deduced protein sequence was compared with a recently cloned chondroitin/dermatan sulphate proteoglycan expressed by a human embryonic fibroblast cell line [14]. The two sequences were 87% homologous at the nucleotide level and 90% homologous at the protein level. The sequence obtained here for the bovine bone core protein was identical with that reported for the 24-residue *N*-terminus of bovine skin proteoglycan [12] and the 14-residue *N*-terminus of bovine tendon proteoglycan [21]. Therefore the hypervariable region (residues 10–20 of the secreted protein) may reflect species differences. Alternatively, since the human fibroblast proteoglycan was derived from transformed cells, sequence variability might also arise from transformation.

Vogel & Fisher [21] compared bovine bone, tendon and cartilage proteoglycans by immunochemical and proteolytic cleavage methods. As suggested by others [2,10], they found that these small proteoglycans were identical both immunochemically and on one-dimensional peptide maps [21]. However, the tendon and cartilage proteoglycans contained a major, rapid, cleavage site for *Staphylococcus aureus* V8 proteinase, whereas that of bovine bone did not [21]. They speculated that this major V8-proteinase site may be blocked in the bone core protein as a result of small differences either in primary structure or in post-translational modification.

Previous studies from our laboratory suggested the existence of one gene copy of PG-II core protein per haploid genome [11]. Northern-blot analyses of various bovine tissues known to contain small proteoglycans revealed similar mRNAs and demonstrated an analogous species in the human, chick and mouse [11]. The high degree of homology between bovine bone and human fibroblast core proteins strongly supports the notion of an evolutionarily conserved single-copy gene for the PG-II core protein. Subtle differences in protein structure or post-translational modification, including the type of glycosaminoglycan chains attached, may convey, however, tissue and functional specificities.

REFERENCES

1. Hascall, V. (1981) in *Biology of Carbohydrates*, vol. 1 (Ginsberg, V., ed.), pp. 1–49, John Wiley and Sons, New York
2. Heinegård, D. & Paulsson, M. (1984) in *Extracellular Matrix Biochemistry* (Piez, K. & Reddi, A., eds.), pp. 277–328, Elsevier/North-Holland, New York
3. Scott, J. & Haigh, M. (1985) *Biosci. Rep.* **5**, 765–774
4. Vogel, K., Paulsson, M. & Heinegård, D. (1984) *Biochem. J.* **223**, 587–597
5. Fisher, L., Termine, J., Dejter, S., Whitson, S., Yanagishita, M., Kimura, J., Hascall, V., Kleinman, H., Hassell, J. & Nilsson, B. (1983) *J. Biol. Chem.* **258**, 6588–6594
6. Schmidt, G., Robenek, H., Harrach, B., Glossl, J., Nolte, V., Hormann, H., Richter, H. & Kresse, H. (1987) *J. Cell Biol.* **104**, 1683–1691
7. Doege, K., Hassell, J., Caterson, B. & Yamada, Y. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3761–3765
8. Reddy, P., Jacquier, A., Abovich, N., Petersen, G. & Rosbash, M. (1986) *Cell* **46**, 53–61
9. Fisher, I., Hawkins, G., Tuross, N. & Termine, J. (1987) *J. Biol. Chem.* **262**, 9702–9709
10. Heinegård, D., Franzen, A., Hedbom, E. & Sommarin, Y. (1986) *Ciba Found. Symp.* **124**, 69–82
11. Day, A., Ramis, C., Fisher, I., Gehron Robey, P., Termine, J. & Young, M. (1986) *Nucleic Acids Res.* **14**, 9861–9876
12. Pearson, C., Winterbottom, N., Fackre, D. & Scott, P. (1983) *J. Biol. Chem.* **258**, 15101–15104
13. Chopra, R., Pearson, C., Pringle, G., Fackre, D. & Scott, P. (1985) *Biochem. J.* **232**, 277–279
14. Krusius, T. & Ruoslahti, E. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7683–7687
15. Brennan, M., Oldberg, A., Pierschbacher, M. & Ruoslahti, E. (1984) *J. Biol. Chem.* **259**, 13742–13750
16. Young, M., Bolander, M., Day, A., Ramis, C., Gehron Robey, P., Yamada, Y. & Termine, J. (1986) *Nucleic Acids Res.* **14**, 4483–4499
17. Young, R. & Davis, R. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1194–1198
18. Maniatis, T., Fritsch, E. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor
19. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
20. von Heigne, G. (1984) *J. Mol. Biol.* **173**, 243–251
21. Vogel, K. & Fisher, L. (1986) *J. Biol. Chem.* **261**, 11334–11340
22. Bourdon, M., Krusius, T., Campbell, S., Swartz, N. & Ruoslahti, E. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 3194–3198
23. Fisher, L. (1985) in *The Chemistry and Biology of Mineralized Tissues* (Butler, W., ed.), pp. 188–196, EBSCO Media, Birmingham

---

Received 26 August 1987/5 October 1987; accepted 6 October 1987